frontiers | Frontiers in Earth Science

TYPE Original Research
PUBLISHED 28 March 2023
DOI 10.3389/feart.2023.1115805

Check for updates

OPEN ACCESS

EDITED BY
Yunhui Zhang,
Southwest Jiaotong University, China

REVIEWED BY
Jingren Zhou,
Sichuan University, China
Xin Wang,
Jilin University, China

*CORRESPONDENCE
Zhengang Jiang,
✉ jiangzhengang@cust.edu.cn
Hua Zhang,
✉ zhanghua@ccit.edu.cn

SPECIALTY SECTION
This article was submitted to Environmental Informatics and Remote Sensing, a section of the journal Frontiers in Earth Science

RECEIVED 04 December 2022
ACCEPTED 23 February 2023
PUBLISHED 28 March 2023

CITATION
Zheng G, Jiang Z, Zhang H and Yao X (2023), Deep semantic segmentation of unmanned aerial vehicle remote sensing images based on fully convolutional neural network.
Front. Earth Sci. 11:1115805.
doi: 10.3389/feart.2023.1115805

Deep semantic segmentation of unmanned aerial vehicle remote sensing images based on fully convolutional neural network

Guoxun Zheng[1,2,3], Zhengang Jiang[1]*, Hua Zhang[1,2,3]* and Xuekun Yao[2,3]

[1]School of Computer Science and Technology, Changchun University of Science and Technology, Changchun, China, [2]School of Computer Technology and Engineering, Changchun Institute of Technology, Changchun, China, [3]Jilin Provincial Key Laboratory of Changbai Historical Culture and VR Reconstruction Technology, Changchun Institute of Technology, Changchun, China

In the era of artificial intelligence and big data, semantic segmentation of images plays a vital role in various fields, such as people's livelihoods and the military. The accuracy of semantic segmentation results directly affects the subsequent data analysis and intelligent applications. Presently, semantic segmentation of unmanned aerial vehicle (UAV) remote-sensing images is a research hotspot. Compared with manual segmentation and object-based segmentation methods, semantic segmentation methods based on deep learning are efficient and highly accurate segmentation methods. The author has seriously studied the implementation principle and process of the classical deep semantic segmentation model—the fully convolutional neural network (FCN), including convolution and pooling in the encoding stage, deconvolution and upsampling, etc., in the decoding stage. The author has applied the three structures (i.e., FCN-32s, FCN-16s, and FCN-8s) to the UAV remote sensing image dataset AeroScapes. And the results show that the accuracy of vegetation recognition is stable at about 94%. The accuracy of road recognition can reach up to more than 88%. The mean pixel accuracy rate of the whole test dataset is above 91%. Applying full convolution neural network to semantic segmentation of UAV remote sensing images can improve the efficiency and accuracy of semantic segmentation significantly.

KEYWORDS

deep learning, unmanned aerial vehicle, remote sensing, semantic segmentation, FCN

1 Introduction

In remote sensing, satellite remote sensing images or UAV remote sensing images can be used for ground object recognition. Satellite remote sensing images have a low resolution for low-altitude targets. They are often affected by weather factors and obscure ground objects, resulting in difficulties in ground object recognition. UAV remote sensing technology takes low-speed unmanned aircraft as the aerial remote sensing platform, captures aerial image data with infrared and camera technology, and processes the image information by computer. Compared with satellite remote sensing platforms, UAVs fly at a lower altitude and can fly close to the ground to improve the resolution of objects (Liu et al., 2021). And their close-range image resolution can reach the centimeter level, which can quickly and economically collect low-altitude high-resolution aerial images. UAV remote sensing can be applied to environmental monitoring (Green et al., 2019). It can fast update, correct, and upgrade geo-environmental information and outdated GIS databases, providing timely technical assurance for government and related departments' administration, land, and geo-environmental management. In addition, UAV remote sensing can also be applied to electric power inspection (Zhang et al., 2017), agricultural monitoring (Zhang et al., 2021), high-speed patrol (Yang et al., 2021), disaster monitoring and prevention (Kamilaris Prenafeta-Boldú, 2018), meteorological detection (Funk and Stütz, 2017), aerial survey (De Benedetti et al., 2017), etc. In recent years, UAV remote sensing has become a hot topic for global research due to its mobility, speed, and economic advantages. It has gradually developed from research and development to the practical application stage, becoming one of the future leading aerial remote sensing technologies.

With the development of deep learning and the Internet of Things (IoT), the research on the integration of UAV remote sensing and artificial intelligence has become more and more abundant. Many researchers have succeeded in automatic target recognition of UAV remote sensing images based on convolutional neural networks with the help of deep learning methods, such as Region-Convolutional neural network (R-CNN), Fast Region-based Convolutional Network (Fast R-CNN), Faster Region-based Convolutional Network (Faster R-CNN), Single Shot Mutibox Detector (SDD), You Only Look Once (YOLO) and other frameworks (Xu et al., 2017; Li et al.; Liu et al., 2020). Xu et al. (2017) extended the framework of Faster R-CNN for detecting cars from low-altitude UAV images taken over signal intersections and demonstrated that Faster R-CNN has excellent potential for parking lot car detection. Li et al. (2020) proposed a method for UAV monitoring railroad scenes based on SSD detection of small objects. Liu et al. (2020) developed a special detection method for small targets in UAV view based on YOLOv3. All these methods used regular rectangles to frame and identify targets. Still, in many cases, one would like to be able to use the shape of the target itself to locate and precisely segment it, for example, to precisely distinguish the shape of each building, road, river, vehicle, etc., itself, i.e., to achieve semantic segmentation.

Semantic segmentation has been a research hotspot in artificial intelligence. It takes some raw data (e.g., a flat image) as input and converts them into a mask with highlighting by finding the location of all pixels and what they represent, thus understanding the meaning of the image. Semantic segmentation can be used in land monitoring (Mohammadimanesh et al., 2019), autonomous driving (Li et al., 2021), face recognition (Meenpal et al., 2019), precision agriculture (Milioto et al., 2018), etc., and plays a vital role in social development and people's life. This paper discusses applying the fully convolutional neural network in deep learning to semantic segmentation of UAV remote sensing images to accurately extract vegetation, roads, and other targets in the features and improve the segmentation accuracy (Li et al., 2019; Li et al., 2019).

The contributions of this paper are as follows.

(1) The author has conducted an in-depth study on the structure of the fully convolutional neural network and meticulously analyzed the encoder and decoder components. And it forms a network structure image with clear feature map size variation, convolutional kernel size, and fusion ideas.
(2) Under the premise of maintaining the classification balance to the greatest extent, the training set, validation set, and a test set of the AeroScapes dataset are re-divided, and the image files and labels are normalized and label encoded before semantic segmentation, and a complete set of data processing algorithm flow is refined.
(3) The experiments of applying three fully convolutional neural networks (FCN-32s, FCN-16s, and FCN-8s) on the UAV remote sensing image dataset AeroScapes are completed, and the results show that the segmentation effect is good.

