Sensors and Materials, Vol. 34, No. 2 (2022) 467–491
MYU Tokyo

S & M 2822

Semantic Image Segmentation in Similar Fusion Background for Self-driving Vehicles

ChienHsiang Wu,[1] TzuChi Tai,[2] and ChinFeng Lai[1*]

[1]Department of Engineering Science, National Cheng Kung University, Tainan 7014021, Taiwan ROC
[2]Taiwan Semiconductor Manufacturing Co. Ltd., Hsinchu 300091, Taiwan ROC

(Received May 6, 2021; accepted December 1, 2021)

Keywords: image segmentation, image enhancement, self-driving vehicle, similar fusion background image

Self-driving vehicles have become increasingly popular in recent years. Because of this, the information fusion sensing method using radar and cameras has been widely adopted in vehicles. We use the vehicle camera sensor and robust image segmentation technology to solve its inherent shortcomings. The images used for image segmentation are obtained under adverse weather conditions, or the image object's color and texture resemble the background. For such images, using the convolutional layer model for image segmentation as a feature extraction method usually leads to error. Any highly robust algorithms for image enhancement for self-driving operation will help alleviate problems related to driving safety. To ensure that the final image segmentation achieves the desired effect and reduces the error rate, we propose a new segmentation-twice method, which correctly classifies the object's label. The test results of the simulation described in this paper show that this experiment correctly classifies the object's label. It can provide accurate environmental perception information for autonomous vehicles, improve the segmentation effect of similar fusion background images, and reduce the error rate.

1. Introduction

With the rapid development of technology, self-driving vehicles are undoubtedly one of the most significant technological inventions in recent years.(1) Self-driving cars rely on various sensors to perceive their environment. The environment perception technology of cars is mainly based on sensors for obtaining obstacle movement information. However, these sensors may significantly impact the safety of the control system of autonomous vehicles because of insufficient sensing information and low accuracy. This feature also restricts the popularity of self-driving vehicles.

Sensors are necessary hardware in autonomous vehicles and each sensor has its specific characteristics. The architecture of the self-driving platform is shown in Fig. 1. For example, the advantage of automotive radar is that it has good weather adaptability. Its performance does not degenerate at night and it can accurately obtain the position and speed of the target. However,

[*]Corresponding author: e-mail: cinfon@ieee.org
https://doi.org/10.18494/SAM3427

ISSN 0914-4935 © MYU K.K.
https://myukk.org/

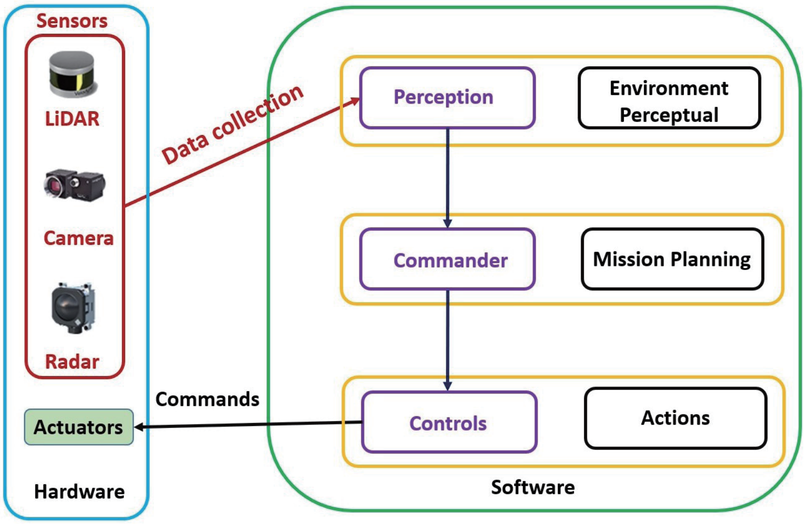


Fig. 1. (Color online) Architecture of the self-driving platform.

there is difficulty achieving target classification for distinguishing stationary vehicles and road traffic signs.

Light detection and ranging (LiDAR) is an optical remote sensing technology that measures parameters,(2) such as the distance to the target, by irradiating a beam of light, usually a pulsed laser, onto the target. The advantage of such a sensor is that it can accurately obtain the three-dimensional information of the object through long-distance detection. However, its disadvantage is that it will be affected by small particles in air.

Low cost, a large amount of data, and easy perception classification are the advantages of car cameras. The disadvantages are poor adaptability to the lighting condition and low accuracy. Therefore, the purpose of this study is to improve the inherent shortcomings of the vehicle camera sensor, particularly for the image segmentation processing of image capture, in order to improve the adaptability and accuracy of object discrimination.

We use the NVIDIA GTX 1070 GPU, and the sampling system is equipped with a Mako G-319 digital camera. Experimental results show that this architecture can achieve a high frame rate of 33.45 fps. The processing speed of such an image is sufficient for real-time segmentation operation.

The technology used by the self-driving system to identify objects on the road is inseparable from semantic image segmentation.(3) For deep learning, image processing technology has been widely utilized in modern machine vision and image classification and recognition. It is also one of the essential application technologies. The convolutional neural network (CNN) has achieved great success in various computer vision tasks.(4) The semantic segmentation task(5) system in self-driving consideration is essential for recognizing object detection and classification images.(6)

The traditional image segmentation mainly involves feature extraction and classification. These features and corresponding classification labels train a classification model commonly found in support vector machines (SVMs) and random forests (RFs). This feature extraction uses an unsupervised learning method. The image classification labels are unused in the extraction process after the feature extraction. The feature extraction method mentioned above is unable to adjust in accordance with the label of the image. If the feature selection is not representative of various categories, the model's accuracy will inevitably deteriorate. During training, CNN can avoid such problems and is able to understand the relationship between learning features and various classification labels from large-scale image datasets. The performance of the convolutional layer in feature extraction is both accurate and fast. Therefore, CNN has gradually replaced the traditional image classification method to become the mainstream algorithm for image processing and also often for intelligent image recognition.(7)

In recent years, self-driving car systems have become increasingly popular. The technology used to recognize objects on the road is also changing with each passing day.(8) Object recognition technology and semantic image segmentation are inseparable. Google's open-source DeepLab is the premier deep learning model for semantic image segmentation. It has four stages: DeepLabv1, DeepLabv2, DeepLabv3, and DeepLabv3+. The introduction of commonly used encoder decoders for semantic segmentation under these architectures is efficient and straightforward, leading to the improvement of the segmentation result.(9)

There are also shortcomings in DeepLab's use of semantic image segmentation. When the color and texture of the object in the image and the background are similar, it will significantly impact the cut-out of the mask and recognition results, and the overall effect will become poor.

To improve the image segmentation issues of objects in a similar fusion background, we design a new preprocessing method to segment the image. The image is preprocessed and then sent to the image segmentation process. Therefore, our method can be applied to any image segmentation method to fundamentally solve object fusion in the background and we apply the preprocessing system to DeepLab for inspection.

We propose an image segmentation method based on similar fusion backgrounds in order to study images that are likely to cause errors in models that use convolutional layers for feature extraction. The main algorithm uses two CNN models for image processing. The PyNET model is utilized to separately enhance the features of the object and background in the image and overlap the enhanced image with the original image in accordance with a certain overlap weight.

The features of the original image and those of the enhanced images are handled simultaneously by overlapping the image. The convolutional layer allows an easy extraction of different features from the object and background. Therefore, the object will not be fused into the background and hence will not be ignored.

To ensure that the final image segmentation will yield the best results, in this study, the image is segmented twice, that is, the approximate object shape is first cut out and then subdivided, and the object label is correctly classified. From the experimental results, the IoU score is determined and used as an indicator to evaluate the performance of the U-Net, DeepLab, and fully convolutional network (FCN) image segmentation models.

Intersection over union (IoU) is a standard measure of the accuracy of the detection of corresponding objects in a specific dataset. It is a simple measurement standard as long as the task obtains a bounding box in the output that can be used for measurement.

$$IoU = \frac{area\,of\;overlap}{area\,of\;union} = \frac{Detection\,Result \cap Ground\,Truth}{Detection\,Result \cup Ground\,Truth} = \qquad (1)$$

The IoU value will be between 0 and 1. The value 0 means that the predicted position deviates entirely from the standard answer and does not even touch one edge. The value 1 means that the predicted position exactly matches the standard answer.

Generally, the benchmark for the judgment of the recognition rate is an IOU value greater than or equal to 0.5.

The research process and contributions are as follows.

- Use the architecture of the CNN model to enhance the features of the original image:
  Since the main reason for the object fusion background is that the object's attributes are too similar to the background features, there is a severe cutting problem in the image segmentation.
- Evaluate the weights suitable for image segmentation:
  The enhanced image will be overlapped with the original image. This overlapping image will have a different intensity depending on the weights of the two images. The adjustment of this weight is an essential factor in the success of image segmentation.
- Use the architecture of another CNN model to implement a mask to reduce the background influence:
  Overlap the original with the enhanced feature image to produce a new image. This new image retains the original features but has enhanced features of the object and weaker features of the background. This overlapped image is fed into the model to generate a mask. The purpose of the mask is to delineate the overall shape of the object to reduce the influence of the background and then improve the result of the segmentation.

The remainder of this paper is organized as follows. In Sect. 2, the research and methods for describing the research background and related algorithms are presented. In Sect. 3, the experiment results are analyzed, and image segmentation and recognition are discussed. Finally, the conclusions are presented in Sect. 4.

2. Research and Methods

To better improve the perception accuracy and reliability of the self-driving vehicle, our method uses the image data obtained by the camera on the vehicle. It uses the trained vision algorithm to obtain the complete information of the target, such as a better identified dynamic target type, a more accurate judgement of the static interference target (e.g., guardrail and traffic sign), and enhanced perceived target recognition. This research is aimed primarily at discriminating similar fusion background images, and machine vision is one of the indispensable technologies for such a purpose.

This work focuses on improving the performance of the existing CNN algorithm for images with similar fusion backgrounds. Such a framework should be effective in enhancing the attention model mechanism at different layers. It is expected that machine vision will approach or even exceed human vision capabilities. The workflow of related research is shown in Fig. 2.

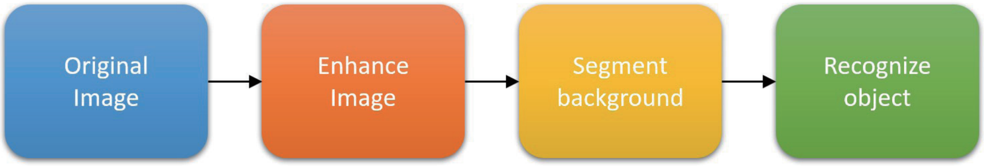


Fig. 2. (Color online) Research workflow.

