ISSN 1330-3651 (Print), ISSN 1848-6339 (Online)

https://doi.org/10.17559/TV-20231023001051
Original scientific paper

Advancing UAV Image Semantic Segmentation with an Improved Multiscale Diffusion Model

Wang WANG*, Chen ZHOU, Hua HE, Changsong MA

Abstract: This study explores the challenges of image semantic segmentation in autonomous driving across varied campus environments. We introduce a specialized dataset consisting of 400 drone-captured images from different campuses. These images have been meticulously labelled into five categories: buildings, vegetation, ground, playgrounds, and lakes. These categories are essential for precise semantic segmentation tasks which are crucial to autonomous driving applications. To address the segmentation challenges presented by the unique and diverse features of campus environments, we propose an innovative algorithm. This algorithm is based on an enhanced diffusion model that is adept at handling multi-scale features inherent in campus environments. By incorporating scalable jump-connection layers in the denoising probability diffusion model, the proposed algorithm not only achieves superior accuracy but also demonstrates a significant improvement in recognition precision within the dataset, resulting in an average *mIoU* of 85%. The results underscore the algorithm's effectiveness and its potential as a robust solution for semantic segmentation tasks in autonomous driving within campus settings, paving the way for further research and application in real-world scenarios.

Keywords: image semantic segmentation; multiscale diffusion model style; reviewing; UVA

1 INTRODUCTION

In recent years, with the rapid development of artificial intelligence and deep learning technologies, autonomous driving has transformed from a science fiction concept to a viable solution in the real world, attracting a large number of researchers' research and industry attention. In autonomous driving systems, neural networks not only are widely used in path planning and decision making, but also play a key role in scene understanding, object detection and semantic segmentation.

The application of neural network in automatic driving system is mainly reflected in the following aspects:

Path planning and decision making: Neural networks can learn a large amount of driving data to extract the patterns and laws of driving behaviour, so as to realize the path planning and decision making of autonomous vehicles. For example, recurrent neural networks (RNNS) can be used to predict the trajectory of a vehicle, or reinforcement learning algorithms can be used to optimize driving decisions.

Scene understanding: The neural network can realize the understanding and cognition of the scene through the analysis and processing of the perception data. For example, convolutional neural networks (CNNS) can be used to identify objects in a scene such as roads, traffic signs, pedestrians, etc, to help autonomous driving systems make decisions accordingly.

Object detection: Neural networks can be trained to recognize and detect various objects on the road, including vehicles, pedestrians, bicycles, etc. Common object detection algorithms include object detection algorithms based on convolutional neural networks, such as RCNN, Fast R-CNN, Faster R-CNN, etc.

Semantic segmentation: Neural networks can segment images into different semantic regions to achieve accurate recognition and understanding of objects such as roads, vehicles, and pedestrians. Common semantic segmentation algorithms include those based on convolutional neural networks, such as FCN and SegNet.

In general, the application of neural networks in autonomous driving systems can help vehicles realize the perception and understanding of the scene, so as to make corresponding decisions and driving planning. With the continuous development of artificial intelligence and deep learning technology, it is believed that there will be more breakthroughs and innovations in the application of neural networks in the field of autonomous driving.

Long et al [1] presented fully convolutional networks which excellently transform pixels to pixels and outperform the semantic segmentation by merging deep semantic and shallow appearance details for improved precision. The implementation achieved superior performance on PASCAL VOC [2], NYUDv2 and SIFT Flow.

In 2015, a novel network architecture, called UNet [3], was proposed to perform image segmentation tasks on large-scale images. This architecture relies on a contracting path to capture context, and an expansive path to make localization possible, thereby creating a u-shaped structure. A key feature of the architecture is the overlap-tile strategy, which ensures seamless segmentation of large images. SegNet [4], a novel deep convolutional neural network for semantic segmentation, was presented, featuring a unique encoder-decoder architecture. Its decoder's up sample method is to utilize pooling indices from the encoder's max-pooling step for non-linear upsampling, thereby avoiding learning to upsample. This mechanism resulted in efficient memory and computational performance during inference, with fewer trainable parameters compared to rival architectures.

Introduced DeepLab system enhanced semantic image segmentation via atrous convolution [5], allowing controlled feature response resolution and expanded filter field of view without increased computation or parameters. Integration of DCNNs with Conditional Random Fields (CRFs) refined object boundary localization, achieving a 79.7% *mIOU* on the PASCAL VOC-2012 test set and improved results on PASCAL-Context, PASCAL-Person-Part, and Cityscapes datasets [6].