The remainder of this paper is organized as follows: Section 2 introduces the work related to UAV remote sensing images, semantic segmentation, and FCN; Section 3 explains the dataset used and the processing method of the dataset, the semantic segmentation method of UAV remote sensing images and the implementation steps; Section 4 shows the experimental results and the related discussion; Section 5 concludes the work of this paper.

2 Related works

This section introduces the research related to this paper, which includes semantic segmentation models and fully convolutional neural networks.

2.1 Semantic segmentation models

Semantic segmentation is one of the critical tasks in computer vision. It is the process of classifying each pixel in an image and linking each pixel to a category label, which is widely used in medical image analysis (Yang and Yu, 2021), unmanned driving (Feng et al., 2020), geographic information systems (Li et al., 2019), etc., Various models, such as FCN, SegNet, U-Net, DeepLab, etc., can achieve semantic segmentation.

FCN is the cornerstone of deep learning techniques applied to semantic segmentation problems (Shelhamer et al., 2017), building a fully convolutional neural network. The convolutional model with images introduces conditional random fields (CRF) as a post-processing module in the CNN to tune the output of the segmentation architecture and enhance it to capture fine-grained information. The decoder encoder model is divided into encoder structure and decoder structure. SegNet encoder follows the network model of VGG16, which mainly categorizes and analyzes the low-level local pixel values of the image to obtain higher-order semantic information to achieve parsing of object information. Pyramid Scene Parsing Network (PSPN) (Zhao et al., 2017) is a multi-scale and pyramid network-based model that uses ResNet, a covariance network with null convolution, for feature extraction and better learning of global information. The null convolution model can exponentially expand the field of perception without losing resolution. The most typical model of null convolution is DeepLab and its upgraded version. DeepLab V1 (Chen et al., 2014) uses null convolution and CRF to solve the problem of information loss and probabilistic model between labels not being applied due to previous model pooling. DeepLab V2 (Chen et al., 2017a) introduces Atrous Spatial Pyramid Pooling (ASPP), which extracts features using multiple sampling rates of null convolution in parallel, and then fuses the features and changes the base layer from VGG16 to ResNet. DeepLab V3 (Chen et al., 2017b) proposes a more

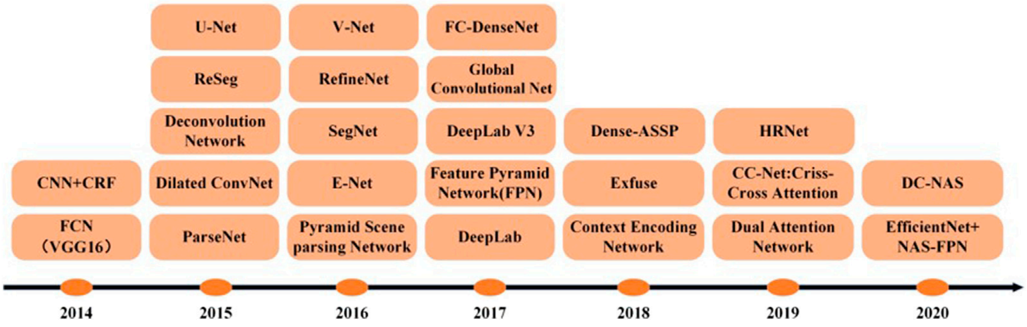

FIGURE 1
The timeline of DL-based segmentation model for 2D images.

general framework that applies to any network. DeepLab V3+ (Chen et al., 2018) adds encoder-decoder constructs to achieve accuracy and time balance by changing the Atrous rate. The recurrent neural network-based model RNN successfully models global contextual information and improves segmentation results by linking pixel level with local information. The attention mechanism model (Chen et al., 2016) assigns different weights to different scale images, e.g., large weights to small-scale targets to achieve large scaling and small scaling to closer targets in the image. Learning active contour models ACMs (Chen et al., 2019) propose a new loss function that considers the information of boundary line length and region and a convolutional neural network based on Dense U-Net. In addition, there are segmentation models based on the GAN for semi-supervised semantic segmentation. The semantic segmentation model based on deep learning is shown in Figure 1 (Minaee et al., 2021).

2.2 Fully convolutional neural network

FCN is the first neural network applied to image semantic segmentation, which replaces the fully connected layers of VGGNet with convolutional layers to build a deep, fully convolutional neural network. FCN adopts "end-to-end" feature learning, which reveals the non-linear features hidden in the data through multi-layer feature extraction, and can automatically learn global features from a large number of training sets to realize the transformation of feature models from manual to learned features.

In the FCN, the last three layers of the CNN network are all transformed into multi-channel convolutional layers of equivalent vector length corresponding to 1 × 1 convolutional kernels (Shelhamer et al., 2017). The network model consists entirely of convolutional layers, and no fully connected layers generate vectors. CNN is an image-level recognition, that is, from image to result. At the same time, FCN is a pixel-level recognition, labeling which category each pixel on the input image is most likely to belong to. FCN changes the classification network into a fully convolutional neural network, specifically transforming the fully connected layers into convolutional layers with up-sampling by deconvolution, fine-tuning using migration learning methods, and using jump structure. Thus, semantic information can be combined with symbolic information to produce accurate and surprising segmentations. The first half of the FCN network is based on the convolutional layer of VGG, so the weight parameters of the VGG network are directly referenced as the pre-training parameters of the FCN and then fine-tuned. The structure of the FCN and the coding and decoding process are introduced in Section 3.2. Like other segmentation network models, the FCN model loss function is a pixel-level cross-entropy loss function and also uses a stochastic gradient descent optimization algorithm for momentum. Instead of following the previous interpolation of interpolation upsampling, FCN proposes a new upsampling, i.e., deconvolution, which can be understood as the inverse operation of the convolution operation. The deconvolution cannot compound the loss of values due to the convolution operation and simply reverses the steps in the convolution process to transform once, so it is also called transposed convolution. The convolution formula is shown in formula 1, and the deconvolution formula is shown in formula 2:

$$F = \frac{i - k + 2 \times p}{S} + 1 \tag{1}$$

Where F is the output image size. i is the input image size. k is the convolution kernel size. p is the padding, the complementary zero size. S is the step size.

$$F = (i - 1) \times S + k - 2 \times p \tag{2}$$

Formula 2 is to exchange the input and output in formula 1 into a sparse state.