2.1 Related algorithms research

In Fig. 3, the development of image recognition is presented from left to right. In the beginning, the image was only classified, and the probability of identifying a single object in the image was relatively high. Next, the position of the object in the image was determined and isolated. Finally, each complex pixel corresponded to its relevant part. With a correct classification, the boundary of the object can be determined more accurately. As object recognition becomes increasingly accurate, the improvement of the recognition speed and the reduction in the number of hardware requirements become the goals of subsequent research. In this section, we briefly describe the mechanism and related algorithms, and explain image enhancement, segmentation, and test datasets.

The You Only Look Once (YOLO) model developed by Redmon *et al.*(10) uses the entire image directly as the neural network input at each position of the image. The use of the regression method to identify the target's boundary at this location and the category to which the target belongs markedly improves the overall recognition speed.

Simonyan and Zisserman(11) improved the Visual Geometry Group (VGG) by using AlexNet. The algorithm utilizes continuous small, instead of large, convolution kernels. The advantage of small convolution kernels is that they can increase the depth of the neural network, thereby enhancing the learning effect while reducing the parameters, and the error rate in Top-5 decreases to 7.3% (Top-N is described in the footnote under Table 1). Small convolution kernels also reduce problems such as vanishing gradient problems. However, the VGG architecture requires much memory and is time-consuming because of the excessive number of NN parameters.

The ResNet method skips the connection of the convolutional layer and calls it the residual.(12) Some of the input data do not go through the neural network and jump directly to the output. This method can prevent the vanishing gradient problem during backpropagation while retaining part of the original information. Therefore, the neural network can be made deeper without causing a decrease in accuracy. The depth of the neural network can reach 152 layers.

In the Xception architecture, each channel performs completely independently using a separate spatial convolution kernel.(13) This approach reduces the coupling between different operations and can effectively utilize the existing computing performance.

ResNeXt is based on the ResNet architecture, with each unit expanded horizontally.(14) Simultaneously, different convolution kernel structures extract different features and finally merge them. The ResNeXt structure can achieve a lighter-weight convolution kernel without

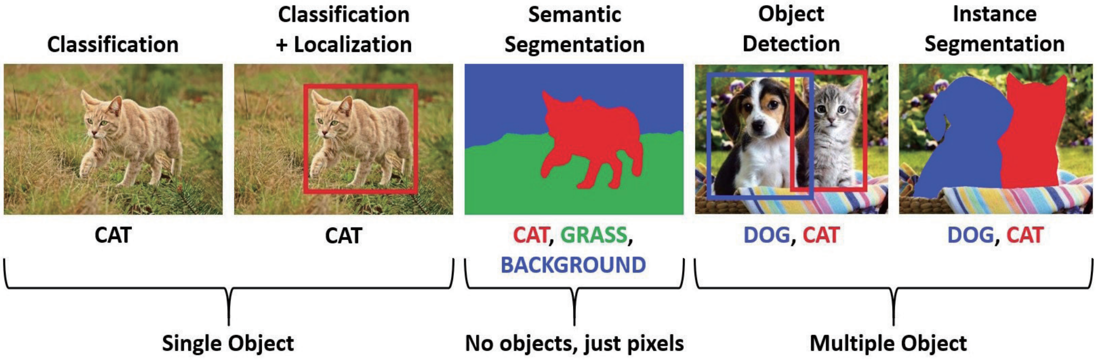


Fig. 3. (Color online) Development and evolution of object recognition.

Table 1
Performance characteristics of various CNN architectures.

Model	Top-1 accuracy	Top-5 accuracy	Parameters	Size (MB)	Depth
VGG16	0.713	0.901	138357544	528	23
Inception v3	0.779	0.937	23851784	92	159
ResNet50	0.749	0.921	25636712	98	—
Xception	0.790	0.945	22910480	88	126
InceptionResNet v2	0.803	0.953	55873736	215	572
ResNeXt50	0.777	0.938	25097128	96	—

Top-1 accuracy is the conventional accuracy: the model result (the one with the highest probability) must be exactly the expected answer.
Top-5 accuracy means that any of the five highest probability results of the model must match the expected answer.

increasing the complexity of the parameters and the same amount of calculation yields an improved prediction accuracy. The performance characteristics of the CNN algorithms mentioned above are compared. The main items to be compared are accuracy and calculation volume, as shown in Table 1.

As seen from Table 1, the earliest VGG16 architecture is the largest and has the lowest accuracy. While reducing the parameters, the Inception v3 architecture also significantly increases the neural network depth and improves the accuracy.(15) ResNet50 is also a deep structure, and its accuracy is not reduced owing to the design of the residual network but is increased compared with the VGG16 architecture.(16) Xception is superior to the Inception architecture, showing improved overall accuracy with reduced neural network depth and parameters. InceptionResNet v2 is composed of Inception and ResNet architectures and exhibits the highest accuracy among the above architectures. Simultaneously, its neural network is the largest in terms of parameters, size, and depth. ResNeXt50 is composed of ResNet and Inception architectures. The overall performance is not outstanding, but it is better than that of the ResNet infrastructure.

In this experiment, a vehicle-mounted camera with a GPU computing architecture is used to intercept and segment image data. The aim is to segment the unusual images observed from the vehicle. It is difficult to obtain the large number of unusual scene images needed to meet the required test image dataset quantity from the image data captured in the experiment. Hence, in this experiment, Oxford IIIT is used as the test dataset. It has more than 7000 pictures and contains various objects and similar background images. This dataset is sufficient for training.

2.2 Image enhancement

It is difficult to use image segmentation to separate images with similar fusion backgrounds. The image semantic segmentation often ignores the object as part of the background or combines it with the background, classifying it as an incorrect object label.(17) In this work, we expect to start from the original image to solve this problem of a fusion background at a fundamental level. Suppose we preprocess the original image. There is a specific gap feature between the object in the image and the background. That is, the object and background cannot be easily distinguished. If the convolutional layer can distinguish both when extracting features, the fusion background problem can be solved. In this section, we introduce image enhancement by calculation and image overlap.

2.2.1 PyNET

The typical purpose of image signal processing (ISP) is to avoid low pixel values and blur. PyNET was developed to solve blurred images(14) by utilizing the pyramid-shaped CNN architecture designed for fine-grained image restoration. It includes ISP strategies such as demosaicing, denoising, white balance, color contrast correction, and color gray adjustment.

The PyNET model is trained sequentially from the lowest fifth level to improve reconstruction at a lower image resolution.(18) The trained model can directly improve the image resolution. The overall PyNET model architecture is shown in Fig. 4.

Levels 4 and 5 deal with images reduced by 8 and 16 times, respectively. Therefore, the two models correct global color and brightness, contrast, and gamma. Because these perceptual losses are not particularly obvious at this scale, the loss function used to train the model depends on the number of corresponding levels of the generated images divided by the level-to-scale proportion. The goal of this training is to minimize the mean square error (MSE).

Levels 2 and 3 deal with images reduced by 2 and 4 times, respectively. They mainly deal with global contextual information. These two-layer models need to consider a variety of semantic information on the image to improve the image quality of the objects by considering various colors, shapes, and attributes. The perception of VGG and the MSE loss function are incorporated into combined training at a ratio of 4 to 1.

The image processed at Level 1 gives the original ratio. After training, local image correction can enhance the object's texture, noise, and local processing color. Each layer model will be trained simultaneously with the lower-level model to ensure a deeper connection between them.

PyNET requires a long learning rate training time, especially at high-resolution levels. In this experiment, the model of each level is trained for 16 epochs. With the dataset of more than 7000 real pictures, the learning rate of each layer starts from 5.0×10^{-5}. The maximum learning rate at the beginning of training using 1500 real pictures is 3.0×10^{-4}. It gradually decays to 1.0×10^{-6} until the end of the training, as shown in Fig. 5.

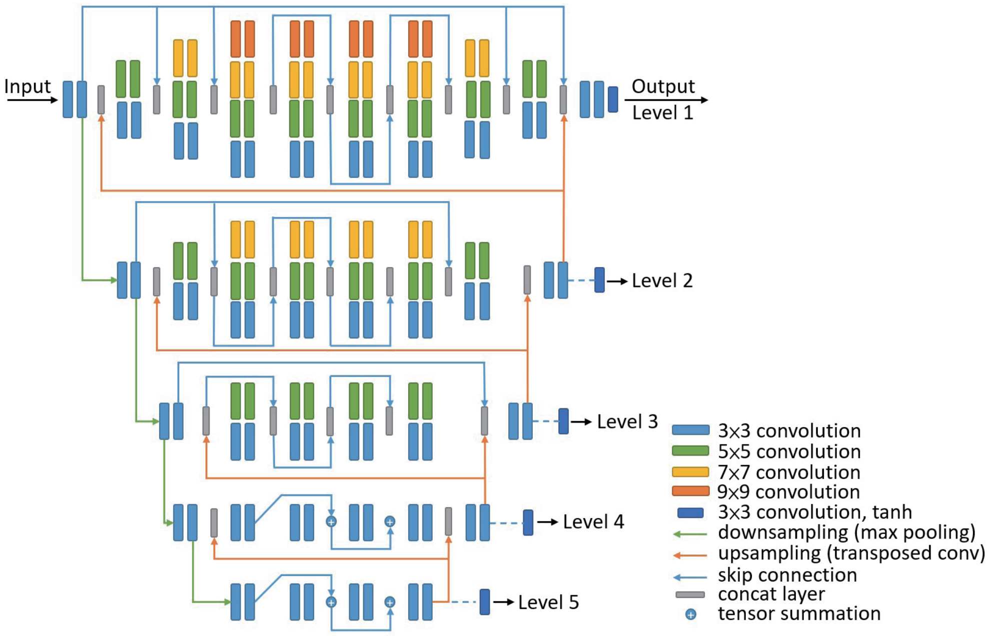


Fig. 4. (Color online) PyNET architecture.[18]

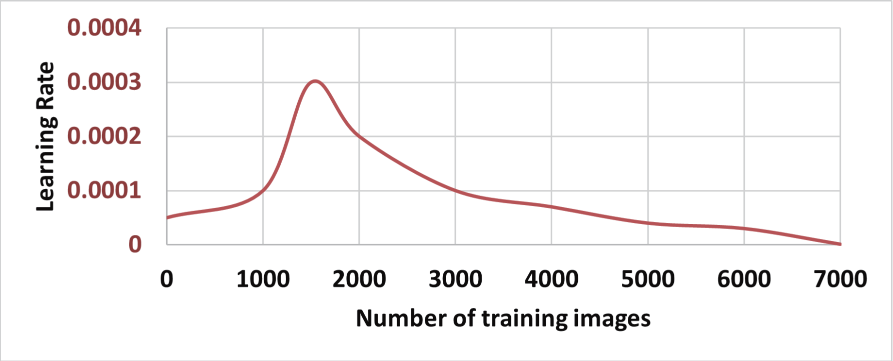


Fig. 5. (Color online) Learning rate for training PyNET.

