



A Multi-Objective Semantic Segmentation Algorithm Based on Improved U-Net Networks

Xuejie Hao ^{1,2,†} , Lizeyan Yin ^{3,†}, Xiuhong Li ², Le Zhang ¹ and Rongjin Yang ^{1,*}

¹ State Key Laboratory of Environmental Criteria and Risk Assessment, Chinese Research Academy of Environmental Sciences, No. 8, Da Yang Fang, An Wai, Chao Yang District, Beijing 100012, China

² State Key Laboratory of Remote Sensing Science, College of Global Change and Earth System Science, Beijing Normal University, No. 19, Xijiekou Wai Street, Haidian District, Beijing 100875, China

³ Institute of Computing, Modeling and Their Applications, ISIMA, University Clermont Auvergne, 63000 Clermont-Ferrand, France

* Correspondence: yangrj@craes.org.cn; Tel.: +86-136-2116-6693

† These authors contributed equally to this work and should be considered co-first authors.

Abstract: The construction of transport facilities plays a pivotal role in enhancing people's living standards, stimulating economic growth, maintaining social stability and bolstering national security. During the construction of transport facilities, it is essential to identify the distinctive features of a construction area to anticipate the construction process and evaluate the potential risks associated with the project. This paper presents a multi-objective semantic segmentation algorithm based on an improved U-Net network, which can improve the recognition efficiency of various types of features in the construction zone of transportation facilities. The main contributions of this paper are as follows: A multi-class target sample dataset based on UAV remote sensing and construction areas is established. A new virtual data augmentation method based on semantic segmentation of transport facility construction areas is proposed. A semantic segmentation model for the construction regions based on data augmentation and transfer learning is developed and future research directions are given. The results of the study show that the validity of the virtual data augmentation approach has been verified; the semantic segmentation of the transport facility model can semantically segment a wide range of target features. The highest semantic segmentation accuracy of the feature type was 97.56%.



Citation: Hao, X.; Yin, L.; Li, X.; Zhang, L.; Yang, R. A Multi-Objective Semantic Segmentation Algorithm Based on Improved U-Net Networks. *Remote Sens.* **2023**, *15*, 1838. <https://doi.org/10.3390/rs15071838>

Academic Editors: Andrea Garzelli and Gwanggil Jeon

Received: 4 March 2023

Revised: 24 March 2023

Accepted: 29 March 2023

Published: 30 March 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Keywords: semantic segmentation; U-Net; data augmentation; virtual sample; construction of transport facilities

1. Introduction

Transport facilities are necessary means of transport (including ships, vehicles and aircraft), land, communications, buildings (consisting of warehouses, ticket offices, stations and waiting areas), machinery and equipment, lines, signage, etc. Transport facilities are a golden gateway to guide an industrial layout, promote the development of rural areas along a route and integrate tourism resources. The construction of transport facilities is a great driving force for economic and social development, and at the same time it can meet the needs of a national defense, guarantee social stability and greatly promote an improvement in people's livelihood. The construction of transport facilities always involves complex and difficult construction conditions and many high-risk hazards. Safety, environmental, schedule and organisational risks are particularly prominent. For construction safety and risk avoidance, it is essential to identify the various target features in the construction area.

The best method for specifying multiple types of target features in the construction area is semantic segmentation. Semantic segmentation is a complex and challenging task in computer vision, requiring advanced algorithms and techniques in order to accurately identify and label the objects in an image. In this process, a set of raw data are taken as input and processed to produce a mask that highlights regions of interest within the image.