FCN training, at least 175 epochs after the algorithm, will perform well. That is, too little will affect the algorithm performance, while too much on the algorithm performance is no more remarkable improvement. And its learning rate can be adjusted after 100 times and is getting smaller. FCN in the upsampling process, only the fusion of pool 5 and pool 4 feature map, pool 3 before the feature map, does not need to be fused. When FCN emerged, it surpassed the most advanced techniques in semantic segmentation. It allows the input of images of arbitrary size, and the output of the same size can be obtained after effective learning, and state-of-the-art results are obtained on the PASCAL VOC, NYUDv2, and SIFT Flow datasets. However, FCN does not have a category-balancing strategy, and the accuracy of semantic

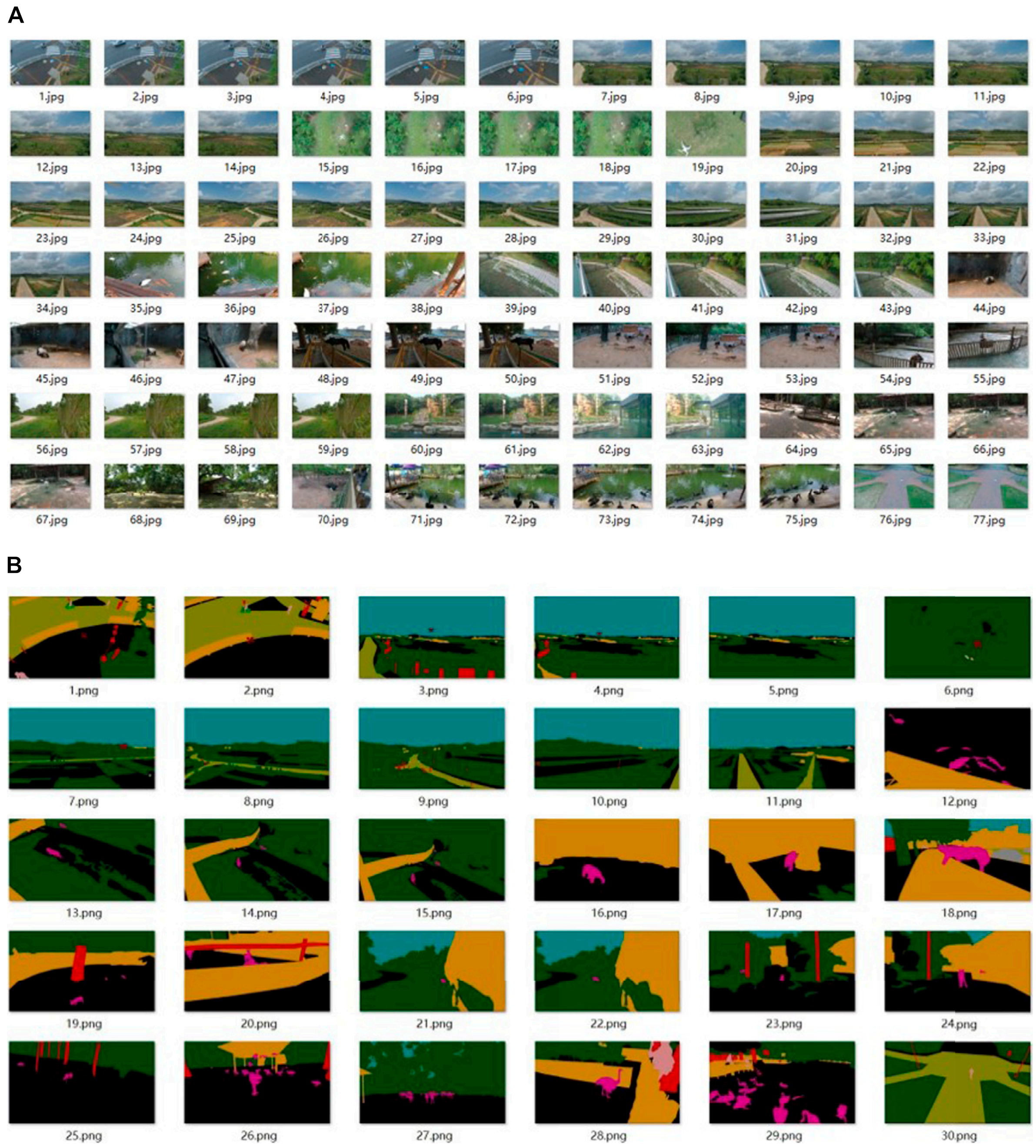

FIGURE 2
Dataset of semantic segmentation: **(A)** partial images in the dataset; **(B)** corresponding label data.

segmentation suffers when the categories in a dataset are not balanced.

3 Data and methods

3.1 Semantic segmentation dataset

The AeroScapes aerial semantic segmentation benchmark includes images captured using commercial UAVs in the altitude range of 5–50 m. The dataset provides 3,269 720p (1,280 × 720) images and 11 categories (excluding background) of ground-truth masks (Figure 2). The dataset offers 3,269 720p (1,280 × 720) images and 11 categories (excluding background) of real masks (Figure 2A), where the 11 categories include Person, Bike, Car, Drone, Boat, Animal, Obstacle, Construction, Vegetation, Road, and Sky. The dataset is provided with a mask map, so no mask map conversion is required. The file structure is as follows.

ImageSets folder: two txt files are stored, dividing the training and test sets.

JPEGImages folder: holds the RGB images.

SegmentationClass folder: holds the mask map of the labels.

Visualizations folder: holds the label images.

This dataset does not divide the training and test sets directly into the corresponding folders, so to use this dataset, it is necessary to read the images according to the divided txt file and distribute each image in the corresponding folder. In addition, the validation set is not provided in this dataset. In this paper, when the original remote sensing images are sent to FCN for training, the author sets the crop to 1,280 × 704 pixels size and divides a part of the images as the validation set under the premise of ensuring the category balance as much as possible in the training set.

It can be seen from Figure 2B that various categories of objects are labeled with different colors, and this is the final output image effect of FCN.

3.2 Semantic segmentation of UAV remote sensing images

UAV remote sensing images usually contain a wide range of complex features, and insignificant differences between features, making it challenging to obtain high segmentation accuracy using manual methods. In the early days, most segmentation networks are classified for pixels by finding a piece of the region containing this pixel and using the category of this region as the category of pixel points, which is memory-consuming and inefficient as the areas may overlap. Therefore, this paper uses deep learning to perform semantic segmentation (Minaee et al., 2021).

3.2.1 The principle and implementation of the three structures of the FCN model

Many network models for deep semantic segmentation include fully supervised learning image semantic segmentation methods and weakly supervised image semantic segmentation methods (Huang et al., 2021). However, the performance of most weak supervised methods still lags behind that of fully supervised methods (Tian et al., 2019). This paper uses a fully supervised learning method, FCN, to perform semantic segmentation on the AeroScapes aerial semantic dataset. FCN is the first attempt to classify pixels directly from abstract semantic features, and the pre-training model of VGGNet is used in advance to greatly reduce the training time without affecting the classification accuracy. The pre-training model of VGG16 is used in this paper. After the first five convolutions and pooling, FCN replaces the original fully-connected layer with a fully-convolutional layer. And then completes the deconvolution operation by bilinear interpolation and sums with the corresponding pooled middle layer information to recover the original image resolution finally and achieves end-to-end, pixel-to-pixel semantic segmentation, and the most crucial feature of this network is that it learns features by itself according to the designed algorithm without human intervention (Shelhamer et al., 2017). In this paper, three fully convolutional neural networks, FCN-32s, FCN-16s, and FCN-8s, have been used to carry out the study, and the structures of the three network models are shown in Figure 3.