2.2.2 Image overlap

We use the Level 1 PyNET model to generate images with enhanced features. If PNG files are enhanced by PyNET, the enhanced images will be biased, which is unsuitable for direct image segmentation. Therefore, here, we carry out extra processing on the enhanced image.

As shown in Fig. 6, the enhanced image is not the actual target result, but the generation of enhanced images by the PyNET model cannot be definitively denied. The image is indeed enhanced.

In this case, both the original image color and the enhanced image are combined, and the convolutional layer is used to extract the features of the overlapping image. Such images with similar fusion degrees of the background can be enhanced to a certain extent.

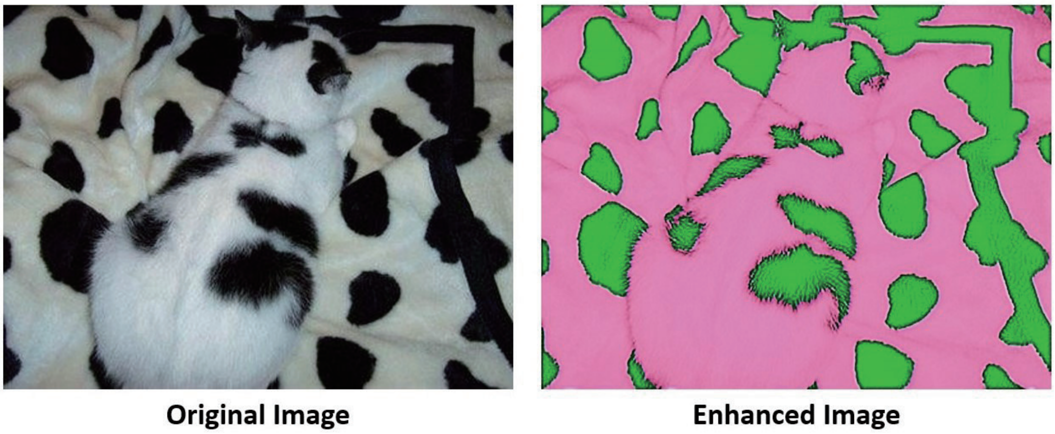


Fig. 6. (Color online) Comparison of the original and enhanced images.

The algorithm for the overlap of the original and enhanced images is shown as

$$overlapping\ (x_i, y_j) = saturated\ (Image1(x_i, y_j) \cdot \alpha + Image2(x_i, y_j) \cdot \beta + \gamma). \tag{2}$$

Here, α is the overlap weight of the original image, β is the overlap weight of the enhanced image, and γ is the pixel bias.

The pixel $(x_i, y_j)^{th}$ of *Image*1 is multiplied by the overlap weight α, then the addition pixel $(x_i, y_j)^{th}$ of *Image*2 is multiplied by the overlap weight β.

Since the sizes of the two images are the same, $(x_i, y_j)^{th}$, the pixel coordinates of the object position will be the same, to which an offset value γ (bias) is finally added.

It is found through experiments that γ will affect the brightness of the overall image. In order to avoid destroying the semantic meaning in the image, we set γ to 0. After the saturated function is calculated, this value may exceed 255 or be less than 0, so this function may need to be adjusted. The formulation is similar to that of the activation function used in the general CNN, as shown below.

$$saturated\left(pixel\right) = \begin{cases} 255 & \text{if } 255 < pixel \\ pixel & \text{if } 0 < pixel < 255 \\ 0 & \text{if } pixel < 0 \end{cases} \tag{3}$$

a. Limit pixel to between 0 and 255 to avoid the difficulty of the subsequent image segmentation.
b. Overlap the images and adjust the overlap weights α and β to produce many pixels with similar colors but without losing the goal of semantic image segmentation.

The algorithm proposed in this paper is aimed at enabling image segmentation in similar fusion background images.

2.3 Image segmentation

Perform the previous procedure, that is, image enhancement and image overlap, and then perform image segmentation. Produce a large number of overlapping images with both original colors and enhanced features. Process the original image using the PyNET model to obtain an enhanced image. Overlap the original and enhanced images in accordance with Eq. (2). After overlapping the images, use the image segmentation model to segment the images and observe whether the image enhancement method can improve similar fusion degree background images. The image segmentation models used in this research are FCN, U-Net, and DeepLab.(19)

2.3.1 Fully convolutional network

We compare the segmentation mask performance, so the experiment is performed using the FCN architecture.(20) In the past, the classified network usually utilized a fully connected layer to convert the original two-dimensional feature map into a one-dimensional fixed-length feature vector. It loses spatial information and finally outputs a specific length vector representing the probability that the input image belongs to each category and uses this as the classification label. This process is called convolutionalization. Since each unit can perform input and output, the whole image is calculated layer by layer instead of batch by batch. Both the forward-propagation and back-propagation calculations are efficient, as shown in Fig. 7.

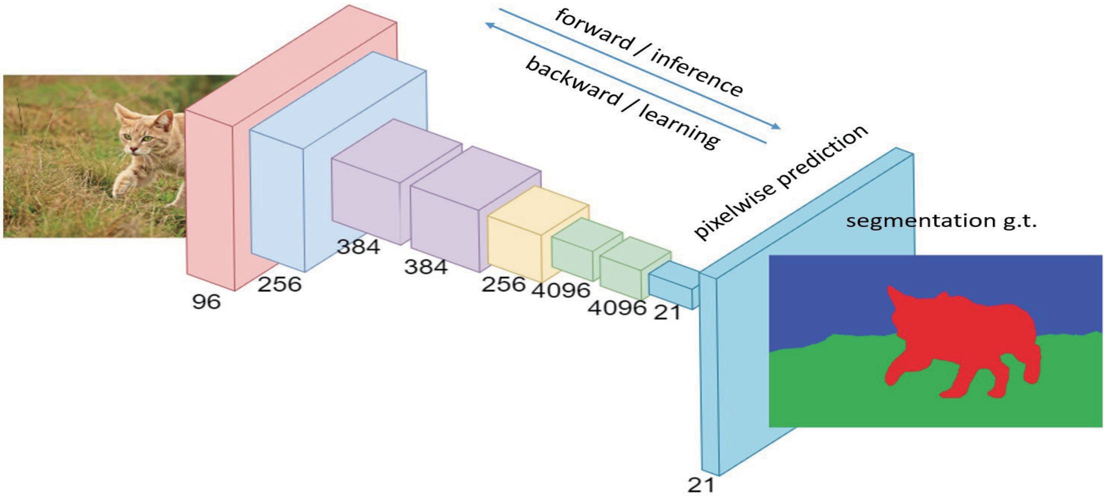


Fig. 7. (Color online) FCN architecture.

2.3.2 U-Net

The construction concept of the U-Net architecture is similar to that of the FCN architecture, as shown in Fig. 8.(17,21) Since the output result is an image, there is no fully connected layer. This algorithm uses a small amount of data for training to obtain accurate segmentation results. The first half of U-Net performs convolution and pooling down-sampling, and the excitation function uses the ReLu function. This process is the extraction of features in the image, while the second half of U-Net performs convolution and up-sampling. The up-sampling method uses deconvolution. The result will be much better than that of the bilinear interpolation method used by the FCN model mentioned above. This improved method solves the dilemma that the original FCN must sacrifice some resolution to obtain more spatial information. U-Net's up-sampling still retains many feature channels, but the disadvantage is that its consumption of computational resources is high. The image segmentation model adopted in this research is constructed on the basis of the U-Net model architectural concept. The selection of the U-Net model architecture can lead to good results with training using only a few datasets. With this model, the Oxford IIIT Pet dataset was used. For all images, the region of interest (ROI) and the entire image segmentation result are considered. This means that images of not only pets can be segmented and that this U-Net architecture can be applied to different types of images. The parameters of the U-Net model formed for this study are presented in Table 2.

After the overlapped image segmentation by this U-Net model, the generated prediction segmentation mask contains two labels: object border and object content. One purpose of this study is to process this segmentation image. First, the two types of labels and the background are

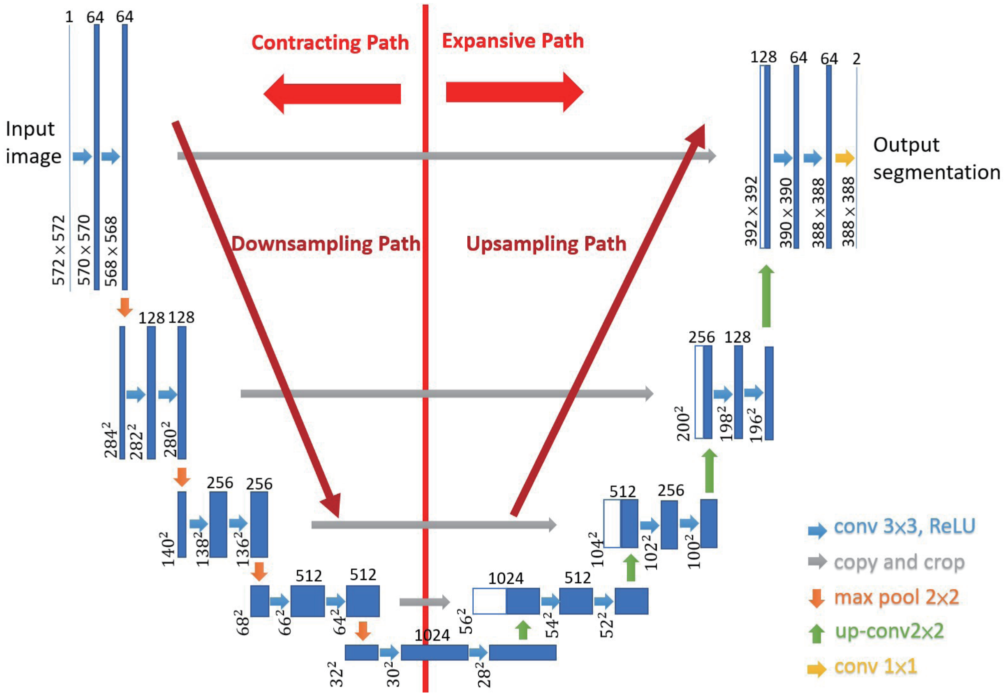


Fig. 8. (Color online) U-Net down-sampling and up-sampling architecture.(21)

Table 2
U-Net model parameter settings.