Semantic segmentation is a particularly complex and challenging problem in the field of computer vision. Unlike traditional image classification techniques, which aim to classify an entire image into a single category, semantic segmentation requires a more granular approach, where each individual pixel within an image must be classified into different classes or objects. This process requires a model that understands the intricacies of complex images and is capable of recognising fine-grained details and spatial relationships between adjacent pixels. In short, the goal of semantic segmentation is to accurately segment the image into different classes, such as people, cars, buildings and other objects. Therefore, there is a need for sophisticated algorithms that can analyze and interpret the complex visual data contained within the image. In this process, the image is decomposed into smaller regions or segments and then each segment is assigned a label based on its visual features. In general, semantic segmentation is an essential task in computer vision that has a broad range of applications in areas, such as autonomous driving, robotics and medical imaging. This is a challenging task requiring a thorough understanding of image processing and machine learning techniques, and researchers are continually investigating new techniques and approaches to improve the accuracy and effectiveness of semantic segmentation algorithms [1]. There are two main categories of semantic segmentation algorithms: traditional methods and deep learning methods [2,3]. Traditional semantic segmentation methods involve two processes: first, feature extraction from the image; then, classification of the image pixels. Feature extraction can be accomplished through manual tagging; however, this method is reliant on the researcher's expertise. Representative algorithms of pixel classification methods include Support Vector Machine (SVM) [4] and Random Forest (RF) [5]. However, traditional semantic segmentation methods have the disadvantages of poor segmentation effects and low segmentation efficiency [2]. Due to the unparalleled success of deep learning techniques in various computer vision and image processing tasks, researchers were driven to extend their superior performance to the field of semantic segmentation as well. The application of deep learning models to semantic segmentation has resulted in significant advancements in this field, making it possible to attain more accurate and efficient segmentation of complex images. The main classical semantic segmentation algorithms based on deep learning are FCN [6], SegNet [7], U-Net [8], DeepLab [9–12] and PSPNet [13].

To fully extract the effective features, strengthen the fusion of the different semantic information, and to solve problems, such as gradient disappearance and overfitting, many researchers have made many improvements based on the U-Net network model. These improvements include performance-based optimisations and module structure-based optimisations. The improvements to optimise performance include: (1) Expanding from 2D to 3D images; (2) strengthening relevant features and weakening irrelevant features; (3) improving the computation speed and memory usage; (4) an improved feature fusion method; (5) improvements in data augmentation methods; and (6) improvements in the generalisation ability. Improvements to the module structure include: (1) Encoder and decoder structure improvements; (2) improvements to the loss function; (3) bottleneck module structure improvements; (4) data flow path improvements; and (5) improvements to the automatic structure search [14]. These enhancements improve the precision and accuracy of the networks to some extent and further improve the segmentation performance of the networks. Among them, in an improvement to the data augmentation methods, Ronneberger O et al. [8] proposed a network and training strategy that relies on data augmentation to make more efficient use of the available annotated samples. Nalepa J et al. [15] proposed a new augmentation technique based on image registration to benefit from the subtle spatial and organisational features captured in the training set. The new augmentation technique is used to extend the dataset for training U-net-based deep networks. The results show that the new augmentation technique achieves performance gains without sacrificing computational speed. Nishio M et al. [16] proposed an improved framework for the U-Net network based on the data augmentation method for automatic pancreas segmentation of Computed Tomography images. Jin Y W et al. [17] further improved the diagnostic potential

of convolutional neural networks (CNNs) for lymph node metastasis detection in breast cancer patients by comprehensively enhancing the input images of multiple segmentation channels. Uysal E S et al. [18] improved the performance of the U-Net model based on extensive data augmentation. Aboudi F et al. [19] proposed two data augmentation methods (AutoReplace and AutoMove) to solve the data efficiency problem. Sfakianakis C et al. [20] developed an important data augmentation method based on medical practices using the convolutional neural network of the U-Net architecture. The study shows that the proposed method has an overall improvement in segmentation accuracy as well as in the estimation of clinical indicators.

Although there are many improved U-Net networks based on data augmentation methods, these data augmentation methods are not suitable for the data in this study. To address the challenges of inadequate training samples and imbalanced training datasets in semantic segmentation, this study proposes a multi-objective semantic segmentation algorithm based on an improved U-Net network. The proposed algorithm aims to enhance the segmentation effect and accuracy of semantic segmentation by incorporating multiple objectives into the segmentation process. However, this study is different from the task of land cover segmentation [21,22]. Firstly, transportation facility construction sites pose unique challenges for land cover segmentation that may not be present in other types of land cover segmentation tasks. For example, transportation facility construction sites often involve complex geometries and frequently change over time, which can make accurate segmentation more difficult. Additionally, the types of land cover within transportation facility construction sites may differ from other types of land cover, such as slag dumps. Secondly, transportation facility construction sites have a significant impact on local communities, including disruptions to traffic and potential environmental impacts. Accurate and timely monitoring of land cover changes within these sites can help mitigate these impacts and ensure compliance with regulations.