These three networks, with identical pre-encoding processes, complete the downscaling and feature extraction of the original image, and the good or bad feature extraction directly affects the final prediction results. As shown in Figure 3, after two convolutional layers, called 1 convolutional block (i.e., Conv1), the feature map is obtained, and after the first pooling layer (i.e., Pool1), the feature map is obtained after 1/2. Similarly, after the second convolution block and the second pooling layer, the feature map becomes 1/4 of the original map; after the third convolution block and the third pooling layer, the feature map becomes 1/8 of the original map; after the fourth convolution block and the fourth pooling layer, the feature map becomes 1/16 of the original map; after the fifth convolution block and the fifth pooling layer, the feature map becomes 1/32 of the original map. This process is called downsampling, i.e., the process of feature extraction. FCN uses the five convolutions and five pooling in the VGG16 model, the feature map size becomes 1/32 of the input image size, and the channels change from 3 to 512. FCN replaces the sixth and seventh fully connected layers in VGG16 with fully convolutional layers, and the feature map size does not change, still 1/32 size, but transforms the number of channels to 4,096. After the final transition convolution, the channel number is converted to the number of label categories of the dataset.

The difference between the three networks lies in the post-decoding process, where the decoding is completed to gradually recover the smaller size feature maps into predicted maps of the same size as the image. FCN-32s is to directly up-samples the encoded feature map by 32 times, i.e., to complete the deconvolution. Then, the obtained feature map is passed to the softmax classifier to output a prediction map of the same size as the input image to get the dense prediction result, which does not use the skip architecture. FCN-16s first up-samples the 1/32 sized encoded feature map by a factor of 2 to 1/16 size, and transform the 1/16 sized feature map after the fourth pooling (Pool 4) into the number of label categories of the dataset by transition convolution, then fuses the two 1/16 sized feature maps (i.e., skip architecture). Finally, up-samples them by a factor of 16 to produce a prediction map of the same size as the input image after the softmax classifier. FCN-8s first up-samples the fused feature map from the previous step by a factor of 2 to 1/8 size, transforms the 1/8 sized feature map after the third pooling (Pool 3) to the number of label categories of the dataset by transition convolution, and fuses the two 1/8-sized feature maps, again with skip architecture. Finally, it up-samples them by a factor of 8 to produce a prediction map of the same size as the input image after the softmax classifier. It is worth noting that in FCN-16s and FCN-8s, when multiple feature maps are fused, it must be ensured that each feature map size is the same. The skip architecture of FCN enables the models to ensure both robustness and accuracy, and all three models achieve end-to-end deep semantic segmentation.

3.2.2 Data preprocessing

The deep semantic segmentation network needs a lot of time and effort to process the dataset before training, and the processing effect of the dataset directly affects the accuracy of semantic segmentation. Each image in the dataset corresponds to a labeled graph, and each category of targets in the image is identified with different colors, i.e., different labels, and the preprocessing process is shown in Figure 4.

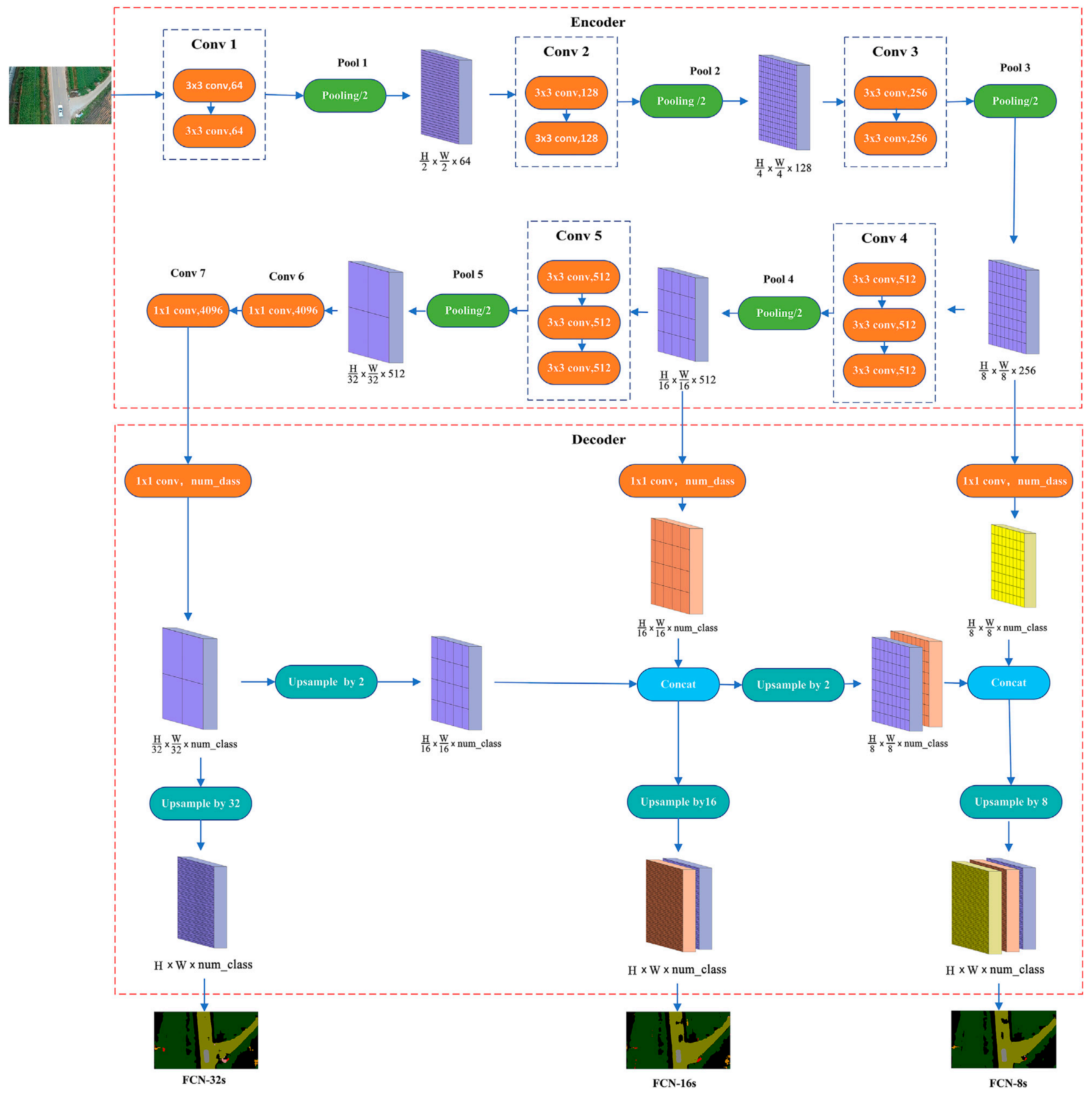

FIGURE 3
Deep learning network structure, including encoder and decoder. The decoder part shows the decoding process of FCN-32s, FCN-16s, and FCN-8s respectively. **(A)** FCN-32s; **(B)** FCN-16s; **(C)** FCN-8s.