Input	Kernel	Stride	Operation	Output
128 × 128 × 3	3 × 3	2 × 2	Convolution	64 × 64 × 96
64 × 64 × 96	3 × 3	2 × 2	Convolution	32 × 32 × 144
32 × 32 × 144	3 × 3	2 × 2	Convolution	16 × 16 × 192
16 × 16 × 192	3 × 3	2 × 2	Convolution	8 × 88576
8 × 8 × 576	3 × 3	2 × 2	Convolution	4 × 4 × 420
4 × 4 × 320	3 × 3	2 × 2	Up-sampling	8 × 8 × 512
8 × 8 × 512, 8 × 8 × 8576	—	—	Concatenation	8 × 8 × 1088
8 × 8 × 108	3 × 3	2 × 2	Up-sampling	16 × 16 × 256
16 × 16 × 256, 16 × 16 × 192	—	—	Concatenation	16 × 16 × 448
16 × 16 × 448	3 × 3	2 × 2	Up-sampling	32 × 32 × 128
32 × 32 × 128, 32 × 32 × 144	—	—	Concatenation	32 × 32272
32 × 32 × 272	3 × 3	2 × 2	Up-sampling	64 × 64 × 64
64 × 64 × 64, 64 × 64 × 96	—	—	Concatenation	64 × 64 × 160
64 × 64 × 160	3 × 3	2 × 2	Up-sampling	128 × 128 × 3

Concatenation: An essential operation of U-Net is concatenation to combine the downward path with the upward path. In this way, the net can learn classification and positioning by an end-to-end training method.

binarized, leaving only the object's overall shape. Then, the mask appearance is fine-tuned by corrosion and expansion to remove some misjudgments caused by noise. In this way, the processed mask can be placed in the original image to complete the background removal process. Next, the background interference is reduced and the objects in the image are more clearly highlighted. This process cleverly eliminates images with similar fusion backgrounds.

2.3.3 DeepLab

Among the image segmentation models, the semantic segmentation results often have some problems. There are two main reasons. The first half of the model is for feature extraction where extensive pooling is continuously performed. This process inevitably leads to the loss of a large amount of spatial information. The second reason is that the label is not sensitive to space, and the predicted pixel label is not processed further.

For DeepLab (from DeepLabv1 to DeepLav2, DeepLabv3, and DeepLabv3+), targeted improvements have been proposed for these two points.(22)

DeepLabv3+ is the latest architecture of the DeepLab series.(23) It is the commonly used encoder/decoder architecture for semantic segmentation. When using DeepLabv3+ as an encoder, adding a simple and effective decoder module can improve the segmentation effect at the object's edge. The overall architecture is shown in Fig. 9.

2.3.4 Test dataset

It is impossible to obtain many image sets that meet the background type of similar fusion degree from the images obtained from a regular driving vehicle. The images in this study are mainly based on the Oxford IIIT Pet dataset. They are suitable for the segmentation of images with similar fusion background images. The test dataset used in this study contains three categories, and each category may cause similar fusion background problems.

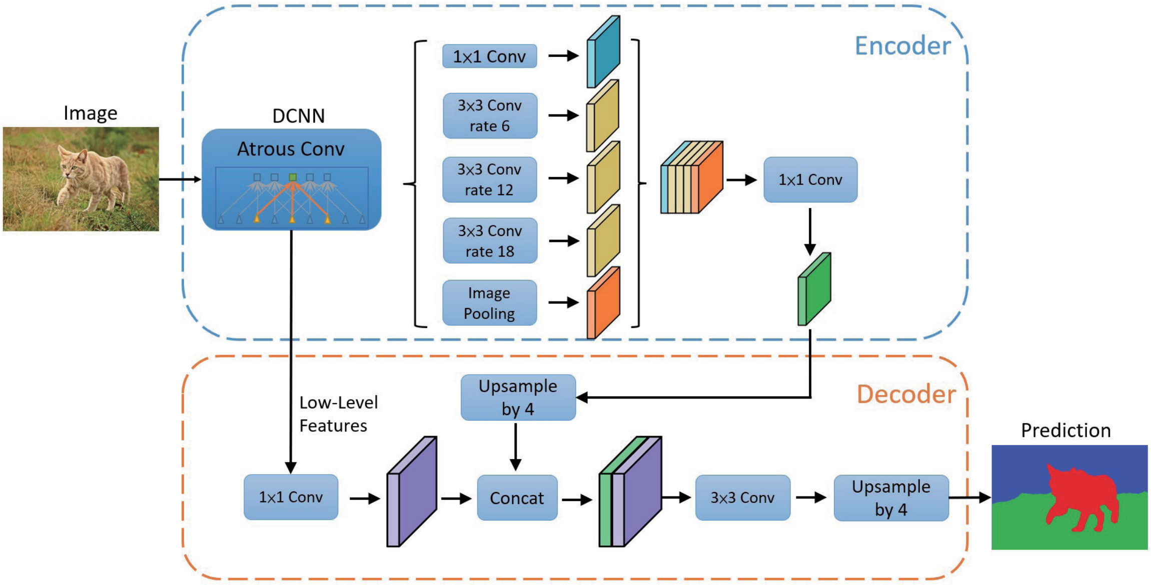


Fig. 9. (Color online) DeepLabv3+ architecture.

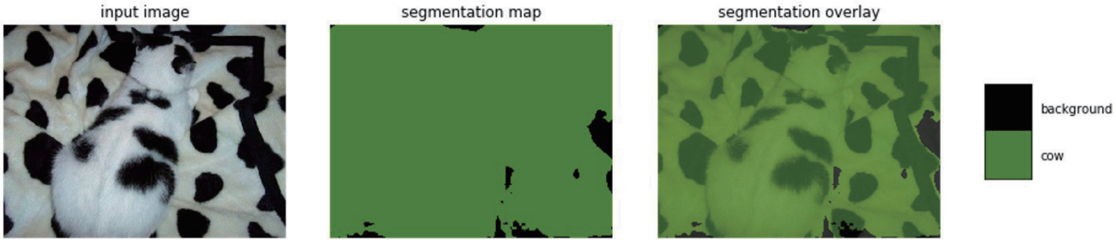


Fig. 10. (Color online) Segmentation of image with similar fusion degree background.

i. Crypsis

Crypsis refers to the color and texture of the object in the image being similar to those of the background. In traditional image segmentation models, such objects in the image are usually merged with the background during the convolution process, resulting in an incorrect final segmentation result. Even the subsequent semantic labels will be erroneous.

As shown in Fig. 10, the DeepLab model ignores the cat in the image, merges the cat and the background together, and identifies the object as a cow.

ii. Camouflage

Since camouflage will blur the border between the object and the background, we also make images with similar fusion backgrounds one of our research targets.

iii. Mimicry

Mimicry refers to the characteristic that a particular species acquires similar characteristics to another species through evolution. There are many strategies for organisms to have hidden effects, such as a body color identical to the background color of the habitat. Images with such specious effects can also cause image segmentation errors. For example, the appearance of a stick insect is just like a branch. In image segmentation, it is difficult to separate the stick insect from the category of its vicinity. It may also lead to the misclassification of semantic labels.

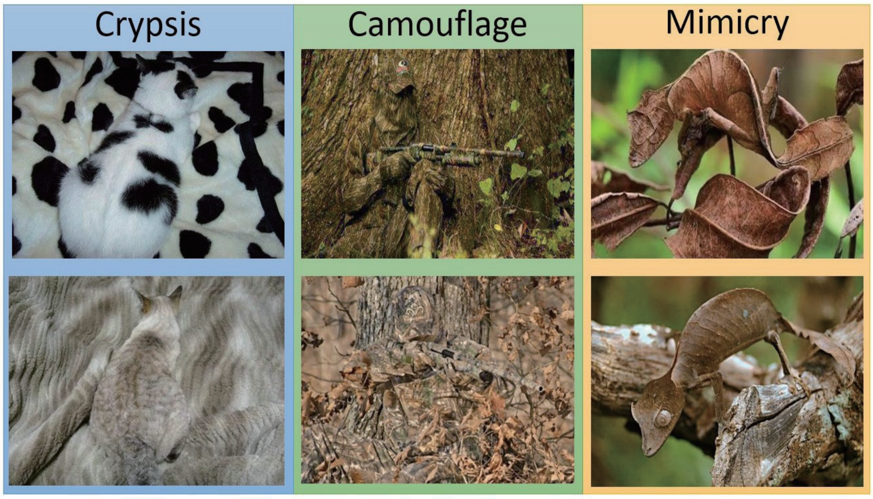


Fig. 11. (Color online) Test dataset.

Figure 11 shows the test dataset images of the above three categories.

We used the image overlap method mentioned in Sect. 2.2.2. Refer to Eq. (2). for its operational method. By adjusting the original image overlap weight α and the enhanced image overlap weight β, the first image of the protective color can be expanded to 1000 images. For other images, only the enhanced image overlap weight β is adjusted and expanded to 100 images. There are 1500 images in the total dataset for the image segmentation test. Each test image will be cut manually to obtain the object's actual shape as reference. After the model segmentation of the original image, the IoU scores of the prediction and reference results will be calculated as the evaluation standards. In addition, these α and β values are not fixed and can depend on the test dataset and the time available for calculation.

3. Experiments and Analysis

In this section, we introduce the experimental design and the expected results of image segmentation and segmentation recognition experiments. Then, we analyze and discuss the experimental results.