DeepLabv3+ was proposed [7], merging spatial pyramid pooling and encoder-decoder structures for enhanced semantic segmentation. It extended DeepLabv3 by incorporating a decoder module, refining segmentation along object boundaries. Additionally, the model utilized the Xception model and depthwise separable convolution in both Atrous Spatial Pyramid Pooling and decoder modules, yielding a more efficient encoder-decoder network. This approach demonstrated effectiveness on PASCAL VOC 2012 and Cityscapes datasets, achieving 89% and 82.1% test set performance without post-processing, respectively.

Proposed SETR [8], employed a pure transformer [9], devoid of convolution and resolution reduction, to approach semantic segmentation as a sequence-to-sequence task. Inspired by ViT's effectiveness [10] and Transformer's success in NLP, SETR encoded images into patch sequences, providing global context at each layer and pairing with a simple decoder. This novel approach achieved state-of-the-art results on ADE20K (50.28% *mIoU*), Pascal Context (55.83% *mIoU*), and competitive performance on Cityscapes, offering a compelling alternative to traditional FCN models.

An innovative adversarial training approach for semantic segmentation models was proposed [11], marking the first application of Generative Adversarial Networks (GANs) to semantic segmentation. The methodology involved training a convolutional semantic segmentation network alongside an adversarial network, which discriminated between segmentation maps derived from ground truth and those generated by the segmentation network.

The introduced denoising diffusion probabilistic models (DDPMs) for high-quality image synthesis draws inspiration from nonequilibrium thermodynamics [12]. These models were trained using a unique connection between DDPMs and denoising score matching with Langevin dynamics, based on a weighted variational bound. This approach resulted in a progressive lossy decompression scheme, generalizing autoregressive decoding.

The author introduced the use of DDPMs for semantic segmentation [13], showcasing effectiveness particularly when labeled data is limited. By investigating intermediate activations from pretrained DDPMs during the Markov step of reverse diffusion, it was found that these activations encapsulate essential semantic information, providing valuable pixel-level representations for segmentation. With minimal training images, the proposed straightforward segmentation technique outperformed existing alternatives on several datasets, given the same level of human supervision.

Image semantic segmentation models face a series of specific challenges in autonomous driving applications for campus environments. For example, due to the complex and changing campus environment, lane lines and traffic signs are often incorrect, blurred, or missing. This makes it more difficult for autonomous driving systems to accurately identify these critical elements. Currently, most of the existing semantic segmentation datasets for autonomous driving focus on urban or highway scenarios, making them less effective when dealing with campus or similar environments. This scenario-specific problem emphasizes the need for more specific datasets and refined models for different application contexts. On the other hand, creating a semantically segmented dataset suitable for traditional deep learning models often demands a lot of time, effort, and labor costs [14]. This is because common neural networks, such as FCN, UNet++ [15], Deeplabv3plus, Transformer [16, 17], require a large dataset to ensure the training accuracy of the model. Additionally, effectively extracting multi-scale features to adapt to different sizes and shapes is also an urgent concern. To address these current limitations, we have created a dataset for image semantic segmentation suitable for campus environments and proposed an algorithm based on an improved diffusion model to achieve accurate semantic segmentation.

For data sets in the field of semantic segmentation, such as cityscapes, there are limitations in the generalization ability of scenes. Currently, it is difficult to build large completed data sets due to limited cost and resources. For UAVid [18], a dataset in the field of semantic segmentation of UAV images, the scenes it contains are more about city and street information, but it does not include categories such as playground and artificial lake, which are common in campuses.

Our key contributions include the following:

We collected drone images of campus environment in different regions (from multiple perspectives), sorted and annotated them to create a data set that solves semantic segmentation in specific scenes. We propose a semantic segmentation algorithm based on an improved multi-scale diffusion model, which achieves excellent recognition accuracy on this dataset by adding additional scale-variable jump join layers on the basis of the de-noising probabilistic diffusion model.