3.2.3 A label processing and encoding

Label encoding is to form a one-to-one correspondence from colors to labels. It needs to store the names and corresponding RGB values of all categories (including background category) in a csv file to form a color map, and hash map each pixel point in the color map to the category it represents using a 256 decimal-like method through a hash function, as shown in formula 3 and formula 4.

$$k = (cm[0] \times 256 + cm[1]) \times 256 + cm[2] \tag{3}$$

$$cm2lbl[k] = i \tag{4}$$

$cm[0]$, $cm[20]$, and $cm[10]$ denote the RGB value in a pixel, k denotes the converted integer, $cm2lbl$ is a hash table constructed using the hash function, and k is used as the index of the pixel in the $cm2lbl$ table to query the category i corresponding to the pixel.

3.2.4 B initializing the dataset

After completing the corresponding processing, initializing the dataset is to divide the images and labels into the training set, validation set, and test set. Firstly, it is necessary to define the crop size and transformation content. The AeroScapes dataset needs to

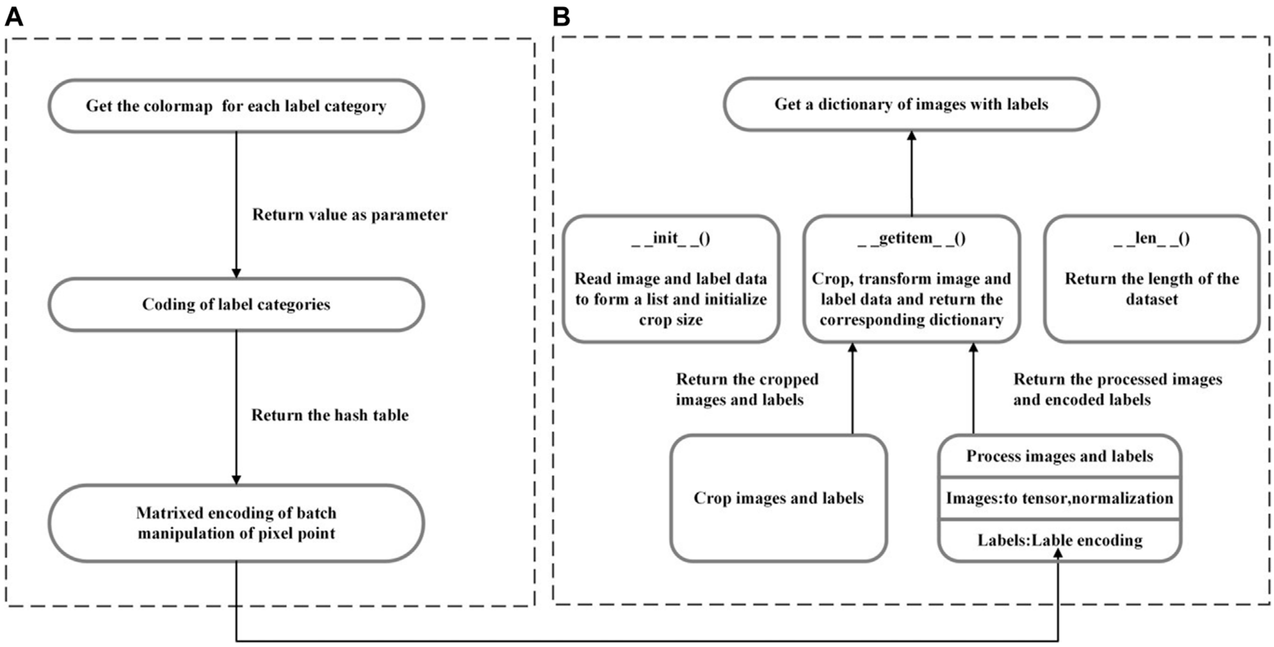
A
Get the colormap for each label category
Return value as parameter
Coding of label categories
Return the hash table
Matrixed encoding of batch manipulation of pixel point
B
Get a dictionary of images with labels
__init__()
Read image and label data to form a list and initialize crop size
__getitem__()
Crop, transform image and label data and return the corresponding dictionary
__len__()
Return the length of the dataset
Return the cropped images and labels
Return the processed images and encoded labels
Crop images and labels
Process images and labels
Images:to tensor,normalization
Labels:Lable encoding

FIGURE 4
Data preprocessing process: **(A)** label processing and encoding; **(B)** initializing the dataset.

crop each image to 1,280 × 704, transform the image into a tensor and normalize it, and encode the labels using the steps in Figure 3A. In this way, the cropped and correspondingly transformed images and labels can be obtained and combined into a corresponding dictionary for subsequent use.

3.2.5 Network training

Semantic segmentation of remote sensing images based on a fully convolutional neural network is trained using PyTorch 1.12 GPU version deep learning framework, Python 3.7, Intel (R) Xeon (R) 12-core processor, NVIDIA Quadro P6000, 24G GDDR5X video memory, and 64G DDR4 memory. In this paper, the training set, validation set, and test set are divided according to the ratio of 6: 2:2, and 1967 training images, 654 validation images, and 648 test images, and the training steps and techniques are shown as follows:

(1) Setting parameters: setting the number of categories, Batch Size, Epoch, initial learning rate value, and image crop size. The AeroScapes dataset contains 11 categories and 1 unlabeled category, 12 in total. In order to improve the training speed and segmentation accuracy, the Batch Size is set to 4; Epoch is set to 175; the learning rate is initialized to 1×10^{-4}, and every time 50 Epochs are completed, the learning rate is reduced to half of the original one. The training of FCN goes through 5 times of pooling, and after each pooling, the feature map size will change to 1/2 of the original one, and to prevent the image size leads to training failure, the image size is uniformly cropped to 1,280 × 704.
(2) Downloading the pre-training model: in order to speed up the training, migration learning can be performed using some pre-trained models to obtain the weights of the model parameters quickly. When training FCN, the pre-encoding process uses the backbone of VGG16, and at the beginning of the first training, the model structure of VGG16 is downloaded from the network first according to the URL https://download.pytorch.org/models/vgg16-397923af.pth, and it only needs to be downloaded once and saved locally. None of the subsequent training needs to be downloaded, which dramatically saves training time, obtains reliable parameter weight values, and increases the training speed of the network. In this paper, the pre-training model uses the network structure of VGG16, and subsequent attempts can be made to use the structure such as VGG19.
(3) Randomly disrupting the order of training images: In deep learning, the model is often "biased" in the training process because the dataset is not disrupted, and the trained network model cannot fit the abstract features of the training set well, and the performance is poor. When FCN is trained, the order of the training images is randomly shuffled. And the model can learn different features of the training set better instead of being limited to some features, which not only enhances the generalization ability of the model but also improves the training accuracy.
(4) Selecting Adam optimizer: after the loss function is calculated in the training of the deep neural network, the optimizer needs to be used to obtain the network parameters with the minimum loss function for backpropagation and complete the update of the network parameters, so as to complete the model training as fast as possible and save computer resources. In this paper, the Adam (Kingma and Ba, 2014) optimizer is chosen, which is simple and efficient, requires less memory, makes the convergence speed fast

while making the fluctuation amplitude small, and achieves parameter self-renewal by the newly added two correction terms.