This paper presents several significant contributions to the field of transport facility planning and management. Firstly, a multi-class target sample dataset was created using remote sensing data from Unmanned Aerial Vehicles (UAVs) and images of transport construction sites. This comprehensive dataset includes different classes of objects, such as roads, construction land, vegetation and water bodies, which are essential for the planning and management of transport facilities. Secondly, this paper proposes a novel virtual data augmentation method based on semantic segmentation of the construction area of the transport facilities. This method generates new virtual data by applying semantic segmentation to the existing images, resulting in an increased number of training samples for deep learning models. This method can complement traditional data augmentation techniques and improve the performance of segmentation models, especially in cases where training datasets are limited or unbalanced. Thirdly, this paper establishes a Semantic Segmentation of Traffic Facilities Model (SSTFM) using data augmentation and transfer learning. The SSTFM uses the proposed dataset and data augmentation methods to train a deep learning model for the semantic segmentation of traffic facility construction areas. The model achieves a high accuracy and efficiency in segmentation, demonstrating the effectiveness of the proposed dataset and the data augmentation method. Finally, this paper provides future research directions for the development of advanced computer vision-based solutions for transport facility planning and management. The proposed dataset, data augmentation method and the SSTFM can be extended and applied to other fields, such as urban planning and environmental monitoring, where similar segmentation tasks need to be performed. Overall, this paper provides valuable insights into the establishment of comprehensive datasets, innovative data augmentation methods and advanced deep learning models for accurate and efficient segmentation in transport facility planning and management.

2. Materials and Methods

This section describes the dataset in detail and describes several methods used to optimize the U-Net networks. The workflow of the SSTFM is shown in Figure 1. The model consists of six modules: (1) dataset; (2) data processing; (3) data augmentation; (4) transfer learning; (5) verification; and (6) model evaluation.

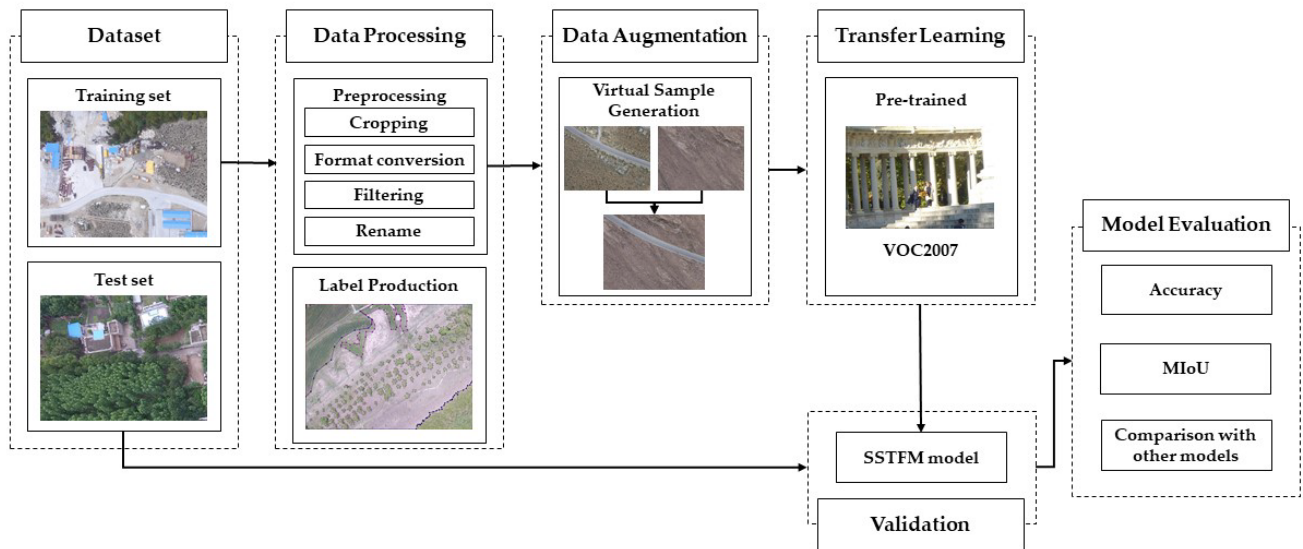


Figure 1. The workflow of the SSTFM.