2 METHOD

The traditional image segmentation algorithms, whether threshold segmentation, region segmentation, edge segmentation or cluster-based methods, almost need to determine a threshold value or define a filter. When it comes to light changes, posture changes or background changes, there is almost nothing to be done, and such prior knowledge often requires expert experience and is difficult to be universal. Therefore, researchers are increasingly focusing on using deep learning networks to solve the problem of image segmentation, from the earliest CNNS and codec structures to the later deep generation models DGM (e.g. GAN, VAE, DDPM, etc.). The basic idea of DGM is to model the real data distribution from the training data, and then in turn use the learned model and distribution to generate and model new data. Its goal is to learn to generate new samples from the same data set as the training data. The core task of the generated model in the training stage is density estimation: learning the mapping relationship of the probability distribution. The model of data generation adopts the method of adding random Gaussian noise. When modelling the distribution, it is necessary to measure the difference between the noise and the training data distribution.

Diffusion model is a new SOTA in depth generation model. Diffusion model exceeds the original SOTA: GAN in image generation tasks, and has excellent performance in many application fields, such as computer vision, NLP, waveform signal processing, multimodal modelling, molecular diagram modelling, time series modelling, adversarial purification, etc [19]. The original diffusion model also has disadvantages, its sampling speed is slow, often requiring thousands of evaluation steps to extract a sample; its maximum likelihood estimation cannot be compared with the model based on likelihood; its ability to generalize to various data types is poor.

A fundamental concept of the diffusion model is the Markov chain (with stationarity). Markov chains are mathematical models that define systems that switch between different states over time. The existing state of the system can only determine the probability of transition to a particular state. The stationarity of Markov chains indicates that if a probability distribution changes with time, then it must tend to a stationary distribution (such as Gaussian distribution) under the action of Markov chains. Over a long enough period of time, the probability distribution will converge to this stationary distribution. This process of convergence is called the forward process. The essence of the transition probability of each time step of the Markov chain is to add noise, which is called diffusion. In contrast to the process of adding noise, a noise sample is obtained from the standard normal distribution, and then denoised step by step, and finally a sample in the data distribution is obtained.

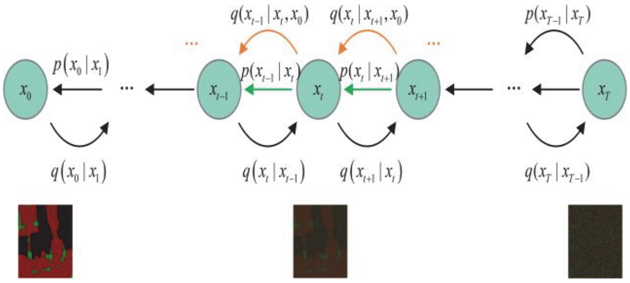

Figure 1 Schematic diagram of the probabilistic diffusion model of denoising

Assuming the original sample is X_0, and after T steps, we obtain a noised image, we can consider the forward diffusion process as a Markov process denoted by $q(xt|x\{t-1\})$. Here, βt is a parameter given in advance to control the progression of noise, and N represents the normal distribution function. Thus, for the forward diffusion process, we can get:

$$q\left(X_{1:T}|X_0\right):=\prod_{t=1}^{T}q\left(X_t|X_{t-1}\right) \tag{1}$$

$$q\left(X_t|X_{t-1}\right):=N\left(X_t;\sqrt{1-\beta_t X_{t-1},\beta_t I}\right) \tag{2}$$

The formula is:

$$X_t=\sqrt{1-\beta_t X_{t-1}}+\sqrt{\beta_t \varepsilon_t} \tag{3}$$

Among $\varepsilon_t\sim N\left(0,1\right)$ .The above formula represents the process of image generating noise, and also gives the method of how the noised image after any t step is calculated from the image of the previous step.

The process of diffusion is easier than denoising. In the denoising process, we know the original data and the previous noise-added data, the dependence of the noise-added data, and the random noise each time to conform to the standard normal. We need a de-noising network, feed data to X_t, the neural network, the network estimates the noise added when the other is born, and then we can remove the noise from X_t to get X_{t-1}. For the denoising process, we can correspondingly remember that the sum is obtained by training, which constitutes the final generative model.