4 Results and discussion

The main evaluation metrics of semantic segmentation are execution time, memory usage, and accuracy, where the accuracy metrics include pixel accuracy (PA), mean pixel accuracy (mPA), and mean intersection over union (mIoU) (Feng et al., 2020).

Assuming that there are k categories (including one background category), p_{ij} represents the total number of pixels that are true for category i but predicted for category j. Specifically, p_{ii} represents true positives, p_{ij} represents false positives, and p_{ji} represents false negatives; the pixel accuracy can be calculated with the formula 5.

$$PA = \frac{\sum_{i=1}^{k} p_{ii}}{\sum_{i=1}^{k}\sum_{j=1}^{k} p_{ij}} \tag{5}$$

The pixel accuracy represents the ratio of the number of correctly classified pixel points to the number of all pixel points.

The mean pixel accuracy can be calculated with the formula 6.

$$mPA = \frac{1}{k}\sum_{i=1}^{k} \frac{p_{ii}}{\sum_{j=1}^{k} p_{ij}} \tag{6}$$

The mean pixel accuracy represents the average ratio of the number of correctly classified pixel points per category and the number of all pixel points in that category.

The mean intersection over the union can be calculated with the formula 7.

$$mIoU = \frac{1}{k}\sum_{i=1}^{k} \frac{p_{ii}}{\sum_{j=1}^{k} p_{ij} + \sum_{j=1}^{k} p_{ji} - p_{ii}} \tag{7}$$

The mean intersection over union represents the average intersection ratios for each category.

4.1 Experimental results

FCN-32s, FCN-16s, and FCN-8s are trained by 1967 UAV remote sensing images. The model parameters are updated using the validation set during the training process, and the loss functions, PA, and mIoU of training and validation are shown in Table 1. The final trained optimal models are used for semantic segmentation prediction of 648 UAV remote sensing images. The PA, mIoU, and mPA are shown in Table 2; the training, validation, and prediction accuracy of 12 categories are shown in Table 3. The results show that the mPA of all three full convolutional neural network structures can be achieved with a rate above 90%. Among all categories, the segmentation accuracy is higher for vegetation, road, and people and lower for animals and boats. By analyzing the training images of the dataset, it can be found that the image number of animals and boats is extremely small, which is the reason for their poor segmentation results.

TABLE 1 The accuracy evaluation indicators of the training and validation datasets.

Network	Training			Validation		
	Loss %	PA%	mIoU %	Loss %	PA%	mIoU %
FCN-32s	0.564	65.730	94.358	17.848	53.222	58.631
FCN-16s	0.613	65.794	94.541	15.816	54.446	62.621
FCN-8s	1.082	65.497	92.898	14.739	54.387	62.418

TABLE 2 The accuracy evaluation indicators of the test dataset.

Network	PA%	mIoU%	mPA%
FCN-32S	35.271	45.509	91.562
FCN-16S	37.853	44.521	91.607
FCN-8S	36.313	43.120	91.200

Due to a large number of predicted pictures, they cannot be displayed entirely. The paper selected ten typical scene pictures as sample data and arranged them in the order of original image, labeled image, FCN-8s prediction image, FCN-16s prediction image, and FCN-32s prediction image for display, as shown in Figure 5.

4.2 Discussion on experimental results

The semantic segmentation of UAV remote sensing images using deep learning is undoubtedly fast and effective, but some problems still deserve further study.

(1) According to the general experimental conclusion, the segmentation effect of FCN-8s should be significantly higher than that of FCN-32s. Still, from the results of Table 2, this is not the case, which may be related to the resolution of our dataset. The author chooses a crop size closest to the resolution of the original image for this experiment. In future research, the researcher can consider a more optimal crop treatment that makes the number of datasets larger and reflects the differences between various models more obviously.

(2) As shown in Table 2, it can be found that the overall segmentation effect of the categories of animals and boats is poor, which is caused by the imbalance of data categories in the AeroScapes dataset, where there are fewer images in these two categories. The category imbalance problem is especially obvious in the detection and segmentation tasks and often requires special attention. In the future, if continuing to use the public dataset, which is no longer able to change its internal results, the author can consider using weights to control the category balance. Those with more category data should have smaller weights to reduce the impact on the overall

TABLE 3 The accuracy evaluation indicators for each category in the test dataset.

Category	Network	Training accuracy%	Validation accuracy %	Test accuracy %
Person	FCN-32s	97.269	76.878	52.674
	FCN-16s	97.433	81.793	54.686
	FCN-8s	96.921	81.635	56.772
Bike	FCN-32s	82.967	46.656	10.873
	FCN-16s	81.844	46.369	15.369
	FCN-8s	80.392	47.730	13.848
Car	FCN-32s	71.140	58.594	38.622
	FCN-16s	73.215	59.028	47.340
	FCN-8s	71.138	58.795	40.827
Drone	FCN-32s	33.753	19.198	13.000
	FCN-16s	32.914	21.207	18.789
	FCN-8s	32.770	19.484	15.954
Boat	FCN-32s	5.956	3.868	2.397
	FCN-16s	5.950	5.549	2.866
	FCN-8s	5.921	3.994	2.760
Animal	FCN-32s	—	—	—
	FCN-16s	—	—	—
	FCN-8s	—	—	—
Obstacle	FCN-32s	95.337	60.213	22.732
	FCN-16s	96.060	66.266	29.772
	FCN-8s	95.923	65.356	20.178
Construction	FCN-32s	95.010	67.319	31.460
	FCN-16s	92.949	67.421	35.959
	FCN-8s	94.974	69.510	35.869
Vegetation	FCN-32s	99.813	96.194	95.903
	FCN-16s	99.789	96.102	93.882
	FCN-8s	99.738	96.076	94.191
Road	FCN-32s	99.873	97.827	86.388
	FCN-16s	99.867	98.246	85.989
	FCN-8s	99.634	98.156	88.054
Sky	FCN-32s	41.911	58.696	33.932
	FCN-16s	43.712	56.921	31.729
	FCN-8s	43.059	57.520	30.989
Background	FCN-32s	99.486	89.135	75.615
	FCN-16s	99.421	89.819	78.092
	FCN-8s	99.186	90.011	77.814

model. If the authors use their dataset, it is necessary to ensure the category balance, either trying to control the category balance when collecting data or considering adding algorithms such as adversarial generative networks to solve the small sample data problem and obtain better semantic segmentation results.