2.1. Study Area and Datasets

The research area is located at the confluence of the primary and secondary plateaus in China, covering a vast and complex mountainous region with varied topography. The area is characterized by a lack of resources, inaccessible transport and communications and a sparse population. The geographical environment is extremely dangerous, the geological conditions extremely complex and the ecological environment extremely fragile.

The dataset for this study is the image data of the study area captured by the UAV. The resolution of the images is 4000×3000 (pixels). The bit depth is 24. The raw data set is 100 images in total. The nine types of target features commonly found in transport construction scenarios are shown in Figure 2.

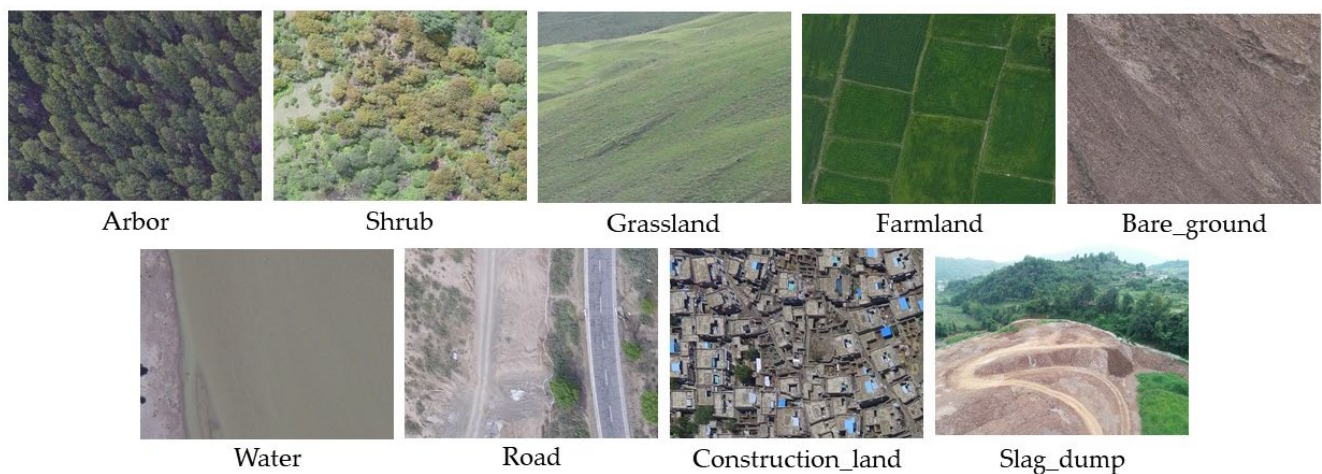


Figure 2. Nine types of target features are commonly found in transport facility construction scenarios.

2.2. Data Processing

Since the semantic segmentation model makes demands on the size and format of the input image, the raw data cannot be used directly. Operations, such as pre-processing, labelling and format conversion of the images are required.

2.2.1. Pre-Processing of Images

Image pre-processing involves several essential steps, including cropping, file format conversion, filtering and renaming the source images. First, the large image is cropped into a small image dataset of the same size, which is convenient for image labelling and training, and dataset A is obtained. Then, the format of the original image is converted to the format required for the model training, and dataset B is obtained. Filtering is performed by deleting the duplicate, incomplete and no research target information images in dataset B to obtain dataset C; finally, the image name is renamed to facilitate data management and processing to obtain the final dataset: dataset D. The filtering operation is the process of deduplication of massive images. In this study, the Locality Sensitive Hashing (LSH) is used for deduplication processing [23]. The LSH is a commonly used approximate nearest neighbour search algorithm, which is mainly used to solve similarity search problems in high-dimensional spaces. Compared with the traditional nearest neighbour search algorithm, LSH can find a vector similar to a given vector in a shorter time, thus improving the search efficiency.