3.1 Experimental method

3.1.1 Image segmentation

The first step is to generate an enhanced image, as shown in Fig. 12. The original image uses PyNET to enhance its features. Then, the original and enhanced images are overlapped, as shown in Fig. 13. A large number of background images with similar fusion degrees are generated using different overlap weights. Finally, these images are input to the U-Net, DeepLab, and FCN models and segmented. The scores obtained from the images with different overlapping weights are compared.[24,25]

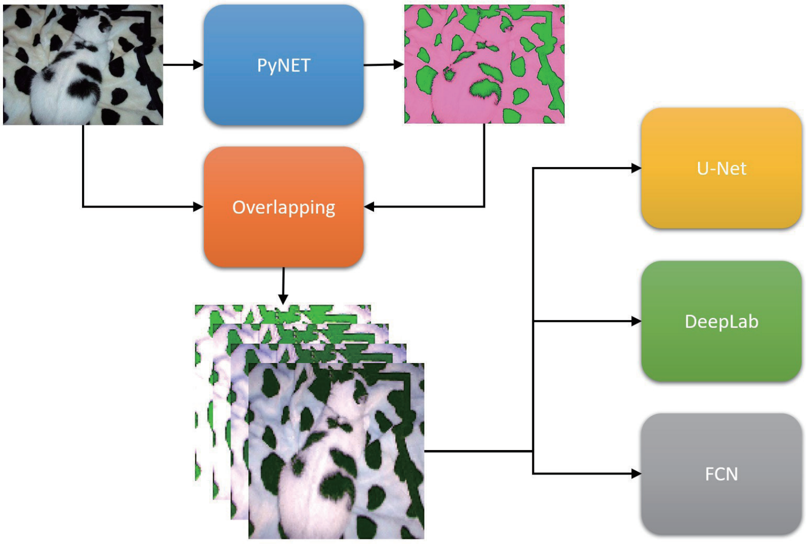


Fig. 12. (Color online) Flow of experiment 1.

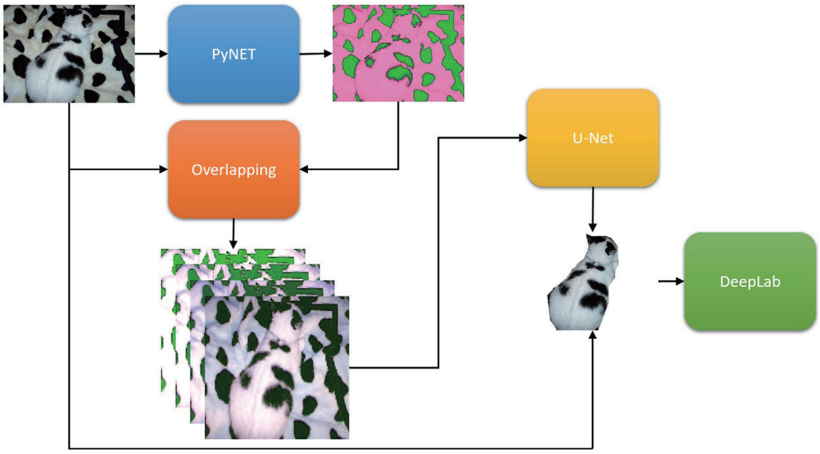


Fig. 13. (Color online) Flow of experimental 2.

3.1.2 Segmentation and recognition

The mask generated by the algorithm designed in this study is applied to the original image. DeepLab is then used for image segmentation to test whether the results obtained with the algorithm can be expected to improve the segmentation and classification results of similar fusion background images.

The features of the original image are enhanced by PyNET and then the original image is overlapped with the enhanced image. A large number of plausible background images with similar fusion degrees are generated using different overlapping weights. The difference from the segmentation image experiment is that the result generated by U-Net is trimmed and overlaid on the original image. The purpose is to cut out the necessary background to reduce the impact of noise on image segmentation. Finally, the DeepLab image segmentation result is obtained.

3.2 Results of experiments

3.2.1 Image segmentation

Below we present the results of the image segmentation comparison experiment. The following three image segmentation models are used: U-Net, DeepLab, and FCN. These segmentation images are generated by adjusting the overlap weights of the original and enhanced images. The resulting image has a fixed overlap weight of the original image and only the overlap weight of the enhanced image is adjusted. The results are plotted in Figs. 14–18, with the overlapping weight on the horizontal axis and the IoU score on the vertical axis.

We used six original images of similar fusion backgrounds in the experiment, two each for crypsis, camouflage, and mimicry. The results are explained as follows. An image with a similar fusion degree background, crypsis image 1, is shown in Fig. 19.

The results of experiment 1 in Figs. 14–18 show that the image segmentation results of U-Net are significantly better than those of DeepLab and FCN. From the experimental results in Fig.

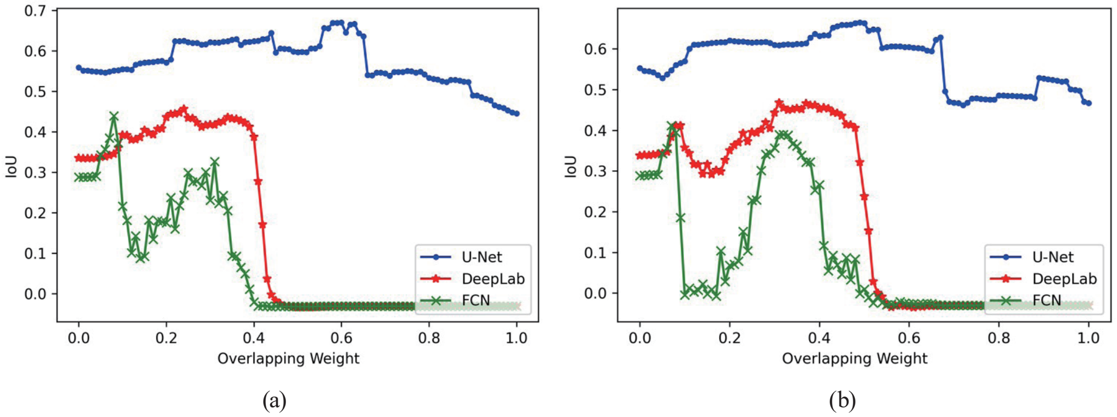


Fig. 14. (Color online) Experiment 1: IoU scores of overlapped image when original image overlapping weights (α) were (a) 1 and (b) 0.9.

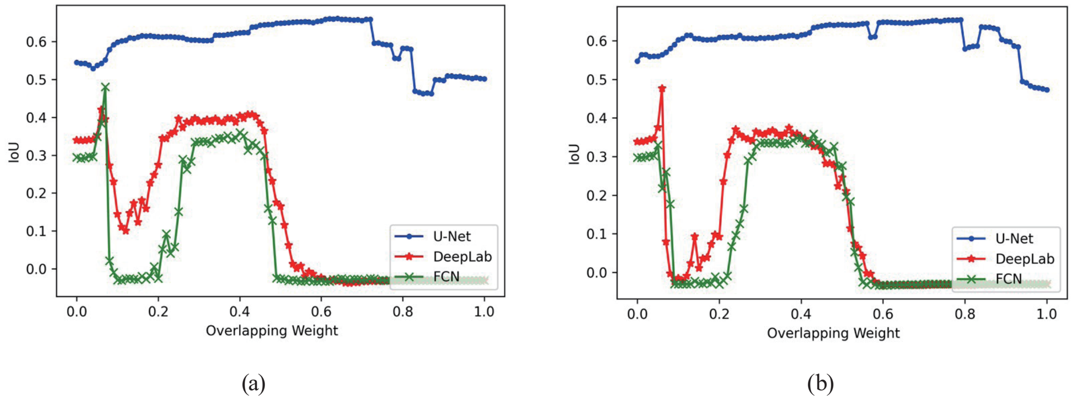
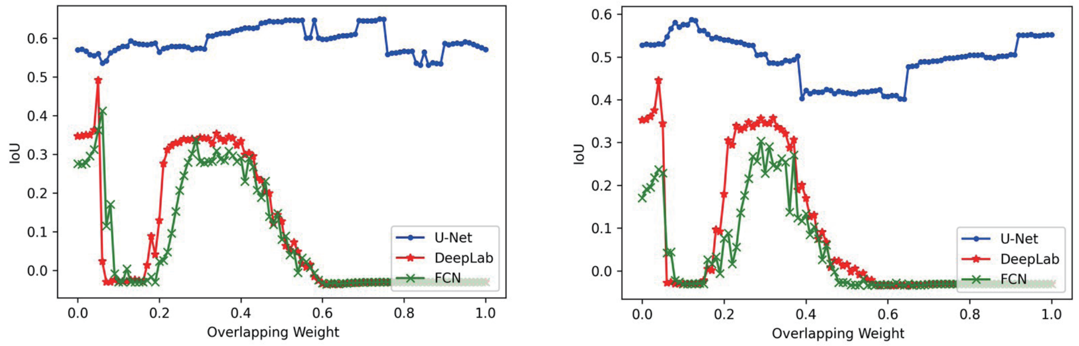
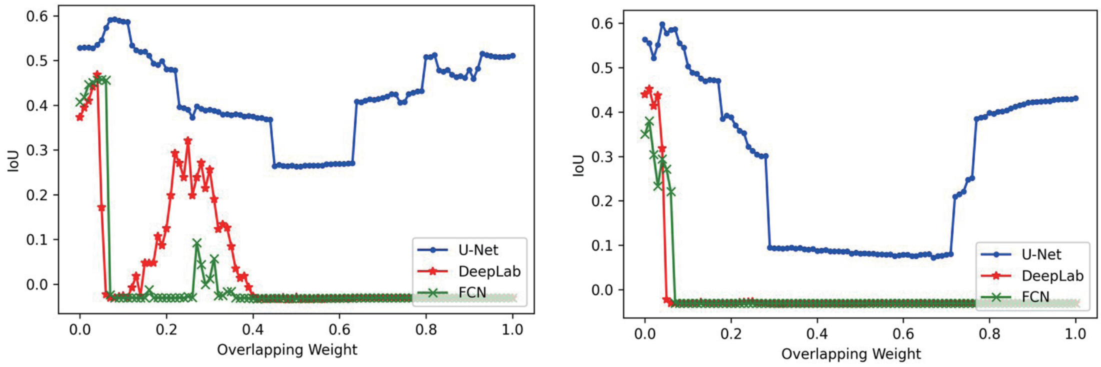


Fig. 15. (Color online) Experiment 1: IoU scores of overlapped image when original image overlapping weights (α) were (a) 0.8 and (b) 0.7.