The loss function is as follows:

$$p_\theta\left(X_{0:T}\right):=p\left(X_t\right)\prod_{t=1}^{T}p_\theta\left(X_{t-1}|X_t\right) \tag{4}$$

$$p_\theta\left(X_{t-1}|X_t\right)=N\left(X_{t-1};\mu_\theta\left(X_t,t\right),\sigma_\theta\left(X_t,t\right)\right) \tag{5}$$

$\mu_\theta\left(X_t,t\right)$ and $\sigma_\theta\left(X_t,t\right)$ are obtained by training, which constitutes the final generative model. The loss function is as follows:

$$\begin{aligned}&E\left[\log p_\theta\left(X_0\right)\right]|\le E_q\left[-\log\frac{p_\theta\left(X_{0:T}\right)}{q\left(X_{1:T}|X_0\right)}\right]\\&\left|=E_q\left[-\log\left(X_T\right)-\sum_{t\ge1}\log\frac{p_\theta\left(X_{t-1}|X_t\right)}{q\left(X_t|X_{t-1}\right)}\right]\right.\\&=-L_{VLB}\end{aligned} \tag{6}$$

In the image semantic segmentation task, the input is a 3 × H × W three-channel color image (H and W represent the height and width of the image respectively, and the unit is the number of pixels), and the output is a single-channel image matrix, whose dimension is H × W. Each element of the output matrix represents the semantic label represented by the pixel at the corresponding position in the original image, usually represented by a numeric code starting from 0.

Our model is based on DDPM. Our model is a U-Net model based on residual block [20] and attention block. U-Net belongs to encoder-decoder architecture, where encoder is divided into different stages, each stage contains a subsampling module to reduce the space size of features (H and W), and decoder gradually recovers the features compressed by encoder. Our proposed model uses sinusoidal position embedding to encode timestep. U-Net uses GroupNorm for normalization. Its core module is residual block (with built-in Attention mechanism).

3 EXPERIMENT

We have collected drone image data in August 2020. Considering the epidemic prevention policies of the campus and the city during COVID-19, we finally chose Wuhan University and Xiamen University to carry out the aerial survey mission. According to the maximum height of the buildings in the site, we set the height of the drone route to 150 meters. In order to enable the data set to better deal with the problem of shooting Angle, we used a five-lens camera (one lens at the bottom, one lens at the left and right directions, and 45° Angle with the horizontal plane) for data acquisition, so that the constructed data set contains images from various shooting angles. Considering the recognition effect of the image, we chose to conduct the experiment under clear and cloudless weather conditions.

Due to the impact of shutter speed, exposure compensation, etc, the original image data collected by UAV contains some poor quality pictures. On the other hand, there are some pictures with high overlap, which is not conducive to the model to better learn the rules of data. A total of about 600 images were manually selected and 400 images were obtained as a data set for training the model. Lableme software is used as the software for semantic segmentation and annotation. We spent about one week annotating all the pictures with detailed semantic segmentation categories. Taking account of the common features of the campus environment, we divide the categories into 5 categories, namely, buildings, vegetation, ground, playground, and lake. We divided the whole data set into a training set, a validation set, and a test set in a ratio of 2:1:1. For the data during training, we reorder the data randomly at the beginning of each epoch.

In order to clearly show the results of semantic segmentation, our dataset customizes the relationship between object category mapping colours and labels.

Table 1 The relationship between object class and RGB color mapping

Class label	Object class	Visual color of semantic segmentation graph
0	Building	(0, 0, 0)
1	Tree/Grass	(128, 0, 0)
2	Ground	(0, 128, 0)
3	Playing-field	(0, 0, 128)
4	Lake	(128, 128, 0)

Ideally, the number of categories should be equal because machine learning algorithms assume that the data is evenly distributed in the class. In the class imbalance problem, the broad problem is that the algorithm will be more biased to predict most classes, resulting in the algorithm not having enough data to learn patterns in a few classes. However, it can be seen from the figure that the categories in this data set are unbalanced, so we use the category weighting algorithm for processing.