2.2.2. Label Production

Once the dataset was determined, the label needs be defined. In practice, the label definition must be combined with the requirements of the project. From the perspective of whether there is supervision or not, deep learning can be divided into three types: supervised, semi-supervised and unsupervised. The main difference between unsupervised and supervised deep learning training is that unsupervised deep learning training does not require annotated data, while supervised training does. In addition, in semi-supervised deep learning training, part of the data are labelled and the other parts are unlabelled. The task of annotating the data requires a significant investment of time and effort. Ideally, the more annotated data available, the better the performance of the trained model. However, this is often not possible. To successfully complete a project, it is essential to balance the use of the available resources with the time constraints. It is also important to consider the impact of data volume on the effectiveness of the model, in line with the accuracy requirements of the project. It is important to combine these two points to obtain an appropriate value.

The effectiveness of the trained model is profoundly impacted by the quality of the annotated data. In general, the following methods are used to improve the quality of the annotation: the selection of experienced annotators; each piece of data being annotated and sampled by more than one person; and the use of a good annotation tool. The software used in this research is Labelme, which is an image annotation tool created by the Computer Science and Artificial Intelligence Laboratory at the Massachusetts Institute of Technology. People can use the tool to create custom annotation tasks or to perform image annotation.

2.3. Data Augmentation

Data augmentation [24] is the process of increasing the amount of data by applying various methods or techniques to the original data. This technique allows the current deep learning model to improve its performance by using a larger amount of data. In essence, data augmentation is a set of manual expansion methods applied to the original labelled training set.

The various data augmentation techniques can be divided into two groups based on their basic principles: data-based augmentation methods and network-based augmentation methods. Data-based augmentation techniques can be divided into four categories: one-sample transformation methods, multi-sample synthesis methods, deep generative model

methods and virtual sample generation methods. Network-based data augmentation approaches can be categorised as either network strategies or learning strategies. Figure 3 shows the data augmentation methods and their respective associations.