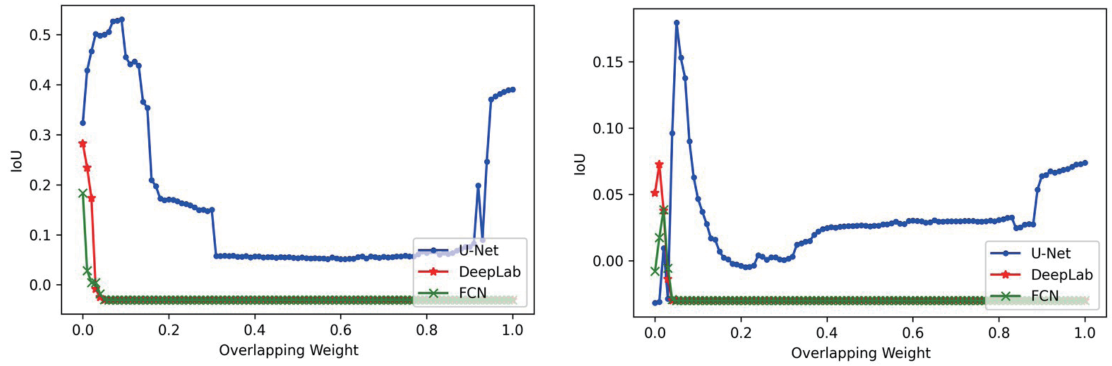
(a) (b)

Fig. 16. (Color online) Experiment 1: IoU scores of overlapped image when original image overlapping weights (α) were (a) 0.6 and (b) 0.5.



(a) (b)

Fig. 17. (Color online) Experiment 1: IoU scores of overlapped image when original image overlapping weights (α) were (a) 0.4 and (b) 0.3.



(a) (b)

Fig. 18. (Color online) Experiment 1: IoU scores of overlapped image when original image overlapping weights (α) were (a) 0.2 and (b) 0.1.

18(b), it is seen that with the original image overlap weight of 0.1, the IoU score curves of U-Net, DeepLab, and FCN cross. Both DeepLab and FCN scores are the lowest. This result indicates that the all-image segmentation effect of this image is poor. FCN limits the size of the sensing area to that of the pixel block. In this way, only some local features can be extracted, which

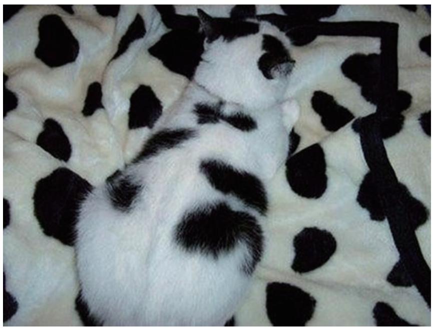


Fig. 19. (Color online) Crypsis image 1 used in experiment 1.

limits the performance of classification. DeepLab needs a more extensive and more profound deep CNN to achieve good segmentation performance.

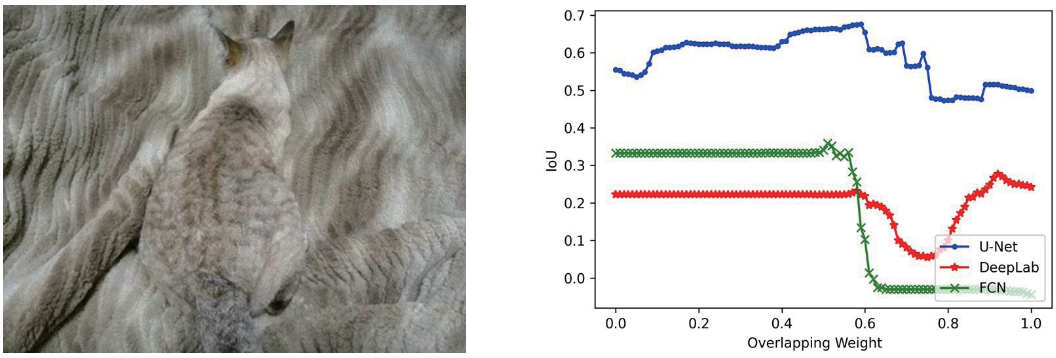
From Table 3, we see that when the original image overlap weight (original weight) is 1, the U-Net enhanced image overlap weight is 0.6 and the obtained IoU score of 0.6696 is the highest. When retaining many original image features, the IoU score will be higher when the enhanced image overlap weight is between 0.5 and 0.8. As the original image overlap weight decreases, the IoU score will also become lower. We conclude that it is preferable to retain the color characteristics of the original image for overlapping images to yield a better segmentation result.

Another image with a similar fusion degree background, crypsis image 2, is shown in Fig. 20(a). The original image overlap weight is constant at 1, and only the enhanced image overlap weight is adjusted. The IoU score after image segmentation model cutting is shown in Fig. 20(b). As the overlap weight of the original image decreases, the IoU score becomes lower. The result shows that the color characteristics of the original image are preserved in the overlapped image. This will lead to better segmentation results. From the results of using different original overlap weights, the U-Net model is found to have the highest IoU score, which is much higher than those of the DeepLab and FCN models.

Figure 21(a) shows camouflage image 1. The original image overlap weight is constant at 1, and only the enhanced image overlap weight is adjusted. The IoU scores obtained after image segmentation model cutting are shown in Fig. 21(b). The experimental results show that in terms of IoU score, the performance of the U-Net model far exceeds that of the DeepLab model.

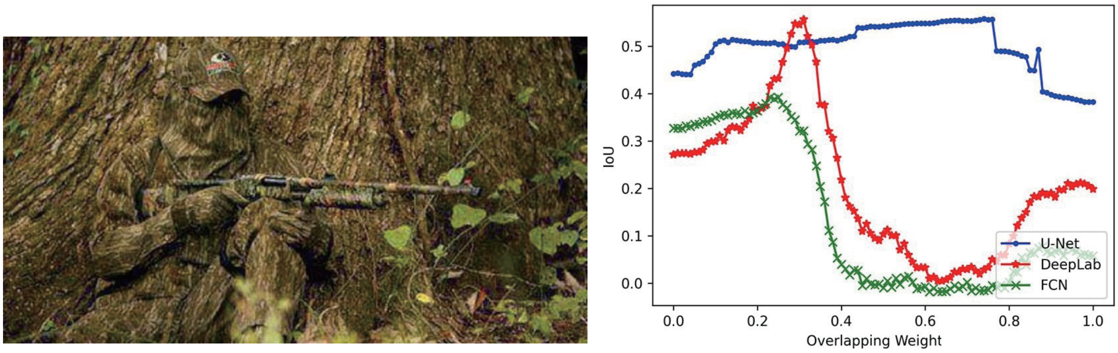
Table 3
Performance of image segmentation models as indicated by IoU score.

Original weight	IoU	U-Net overlapping	IoU	DeepLab overlapping	IoU	FCN overlapping
1	0.6696	0.60	0.4566	0.24	0.4391	0.08
0.9	0.6647	0.49	0.4674	0.31	0.4115	0.07
0.8	0.6607	0.64	0.4197	0.06	0.4800	0.07
0.7	0.6545	0.79	0.4768	0.06	0.3586	0.43
0.6	0.6496	0.74	0.4911	0.05	0.4122	0.06
0.5	0.5873	0.12	0.4451	0.04	0.3037	0.29
0.4	0.5920	0.08	0.4681	0.04	0.4579	0.05
0.3	0.5974	0.04	0.4517	0.01	0.3796	0.01
0.2	0.5308	0.09	0.2882	0.00	0.1832	0.00
0.1	0.1796	0.05	0.0726	0.01	0.0385	0.02



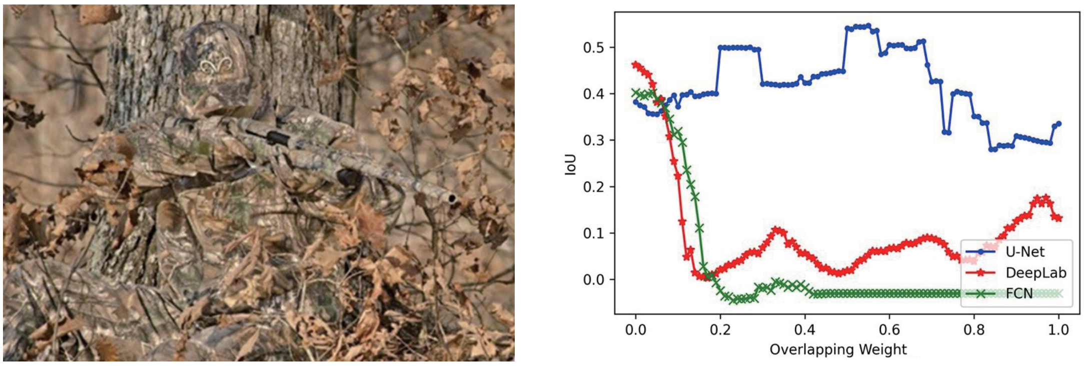
(a) (b)

Fig. 20. (Color online) (a) Crypsis image 2 used in experiment 1. (b) IoU scores for crypsis image 2 in experiment 1.



(a) (b)

Fig. 21. (Color online) (a) Camouflage image 1 used in experiment 1. (b) IoU scores for camouflage image 1 in experiment 1



(a) (b)

Fig. 22. (Color online) (a) Camouflage image 2 used in experiment 1. (b) IoU scores for camouflage image 2 in experiment 1.

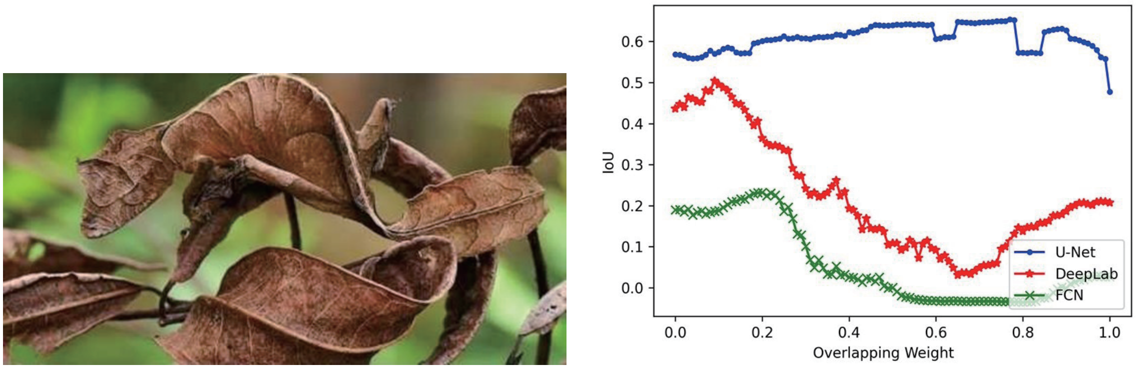
Figure 22(a) shows camouflage image 2. The original image overlap weight is constant at 1, and only the enhanced image overlap weight is adjusted. IoU scores after image segmentation model cutting are shown in Fig. 22(b). It can be seen that when the enhanced images do not overlap in the initial stage, the DeepLab and FCN models perform better than the U-Net model.