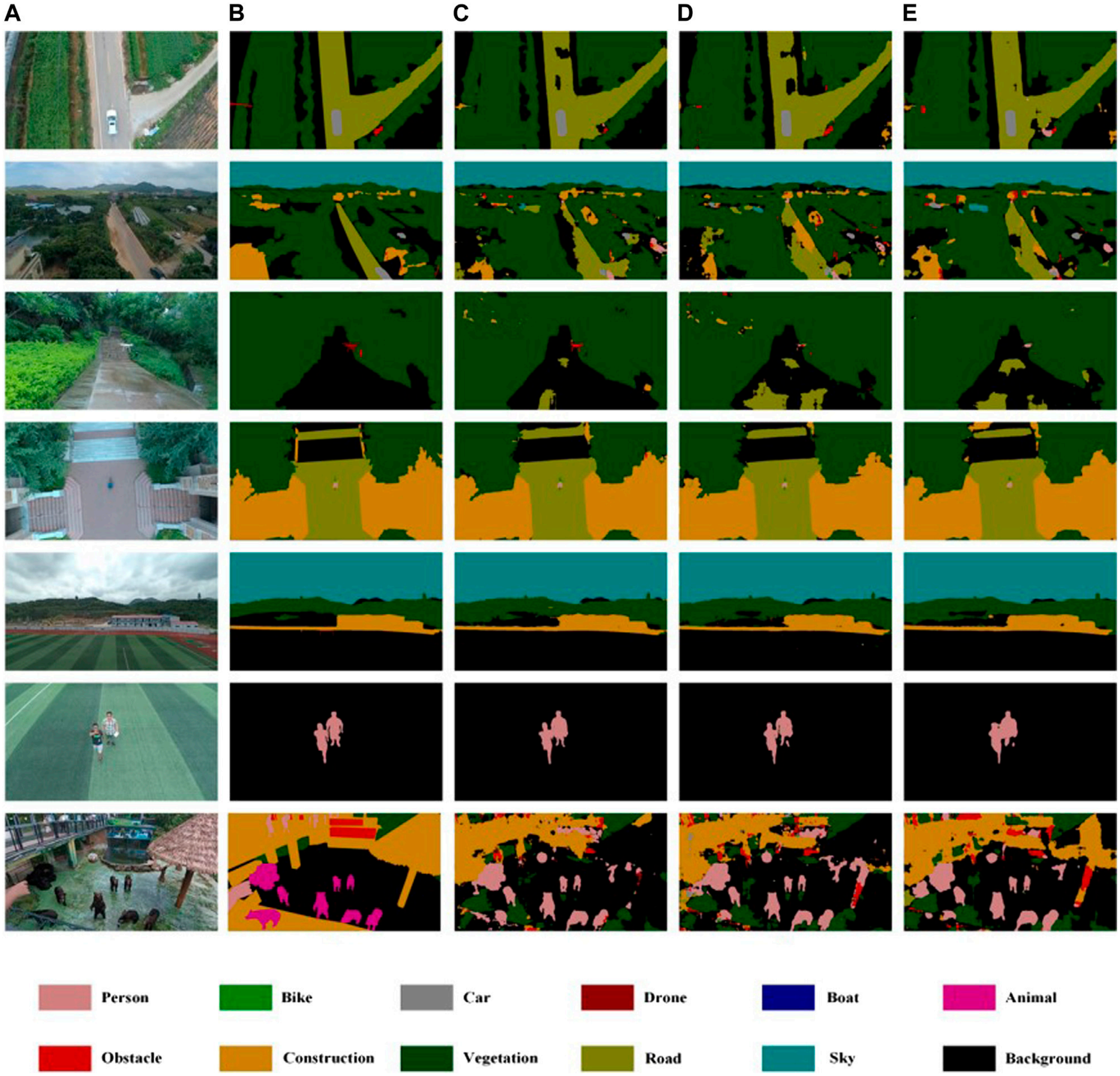

FIGURE 5
Semantic segmentation results in the AeroScapes dataset with FCN-8s, FCN-16s, and FCN-32s, respectively. Different colors represent different labels. **(A)** Image; **(B)** Label; **(C)** FCN-8s; **(D)** FCN-16s; **(E)** FCN-32s.

(3) From the segmentation results of the last two images in Figure 4, it can be easily seen that the animals are not properly labeled, and the same colors are used as persons, which is a common problem of inconsistent scales and small differences between categories in remote sensing images. When the UAV is flying at a low altitude, the images of animals captured are about the same size as the images of persons captured when the UAV is flying at a high altitude. At the same time, the animals standing up have a very high similarity to the appearance of persons, and the differences between the two categories are small, so the segmentation results are wrong. How to deal with this kind of problem needs further study.

(4) During the experiments, only label coding, center cropping, and normalization were done on the dataset, and the dataset was not expanded by data augmentation. In future studies, geometric enhancements such as horizontal flipping, random cropping and scaling, random mirroring, or texture enhancements such as adjusting brightness and contrast can be considered to expand the dataset to make the trained segmentation model have better robustness and generalization performance.

5 Conclusion

This paper has researched the implementation method and process of semantic segmentation of UAV remote sensing images using three fully convolutional neural network structures. And the author has used three models, FCN-32s, FCN-16s, and FCN-8s, respectively, and trained 1,967 images, validated 654 images, and predicted 648 images. The experimental results show that setting the appropriate batch size and initial learning rate value for the fully convolutional neural network and choosing the Adam optimizer can segment the UAV remote sensing images effectively, and the segmentation results of FCN-16s and FCN-8s are better. Compared with traditional semantic segmentation methods such as region-based and SVM methods, deep learning-based segmentation methods do not depend on the quality of features extracted by domain experts, and can solve the problem of automatic feature learning, which is bound to improve the efficiency and accuracy of semantic segmentation significantly. Meanwhile, with the gradual popularization of UAV equipment, it is also easy to obtain high-resolution remote sensing images, which is more beneficial for us to apply deep learning to accomplish more valuable tasks in the image field.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding authors.

Author contributions

GZ was responsible for writing and method. HZ and XY were responsible for data analyzing. ZJ was responsible for review and proofreading.

Funding

This research was jointly supported by the Foundation of Jilin Provincial Science and Technology Department (202002029JC, 20210203103SF, 20200301045RQ), CCIT Science and Technology Project (320200012), and fund of Jiangxi Provincial Department of Science and Technology (20202ACBL214016, 20224ACB204021).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Chen, L. C., Papandreou, G., Kokkinos, I., Murphy, K., and Yuille, A. L. (2017a). DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. *IEEE Trans. pattern analysis Mach. Intell.* 40 (4), 834–848. doi:10.1109/tpami.2017.2699184

Chen, L. C., Papandreou, G., and Kokkinos, I. (2014). Semantic image segmentation with deep convolutional nets and fully connected crfs[J]. arXiv preprint arXiv:1412.7062.

Chen, L. C., Papandreou, G., and Schroff, F. (2017b). Rethinking atrous convolution for semantic image segmentation[J]. arXiv preprint arXiv:1706.05587.

Chen, L. C., Yang, Y., and Wang, J. (2016). “Attention to scale: Scale-aware semantic image segmentation[C],” in *Proceedings of the IEEE conference on computer vision and pattern recognition* (Las Vegas, USA: IEEE), 3640–3649.