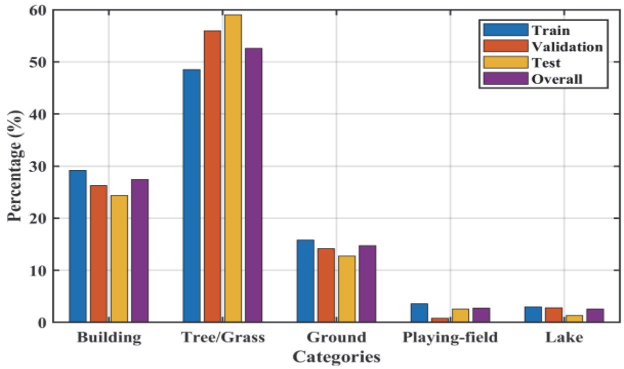

Figure 2 Histogram of the proportion of pixels in the data set

According to the introduction in the previous section, the network structure diagram of the diffusion model we built for semantic segmentation is shown in the figure below. On the whole, it is an UNet structure where the input raw RGB image is fed into the model through a diffusion process, superimposed with the standard normal distribution noise. Then, after ResNetBlock and downsampling for several times, the output after each downsampling structure is connected with the module of the upsampling part, that is, skip connection, so that the model can pay balanced attention to images of different levels and scales. It should be noted that the ground truth of semantic segmentation has never been input into the model, and it exists only to calculate the noise loss to ensure that the optimization direction of the model is correct.

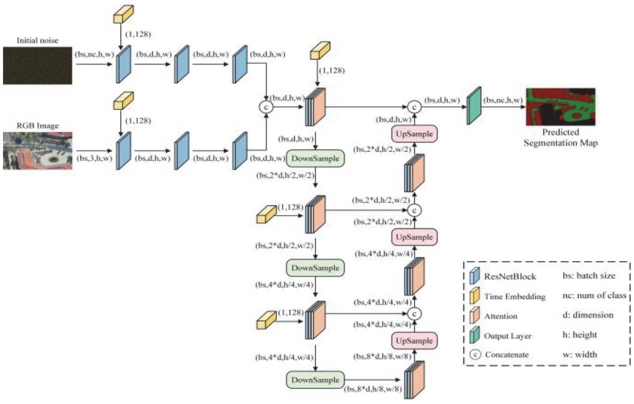

Figure 3 Model structure visualization (based on Unet) of DDPM

We perform a series of data enhancements to the dataset, including horizontal and vertical inversion of the image and rotation at any Angle to improve the generalization ability of the model. To check the performance of the model, we use $F1$ scores and $mIoU$ as evaluation metrics rather than accuracy alone. At present, precision and recall are often mentioned respectively in the algorithms related to recognition and detection.

$$precision = \frac{TP}{TP + FP} \tag{7}$$

$$recall = \frac{TP}{TP + FN} \tag{8}$$

TP (true positive) indicates that the ground truth is true and the output of the model is true. FP (false positive) indicates that the ground truth is false, but the output of this model is true; FN (false negative) means that the ground truth is false and the output of this model is false. F-score can balance the accuracy and recall rate of the two values, reflecting the accuracy of the algorithm.

$$F\text{-score} = \frac{\left(1+\beta^2\right) precision \times recall}{\beta^2 precision + recall} \tag{9}$$

from which β is a parameter set by itself. When $\beta \to 0$, F-score degenerates to precision. At $\beta \to \propto$, the F-score degenerates to recall. Moreover, when either precision or recall approaches 0, F-score will approach 0 at the same time, indicating that the accuracy of the algorithm is very low. In order to balance the influence of recall and precision, so that both indexes are high, $F1$-score is generally adopted to evaluate the performance of the model. Specifically, the F-score of the hour is $F1$-score. It means that we use the same weight for precision and recall, and is the harmonic average of precision and recall, which is the key to measuring the class imbalance problem. The ideal value of $F1$-score is close to 1, which requires precision and recall to be 1, which means that the model has the best performance.

Intersection over Union (IoU) is an indicator used to describe the degree of overlap. This was originally an indicator in the field of object detection, representing the ratio of the intersection area to the sum of the predicted area and the ground truth area.

$$IoU = \frac{TP}{\left(TP + FP + FN\right)} \tag{10}$$

is the average value of each type of IoU in the semantic segmentation result. It is calculated by the following formula.

$$mIoU = \sum_{i=1}^{N} IoU_i \tag{11}$$

where N represents the total number of categories in the semantically segmented dataset, which takes the value 5 for this dataset.

Due to the special principle and mechanism of diffusion model, it has high requirements for hardware memory and computing speed. On the other hand, the width and height of our drone images are 3000 and 2000, respectively. This is a very high resolution for an image generation model. Therefore, we trained with two high-performance servers, each containing two Tesla A100 graphics cards. Considering that Pytorch deep learning framework supports multi-machine and multi-card parallel training well, we use torch to write and train the model. In the training process, AdamW algorithm was finally adopted as the optimizer with a learning rate of 0.0005, weight decay of 0.001 and batch size of 16. The timesteps used in training are 32. The training rounds are 80. The total training duration is two days.