However, as the overlap weight of the enhanced image increases, the IoU score of the U-Net model is generally higher than those of the DeepLab and FCN models.

Figure 23(a) shows mimicry image 1. The original image overlap weight is constant at 1, and only the enhanced image overlap weight is adjusted. The IoU scores obtained after image segmentation model cutting are shown in Fig. 23(b). It can be seen that the IoU scores of the U-Net model exceed those of the DeepLab and FCN models. The difficulty in the segmentation of this image is that the lizard's tail is easily overlooked, causing the score to decrease rapidly.

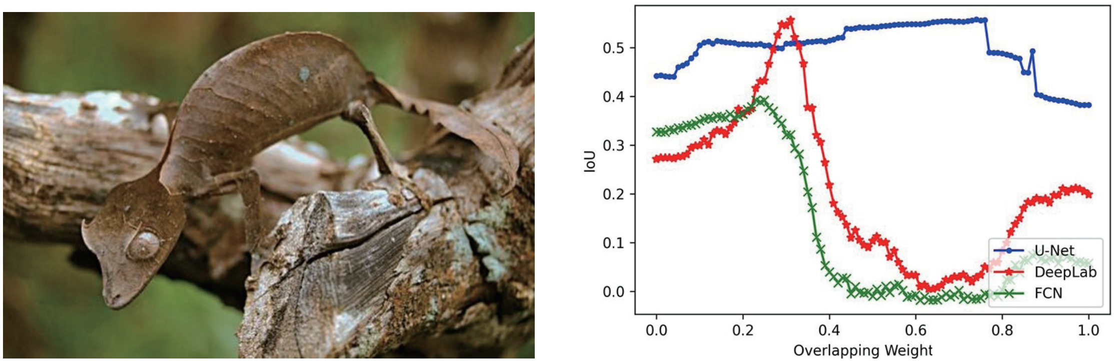
Figure 24(a) shows mimicry image 2. The original image overlap weight is constant at 1, and only the enhanced image overlap weight is adjusted. The IoU scores obtained after image segmentation model cutting are shown in Fig. 24(b). It can be seen that the score curves cross each other multiple times. Therefore, the overlap weight of the enhanced image affects the result of the complete image segmentation.

For the DeepLab model, the IoU score–enhanced image overlap weight curve crosses those of the other models multiple times. The U-Net model obtains the highest IoU score of 0.5579 when the enhanced image overlap weight is 0.74, similar to the case of a weight of 0.77. It may be that the two images have a high degree of similarity.



(a) (b)

Fig. 23. (Color online) (a) Mimicry image 1 used in experiment 1. (b) IoU scores for mimicry image 1 in experiment 1.



(a) (b)

Fig. 24. (a) (Color online) Mimicry image 2 used in experiment 1. (b) IoU scores for mimicry image 2 in experiment 1.

Table 4
Best IoU scores of each model for each type of image.

Original	IoU	U-Net overlapping	IoU	DeepLab overlapping	IoU	FCN overlapping
Crypsis-1	0.6696	0.60	0.4566	0.24	0.4391	0.08
Crypsis-2	0.6762	0.59	0.2768	0.92	0.3586	0.51
Camouflage-1	0.7650	0.60	0.4253	0.03	0.3443	0.02
Camouflage-2	0.5464	0.55	0.4615	0.00	0.4023	0.00
Mimicry-1	0.6537	0.77	0.5043	0.09	0.2323	0.20
Mimicry-2	0.5579	0.74	0.5564	0.31	0.3923	0.23

Unlike mimicry image 1, this lizard’s head is awkward to process because of the light and shadow effect and is easily lost when distinguishing the boundary. Another difficulty in segmentation is that the lizard’s tail is easily overlooked, resulting in a rapid decline in score.

Table 4 shows the best IoU scores of each model for each type of image, where only the enhanced image overlap weight was adjusted while keeping the original image overlap weight constant at 1.

3.2.2 Segmentation and recognition

The test results of the image segmentation model presented in Sect. 3.2.1 are used to train the U-Net model to process the enhanced overlapping images to obtain the highest IoU score. Therefore, in this experiment 2, we continue to use the best segmentation result of the U-Net model obtained from experiment 1, then overlay the mask predicted by the model on the original image and perform image segmentation and recognition again with DeepLab. We use segmentation-twice methods to significantly reduce background interference in images with similar fusion backgrounds, improving the image segmentation results.

In our work, the image segmentation results are divided into four parts.

- Input image: The input image is obtained by superposing the U-Net prediction mask onto the original image.
- Segmentation map: The segmentation map is the prediction mask used by DeepLab to segment the input image.
- Segmentation overlay: This is the result of overlaying the segmentation map on the input image and is used to evaluate the segmentation effect.
- Semantic label classified by DeepLab, shown on the far right of Figs. 25–30.

The image segmentation results for crypsis images 1 and 2 are shown in Figs. 25 and 26, respectively. As seen in Figs. 25 and 26, after the segmentation process using the U-Net model, the background interference in the image is reduced. The enhanced image method can indeed improve the segmentation of images with similar fusion backgrounds.

The image segmentation results for camouflage images 1 and 2 are shown in Figs. 27 and 28, respectively. As seen in Figs. 27 and 28, the U-Net model cuts out the rough outline of a person. However, the DeepLab model cannot perform image segmentation for camouflage images because the object pixels are still too complex.

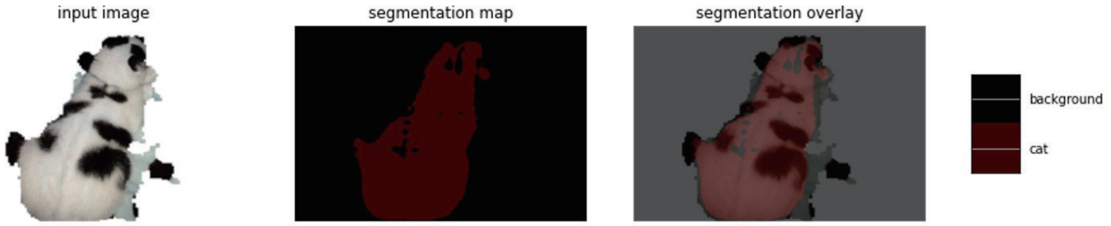


Fig. 25. (Color online) Results of experiment 2 using crypsis image 1.

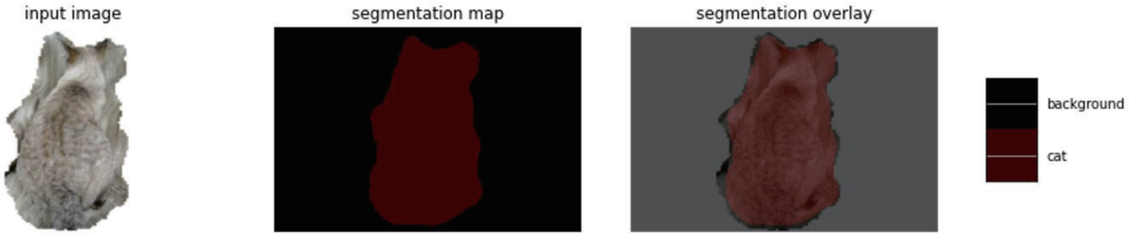


Fig. 26. (Color online) Results of experiment 2 using crypsis image 2.

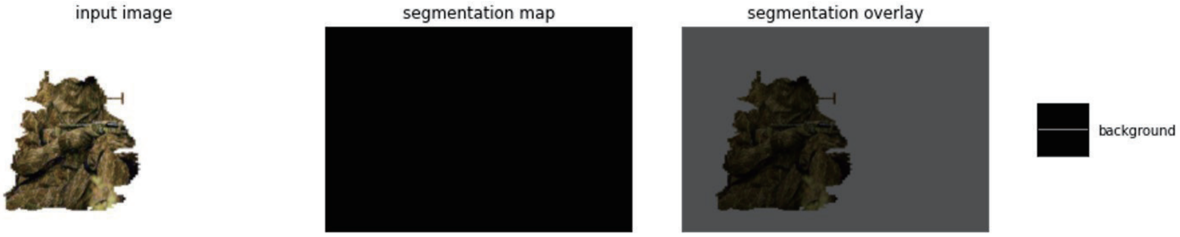


Fig. 27. (Color online) Results of experiment 2 using camouflage image 1.

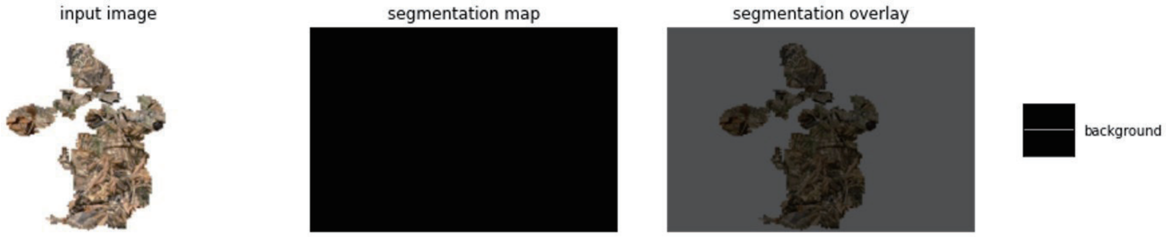


Fig. 28. (Color online) Results of experiment 2 using camouflage image 2.

The image segmentation results for mimicry images 1 and 2 are shown in Figs. 29 and 30, respectively. As seen in Figs. 29 and 30, although the DeepLab model has segmented rough object outlines, the semantic label classifications are incorrect.