Chen, L. C., Zhu, Y., and Papandreou, G. (2018). “Encoder-decoder with atrous separable convolution for semantic image segmentation[C],” in *Proceedings of the European conference on computer vision (ECCV)* (Munich, Germany: Spinger), 801–818.

Chen, X., Williams, B. M., and Vallabhaneni, S. R. (2019). “Learning active contour models for medical image segmentation[C],” in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. Long Beach, CA, USA, 15-20 June 2019 (IEEE) 11632–11640.

De Benedetti, M., D’Urso, F., Fortino, G., Messina, F., Pappalardo, G., and Santoro, C. (2017). A fault-tolerant self-organizing flocking approach for UAV aerial survey. *J. Netw. Comput. Appl.* 96, 14–30. doi:10.1016/j.jnca.2017.08.004

Feng, D., Haase-Schütz, C., Rosenbaum, L., Hertlein, H., Glaser, C., Timm, F., et al. (2020). Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges. *IEEE Trans. Intelligent Transp. Syst.* 22 (3), 1341–1360. doi:10.1109/tits.2020.2972974

Funk, F., and Stütz, P. (2017). “A passive cloud detection system for UAV: Weather situation mapping with imaging sensors[C],” in 2017 IEEE Aerospace Conference. Big Sky, MT, USA, 04-11 March 2017 (IEEE), 1–12.

Green, D. R., Hagon, J. J., and Gómez, C. (2019). “Using low-cost UAVs for environmental monitoring, mapping, and modelling: Examples from the coastal zone[M],” in *Coastal management* (United Stated: Taylor & Franci), p465–p501.

HuangWuPeng, L. X. Q., and Yu, X. (2021). Depth semantic segmentation of tobacco planting areas from unmanned aerial vehicle remote sensing images in plateau mountains. *J. Spectrosc.* 2021, 1–14. doi:10.1155/2021/6687799

Kamilaris, A., and Prenafeta-Boldú, F. X. (2018). Disaster monitoring using unmanned aerial vehicles and deep learning[J]. arXiv preprint arXiv, 1807.

Kingma, D. P., and Ba, J. (2014). Adam: A method for stochastic optimization[J]. arXiv preprint arXiv:1412.6980.

Li, J., Jiang, F., Yang, J., Kong, B., Gogate, M., Dashtipour, K., et al. (2021). Lane-DeepLab: Lane semantic segmentation in automatic driving scenarios for high-definition maps. *Neurocomputing* 465, 15–25. doi:10.1016/j.neucom.2021.08.105

Li, W., He, C., Fang, J., Zheng, J., Fu, H., and Yu, L. (2019a). Semantic segmentation-based building footprint extraction using very high-resolution satellite images and multi-source GIS data. *Remote Sens.* 11 (4), 403. doi:10.3390/rs11040403

Li, Y. D., Dong, H., Li, H. G., Zhang, X., Zhang, B., and Xiao, Z. (2020). Multi-block SSD based on small object detection for UAV railway scene surveillance. *Chin. J. Aeronautics* 33 (6), 1747–1755. doi:10.1016/j.cja.2020.02.024

Li, Y., Peng, B., He, L., Fan, K., and Tong, L. (2019b). Road segmentation of unmanned aerial vehicle remote sensing images using adversarial network with multiscale context aggregation. *IEEE J. Sel. Top. Appl. Earth Observations Remote Sens.* 12 (7), 2279–2287. doi:10.1109/jstars.2019.2909478

Liu, M., Wang, X., Zhou, A., Fu, X., Ma, Y., and Piao, C. (2020). UAV-YOLO: Small object detection on unmanned aerial vehicle perspective. *Sensors* 20 (8), 2238. doi:10.3390/s20082238

Liu, S., Cheng, J., Liang, L., Bai, H., and Dang, W. (2021). Light-weight semantic segmentation network for UAV remote sensing images. *IEEE J. Sel. Top. Appl. Earth Observations Remote Sens.* 14, 8287–8296. doi:10.1109/jstars.2021.3104382

Meenpal, T., Balakrishnan, A., and Verma, A. (2019). "Facial mask detection using semantic segmentation[C]," in 2019 4th International Conference on Computing, Communications and Security (ICCCS). Rome, Italy, 10-12 October 2019 (IEEE), 1–5.

Milioto, A., Lottes, P., and Stachniss, C. (2018). "Real-time semantic segmentation of crop and weed for precision agriculture robots leveraging background knowledge in CNNs[C]," in 2018 IEEE international conference on robotics and automation (ICRA). Brisbane, QLD, Australia, 21-25 May 2018 (IEEE), 2229–2235.

Minaee, S., Boykov, Y., Porikli, F., Plaza, A., Kehtarnavaz, N., and Terzopoulos, D. (2021). "Image segmentation using deep learning: A survey," in IEEE Transactions on Pattern Analysis and Machine Intelligence (IEEE) 44 (7), 3523–3542. doi:10.1109/TPAMI.2021.3059968

Mohammadimanesh, F., Salehi, B., Mahdianpari, M., Gill, E., and Molinier, M. (2019). A new fully convolutional neural network for semantic segmentation of polarimetric SAR imagery in complex land cover ecosystem. *ISPRS J. photogrammetry remote Sens.* 151, 223–236. doi:10.1016/j.isprsjprs.2019.03.015

Shelhamer, E., Long, J., and Darrell, T. (2017). Fully convolutional networks for semantic segmentation. *IEEE Trans. Pattern Analysis Mach. Intell.* 39 (4), 640–651. doi:10.1109/tpami.2016.2572683

Tian, X., Wang, L., and Ding, Q. (2019). Review of image semantic segmentation based on deep learning[J]. *J. Softw.* 30 (2), 440–468. doi:10.13328/j.cnki.jos.005659

Xu, Y., Yu, G., and Wang, Y. (2017). Car detection from low-altitude UAV imagery with the faster R-CNN[J]. *J. Adv. Transp.* 2017, 2823617. doi:10.1155/2017/2823617

Yang, J., Ding, Z., and Wang, L. (2021). The programming model of air-ground cooperative patrol between multi-UAV and police car. *IEEE Access* 9, 134503–134517. doi:10.1109/access.2021.3115950

Yang, R., and Yu, Y. (2021). Artificial convolutional neural network in object detection and semantic segmentation for medical imaging analysis. *Front. Oncol.* 11, 638182. doi:10.3389/fonc.2021.638182

Zhang, H., Wang, L., Tian, T., and Yin, J. (2021). A review of unmanned aerial vehicle low-altitude remote sensing (UAV-LARS) use in agricultural monitoring in China. *Remote Sens.* 13 (6), 1221. doi:10.3390/rs13061221

Zhang, Y., Yuan, X., Li, W., and Chen, S. (2017). Automatic power line inspection using UAV images. *Remote Sens.* 9 (8), 824. doi:10.3390/rs9080824

Zhao, H., Shi, J., and Qi, X. (2017). "Pyramid scene parsing network[C]," in *Proceedings of the IEEE conference on computer vision and pattern recognition* (Hawaii, USA: IEEE), 2881–2890.