The following figure shows the changes of the two key indicators of the model in the training process with the training rounds.

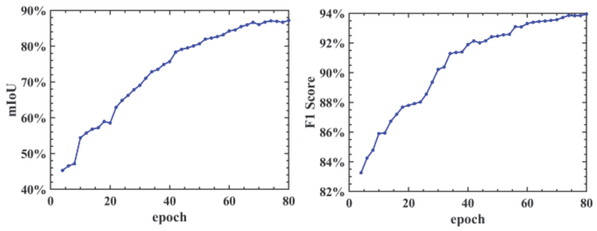

Figure 4 The change curve of *mIoU* and *F*1 score during model training

The $F1$ score and $mIoU$ of the model gradually increased with the increase of rounds, and the index improved rapidly in the first 50 training sessions. After that, with the increase of rounds, the improvement of indicators is relatively gradual. Therefore, we chose the model after 80 training sessions as the final model for testing.

It can be seen from the figure that the $mIoU$ of the model is 85.1%, and the IoU in the categories of buildings, vegetation and playgrounds is greater than 90%, and the largest is the playground, corresponding to 94.7%. This shows that the model has a very good segmentation effect for the above three categories, which can also be found from the previous figure. However, the segmentation effect of the model on the ground and lake is much lower than that of the other three categories, the lowest is 69.1% of the ground.

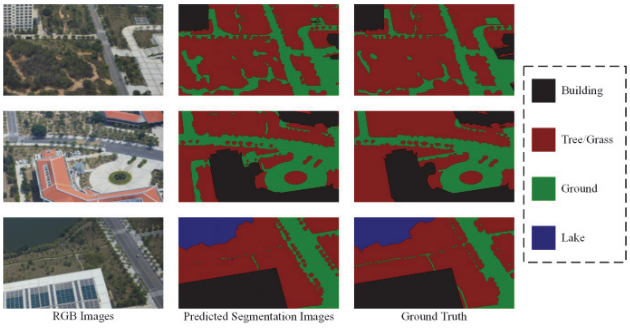

Figure 5 The predictive quality effect of the model on the test set (the first column is the original image, the second column is the semantically segmented image, and the third column is the ground truth)

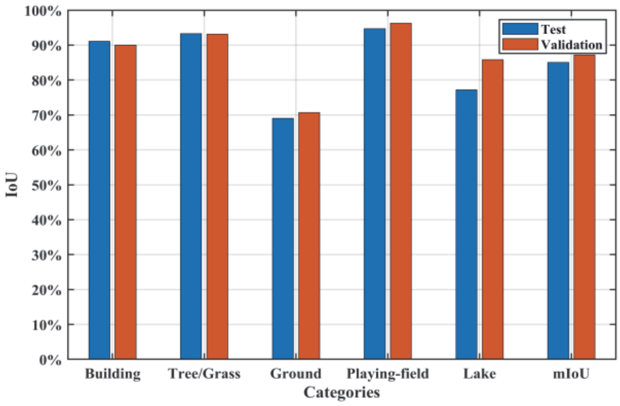

Figure 6 *IoU* for each category (a virtual horizontal line showing the av erage

4 DISCUSSION

We tested our proposed semantic segmentation algorithm based on the denoising probabilistic diffusion model on the data set created by ourselves. The above experimental results show that the model can achieve the semantic segmentation of campus scenes well. The average $mIoU$ is 85%, which is higher than the $mIoU$ index of 83.4% [21] obtained by other models in Cityscape data set at present, and the effect of 72.4% [22] obtained by other models in UAVid. Of course, our dataset contains only 5 categories, whereas UAvid contains 420 images in 8 categories, so the model has an inherent advantage to some extent, but an increase of nearly 10 percentage points is enough to demonstrate the effectiveness of our method.

When we evaluated the entire test set, we noticed a few images in which the model predicted better semantic segmentation than ground truth. One sample is shown in the figure below.