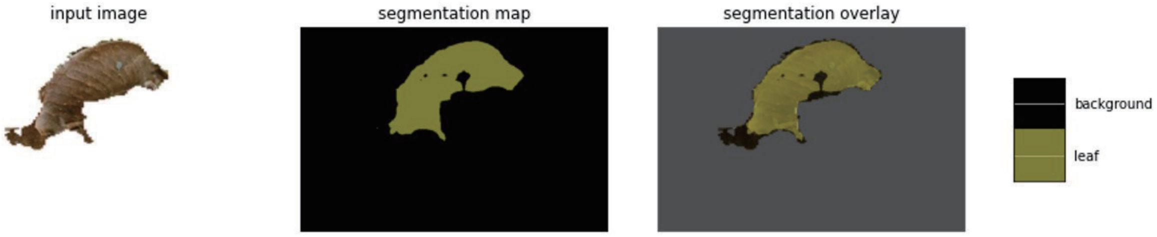


Fig. 29. (Color online) Results of experiment 2 using mimicry image 1.

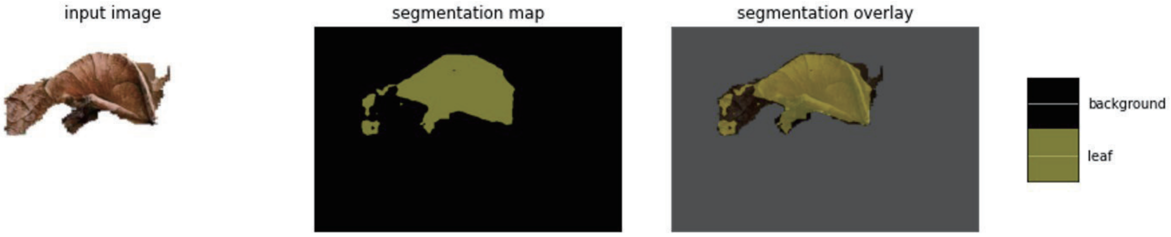


Fig. 30. (Color online) Results of experiment 2 using mimicry image 2.

4. Conclusions

Self-driving vehicles have become a manifestation of the development of modern science and technology. We used lens-related sensors commonly utilized in automobiles to perceive the road environment. Then, path planning and object image segmentation decisions were performed, which are essential to improving autonomous driving safety and reliability.

The primary purpose of our work is to achieve the segmentation of images with similar fusion backgrounds using a CNN hybrid model. Although most of the experiments are based on simulation, we have successfully strengthened the segmentation and classification of objects and backgrounds using mainstream image segmentation models. We utilized a multivariate method to verify the result. Therefore, our proposed method is not limited to the most commonly used data for autonomous driving and can be applied to improve the similarity and fusion of background images.

- Strengthen the similarity and fusion degree of background image features and reduce background interference on object features.
- In mainstream image segmentation models, feature extraction is implemented using a convolutional layer. The convolutional layer design is unable to handle images with objects that are similar to background features. In maximum pooling, the object will be ignored because the object and background appear the same, which leads to an error in image segmentation. Therefore, if the features of the original image are enhanced so that the object and background features are significantly different, the convolutional layer can extract different features. To this end, adjust the overlap weights of the original and enhanced images.

The enhanced image may not be suitable for direct image segmentation because essential features such as the color and texture of the original image may have been destroyed. In order to avoid this problem, we overlap the original and enhanced images so that the overlapped image retains the characteristics of both the original and enhanced images at the same time. In this way, the problem of similar fusion backgrounds is alleviated. Moreover, by adjusting the overlap weight, it is also possible to amplify a sparse number of background images with similar fusion degrees and test the model more extensively. We found that when the original image overlap weight is maintained at 1 and the enhanced image overlap weight is set between 0.5 and 0.8, improved image segmentation results can be obtained.

- Use the U-Net model for the segmentation of overlapping images.

As mentioned above, mainstream image segmentation models are unsuitable for the segmentation of images with similar fusion backgrounds. The same may be said of the segmentation of enhanced overlapping images. The experimental results showed that the U-Net model trained in this research has good IoU score in the segmentation of enhanced images. In order to obtain more scale features, the DeepLab model uses atrous convolution, which is not suitable for the segmentation of overlapping images. As a result, the overall IoU score of the DeepLab model is lower than that of the U-Net architecture. Image segmentation by the FCN model is still relatively rough and the overall performance is poor.

- The segmentation results re-do image segmentation.

The accuracy of image segmentation results can be improved by overlaying the segmentation results of the U-Net model on the original image to greatly reduce the interference caused by the background. Since the effect is equivalent to the preprocessing of the original image, it can be applied to all image segmentation models.

- Improved segmentation results of crypsis color images

The method proposed in this study can better process the similarity and fusion of the background image of the crypsis color type, improving the image segmentation effect. The experimental results show that the DeepLab model can correctly segment the contours and classify the semantic labels of images with objects having protective coloring.

Despite there being many image segmentation methods that can yield better results when processing experimental data, when dealing with complex and dynamic images, there are still some difficulties. In this section, we have summarized the existing challenges and future development trends in the field of vehicle image segmentation.

Although the U-Net model used in this study has achieved good results in segmenting overlapping images, the safety requirements for self-driving car applications are endless. Relevant research on improving the safety of self-driving cars must be continued.

References

1 B. W. Abegaz and N. Shah: 2020 11th IEEE Annu. Ubiquitous Computing, Electronics & Mobile Communication Conf. (UEMCON) (2020) 0486. https://doi.org/10.1109/UEMCON51285.2020.9298141

2 R. Battrawy, R. Schuster, O. Wasenmüller, Q. Rao, and D. Stricker: 2019 IEEE/RSJ IEEE Int. Conf. Intell. Robots Syst. (IROS) (2019) 7762. https://doi.org/10.1109/IROS40897.2019.8967739

3 Y. Ping, J. Xinwei, and M. Yichao: 2008 IEEE Int. Conf. Serv. Oper. Logist. Inform. SOLI (2008) 1915. https://doi.org/10.1109/SOLI.2008.4682844

4 G. Sun and F. Zhang: IEEE Access **8** (2020) 117080. https://doi.org/10.1109/ACCESS.2020.3004860

5 D. Kenjic, F. Baba, D. Samardzija, and Z. Kaprocki: 2019 IEEE 9th Int. Conf. Consum. Electron. (ICCE-Berlin) (2019) 420. https://doi.org/10.1109/ICCE-Berlin47944.2019.8966136

6 Y. Liu, M. Cheng, X. Hu, J. Bian, L. Zhang, X. Bai, and J. Tang: IEEE Trans. Pattern Anal. Mach. Intell. **41** (2019) 1939. https://doi.org/10.1109/TPAMI.2018.2878849

7 L. Khelifi and M. Mignotte: 2017 IEEE Int. Conf. Image Processing (ICIP) (2017) 3080. https://doi.org/10.1109/ICIP.2017.8296849

8 G. Muslu and B. Bolat: 2019 Scientific Meeting on Electrical-Electronics & Biomedical Engineering and Computer Science (EBBT) (2019) 1. https://doi.org/10.1109/EBBT.2019.8741541

9 W. T. Chiu, C. H. Lin, C. L. Jhu, C. Lin, Y. C. Chen, M. J. Huang, and W. M. Liu: 2020 Int. Computer Symp. (ICS) (2020) 535. https://doi.org/10.1109/ICS51289.2020.00110

10 J. Redmon, S. Divvala, R. Girshick, and A. Farhadi: 2016 IEEE Conf. Computer Vision and Pattern Recognition (CVPR) (2016) 779. https://doi.org/10.1109/CVPR.2016.91

11 K. Simonyan and A. Zisserman: arXiv preprint arXiv:1409.1556 (2014). https://arxiv.org/abs/1409.1556

12 K. He, X. Zhang, S. Ren, and J. Sun: Proc. IEEE Conf. Computer Vision and Pattern Recognition (2016) 770. https://arxiv.org/abs/1512.03385

13 F. Chollet: 2017 IEEE Conf. Computer Vision and Pattern Recognition (CVPR) (2017) 1800. https://doi.org/10.1109/CVPR.2017.195

14 X. Yu, Z. Yu, and S. Ramalingam: 2018 IEEE/CVF Conf. Computer Vision and Pattern Recognition (2018) 4432. https://doi.org/10.1109/CVPR.2018.00466

15 A. Demir, F. Yilmaz, and O. Kose: 2019 Medical Technologies Congr. (TIPTEKNO) (2019) 1. https://doi.org/10.1109/TIPTEKNO47231.2019.8972045

16 X. Tian and C. Chen: 2019 IEEE 2nd Int. Conf. Information Communication and Signal Processing (ICICSP) (2019) 34. https://doi.org/10.1109/ICICSP48821.2019.8958555

17 S. K. Panguluri and L. Mohan: 2021 Int. Conf. Computer Communication and Informatics (ICCCI) (2021) 1. https://doi.org/10.1109/ICCCI50826.2021.9402531

18 A. Ignatov, L. V. Gool, and R. Timofte: 2020 IEEE/CVF Conf. Computer Vision and Pattern Recognition Workshops (CVPRW) (2020) 2275. https://doi.org/10.1109/CVPRW50498.2020.00276

19 T. Carneiro, R. V. M. D. NóBrega, T. Nepomuceno, G. Bian, V. H. C. D. Albuquerque, and P. P. R. Filho: IEEE Access **6** (2018) 61677. https://doi.org/10.1109/ACCESS.2018.2874767

20 C. Peng, Y. Li, L. Jiao, Y. Chen, and R. Shang: IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. **12** (2019) 2612. https://doi.org/10.1109/JSTARS.2019.2906387

21 X. Gao and L. Fang: 2020 39th Chinese Control Conf. (CCC) (2020) 7090. https://doi.org/10.23919/CCC50068.2020.9188804

22 W. Xiao, L. Chang, and W. Liu: 2018 IEEE Int. Conf. Consumer Electronics-Taiwan (ICCE-TW) (2018) 1. https://doi.org/10.1109/ICCE-China.2018.8448568

23 S. C. Yurtkulu, Y. H. Şahin, and G. Unal: 2019 27th Signal Processing and Communications Applications Conf. (SIU) (2019) 1. https://doi.org/10.1109/SIU.2019.8806244

24 P. Isola, J. Zhu, T. Zhou, and A. A. Efros: 2017 IEEE Conf. Computer Vision and Pattern Recognition (CVPR) (2017) 5967. https://doi.org/10.1109/CVPR.2017.632

25 G. Huang, Z. Liu, L. V. D. Maaten, and K. Q. Weinberger: 2017 IEEE Conf. Computer Vision and Pattern Recognition (CVPR) (2017) 2261. https://doi.org/10.1109/CVPR.2017.243