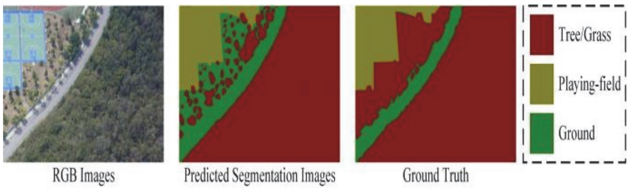

Figure 7 Comparison of a prediction result of the model with the true value

The prediction results of the model are more accurate in the segmentation and classification of many small vegetation areas on the ground, which is quite different from the ground truth. In fact, due to the limitations of energy and time, it is difficult for the ground truth pictures we manually annotated to be correct in every pixel. This is also the advantage of deep generation networks, by iterating more times to get a finer structure. It can be proved that our model can learn the edge details and curves of objects in the original image well, which is not only effective for large objects, but also satisfactory for edge segmentation of small objects with improved multi-scale structure. In addition, this situation in the figure above may be one of the reasons for the low segmentation *IoU* of the ground seen in the experimental results.

Our model also has some shortcomings, for example, semantic segmentation of some objects with sharp edges may result in too smooth or jagged segmentation results. The model has a low recognition rate for water bodies, which may be due to the small proportion of pixels of this category in the data set, and the reflection of trees and buildings in the lake often makes it difficult to correctly identify and segment.

5 CONCLUSION

This paper highlights the efficacy of a semantic segmentation algorithm based on a denoising probabilistic diffusion model within campus environments. The results on our bespoke dataset indicate the model's excellent performance in conducting semantic segmentation within these complex settings, with an average *mIoU* of 85%. Compared to other models' performance on datasets like Cityscape (83.4% *mIoU*) and UAVid (72.4%), our model shows notable improvement. The segmentation effectiveness is especially in categories such as buildings, vegetation, and playgrounds, where the *IoU* exceeds 90%, peaking at 94.7% for playgrounds. Nevertheless, the segmentation performance for ground and lake categories is lower, with the ground scoring the lowest at 69.1%. The model also showcases its proficiency in delineating finer object edges and details in images, offering accurate segmentation and classification for smaller vegetation areas on the ground which differed from the ground truth. Although the model demonstrates promising potential, there are areas for improvement, including segmentation of objects with sharp edges and increasing recognition rates for water bodies. These findings provide valuable insights and a robust foundation for further refining and expanding the application of the proposed semantic segmentation algorithm in varied and dynamic environments, such as campuses.

Acknowledgements

This research was supported by the Geely University of China under Grant (Project No. 2022xjgglz04, Project No. 2023XJG023); The National Key Research and Development Program of China (Project No. 2022YFC3301401); Humanities and Social Science Fund of Ministry of Education, China (Project No. 21XJA630004); Supported by Sichuan Science and Technology Program, China (Project No. 2023JDR0194).

6 REFERENCES

[1] Long, J., Shelhamer, E., & Darrell, T. (2015). Fully Convolutional Networks for Semantic Segmentation. arXiv:1411.4038. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* https://doi.org/10.1109/CVPR.2015.7298965

[2] Everingham, M., Van Gool, L., Williams, C. K. I., Winn, J., & Zisserman, A. (2010). The PASCAL Visual Object Classes (VOC) Challenge. *International Journal of Computer Vision, 88*, 303-338. https://doi.org/10.1007/s11263-009-0275-4

[3] Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. In: Navab, N., Hornegger, J., Wells, W., Frangi, A. (eds). *Medical Image Computing and Computer-Assisted Intervention-MICCAI 2015. MICCAI 2015. Lecture Notes in Computer Science, 9351*. Springer, Cham. https://doi.org/10.1007/978-3-319-24574-4_28

[4] Badrinarayanan, V., Kendall, A., & Cipolla, R. (2017). SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *39*(12), 2481-2495. https://10.0.4.85/TPAMI.2016.2644615

[5] Long Qing, C., Papandreou, G., Kokkinos, I., Murphy, K., & Yuille, A. L. (2016). DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *arXiv (Cornell University), 40*(4).

[6] Cordts, M, Omran, M, Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., & Schiele, B. (2016). The Cityscapes Dataset for Semantic Urban Scene Understanding. *2016 IEEE Conference on Computer Vision and Pattern Recognition(CVPR).* https://doi.org/10.1109/cvpr.2016.350

[7] Chen, L., Zhu, Y., Papandreou, G., Schroff, F., & Adam, H. (2018). Encoder-Decoderwith Atrous Separable Convolution for Semantic Image Segmentation. *Computer Vision-ECCV 2018.* https://doi.org/10.1007/978-3-030-01234-2_49

[8] Zheng, S., Lu, J., Zhao, H., Zhu, X., Luo, Z., Wang, Y., Fu, Y., Feng, J., Xiang, T., & Zhang, L. (2020). Rethinking Semantic Segmentation from a Sequence-to-Sequence Perspectivewith Transformers. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).* https://doi.org/10.1109/CVPR46437.2021.00681

[9] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L. &Polosukhin, I. (2017). Attention Is All You Need. arXiv.org. https://doi.org/10.48550/arXiv.1706.03762

[10] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (n.d.). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale.

[11] Gaurav, S. & Sharma, K. (2021).Semantic Segmentation using Generative Adversarial Networks with a Feature Reconstruction Loss. *2021 Asian Conference on Innovation in Technology (ASIANCON).* https://doi.org/10.1109/asiancon51346.2021.9544954

[12] Ho, J., Jain, A., & Abbee, P. (2020). Denoising Diffusion Probabilistic Models.

[13] Dmitry Baranchuk, Rubachev, I., Andrey Voynov, Khrulkov, V., & Babenko, A. (2021). Label-Efficient Semantic Segmentation with Diffusion Models. arXiv (Cornell University).

[14] Alam, F., Sang Ko, H., Lee, H. F., & Yuan, C. (2023). Deep Learning Approach for Volume Estimation in Earthmoving Operation. *International Journal of Industrial Engineering and Management, 14*(1), 41-50.

https://doi.org/10.24867/IJIEM-2023-1-323

[15] Zhou, Z., Siddiquee, M. M. R., Tajbakhsh, N., & Liang, J. (2019). UNet++: Redesigning Skip Connections to Exploit Multiscale Features in Image Segmentation. *IEEE Transactions on Medical Imaging, 39*(6),1856-1867. https://doi.org/10.1109/tmi.2019.2959609

[16] Niu, B., Feng, Q., Chen, B., Ou, C., Liu, Y., & Yang, J. (2022). HSI-TransUNet: A transformer based semantic segmentation model for crop mapping from UAV hyperspectral imagery. *Computers and Electronics in Agriculture, 201*. https://doi.org/10.1016/j.compag.2022.107297

[17] Chen, J., Lu, Y., Yu, Q., Luo, X., Ehsan Adeli, Wang, Y., Lu, L., Yuille, A. L., & Zhou, Y. (2021). Trans UNet: Transformers Make Strong Encoders for Medical Image Segmentation.

[18] Lyu, Y., Vosselman, G., Xia, G. S., Yilmaz, A., & Ying, Y. (2018a). UAVid: A Semantic Segmentation Dataset for UAV Imagery. arXiv (Cornell University).

[19] Yang, L., Zhang, Z., Song, Y., Hong, S., Xu, R., Zhao, Y., Zhang, W., Cui, B., & Yang, M. (2023). Diffusion Models: A Comprehensive Survey of Methods and Applications. *ACM Computing Surveys, 56*(4), 1-39. https://doi.org/10.1145/3626235

[20] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770-778. https://doi.org/10.1109/cvpr.2016.90

[21] Rizzoli, G., Barbato, F., & Zanuttigh, P. (2022). Multimodal Semantic Segmentation in Autonomous Driving: A Review of Current Approaches and Future Perspectives. *Technologies, 10*(4). https://doi.org/10.3390/technologies10040090

[22] Ding, Y., Zheng, X., Chen, Y., Shen, S., & Xiong, H. (2022). Dense context distillation network for semantic parsing of oblique UAV images. *International Journal of Applied Earth Observation and Geoinformation, 114*. https://doi.org/10.1016/j.jag.2022.103062

Contact information:

Wang WANG, lecturer
(Corresponding author)
Geely University of China,
Chengdu, Sichuan, 641423, China
E-mail: wangwang@guc.edu.cn

Chen ZHOU, Professor, PhD
Wuhan University,
Wuhan, Hubei, 430072, China
E-mail: chenzhou@whu.edu.cnil

Hua HE, Professor, PhD
Chongqing Technology and Business University,
Chongqing, 400067, China
E-mail: huahe@guc.edu.cn

Changsong MA, Professor, PhD
Krirk University,
Bangkok, 10220, Thailand
E-mail: changsongma@guc.edu.cn