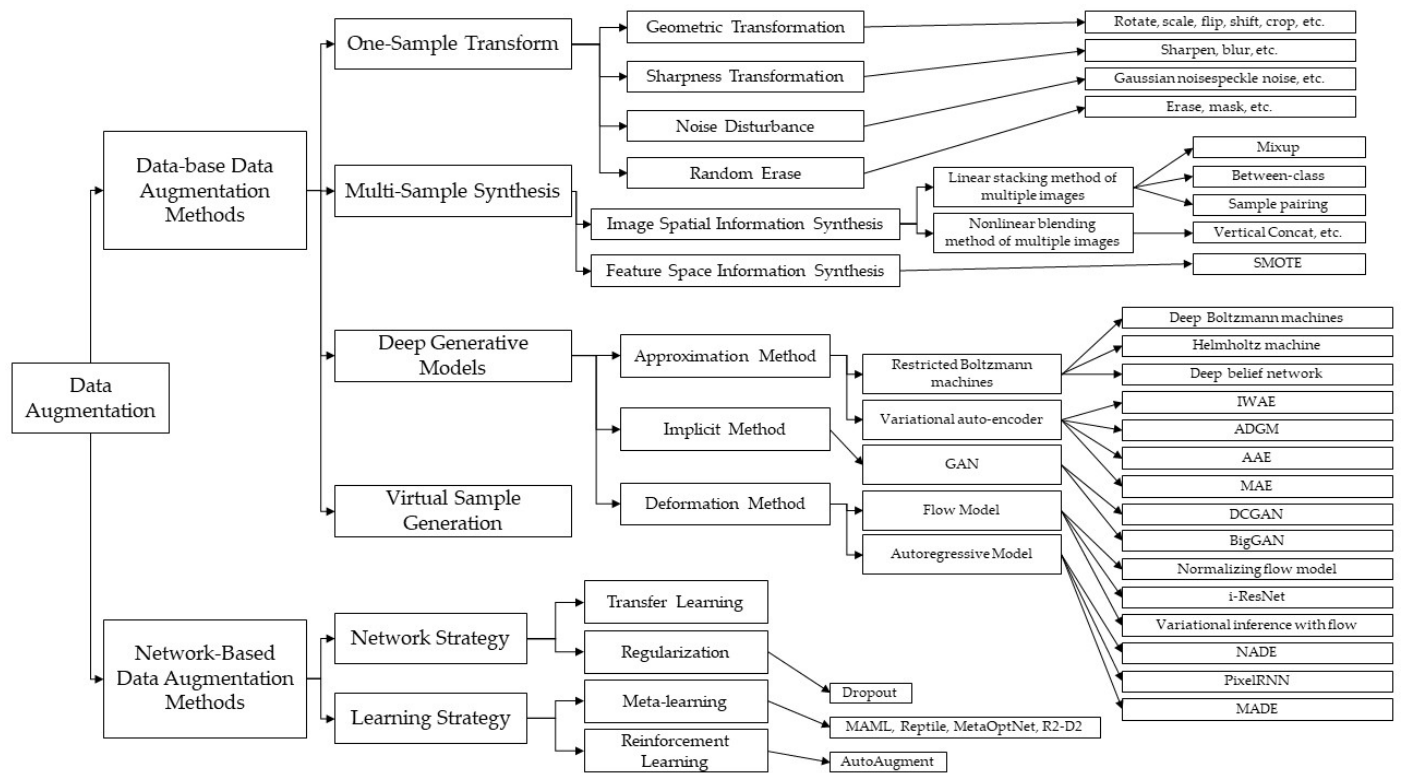


Figure 3. The data augmentation methods and the affiliation between the methods [25].

Data augmentation methods can be divided into two categories: data-based data augmentation methods and network-based data augmentation methods. Data-based data augmentation methods can be divided into methods, such as one-sample transformation, multi-sample synthesis, deep learning generation and virtual sample generation. The one-sample transformation method takes a single sample data as the operation object and changes the original data through geometric transformation (rotate, scale, flip, shift, crop, etc.), sharpness transformation (sharpen, blur, etc.), noise perturbation (gaussian noise, salt and pepper noise, speckle noise, etc.) and random erasure (erase, mask, etc.) to generate new data different from the original data. Multi-sample synthesis is the process of artificially mixing the information from multiple images to generate new data. It can be divided into image spatial information synthesis and feature space information synthesis (SMOTE). Image spatial information synthesis methods can be divided into two types: linear stacking methods of multiple images (Mixup, Between-Class and Sample Pairing) and nonlinear blending methods of multiple images (Vertical Concat, Horizontal Concat, Mixed Concat, Random 2×2 , VH-Mixup, etc.). From the perspective of the maximum likelihood function processing methods, deep generative models can be divided into three types: approximation method, implicit method and deformation method. The approximation method is used to obtain the approximate distribution of the likelihood function by variation or sampling, mainly including restricted Boltzmann machines (Deep Boltzmann machines, Helmholtz machine and Deep belief network) and variational autoencoders (Importance weighted autoencoders (IWAE), Auxiliary deep generative models (ADGM), Adversarial autoencoders (AAE) and Masked Autoencoders (MAE)). The implicit method is a method that avoids the maximum likelihood process, and its representative model is the generative confrontation network (GAN) and its deformation network Deep convolutional generative adversarial networks (DCGAN) and BigGAN. The deformation method is used

to properly deform the probability function. The purpose of deformation is to simplify the calculation, such methods include flow model (Normalizing Flow Model, i-ResNet and Variational Inference with Flow) and autoregressive models (Neural autoregressive distribution estimation (NADE), Pixel Recurrent Neural Network (PixelRNN) and Masked Autoencoder for Distribution Estimation (MADE)). Data augmentation methods based on network strategies are mainly divided into network strategies and learning strategies. Representative methods of network strategies are transfer learning and regularization (Dropout). The most representative methods of learning strategies are meta-learning (MAML, Reptile, MetaOptNet and R2-D2) and reinforcement learning (Auto Augment) [25].

Among the data augmentation methods, transfer learning and one-sample transformation are the most appropriate data augmentation methods for this study. The one-sample transformation method is simple to use but the amount of information it extends is limited. Augmenting the data by increasing the frequency of the repetitions or applying incorrect manipulations can lead to changes in the original semantic annotations of the image [25]. Therefore, this paper proposes a virtual sample data augmentation approach. The flow chart of this method is shown in Figure 4. The virtual sample generation process is as follows: First, the multi-view target objects in the original image are extracted by matting technology, and the target image set A' is obtained. Then, the original background image set B' , containing the research area is selected for easy marking. Finally, the images in set A' and set B' are randomly combined to obtain the virtual sample dataset C' . An example of a virtual sample generation diagram is shown in Figure 5.

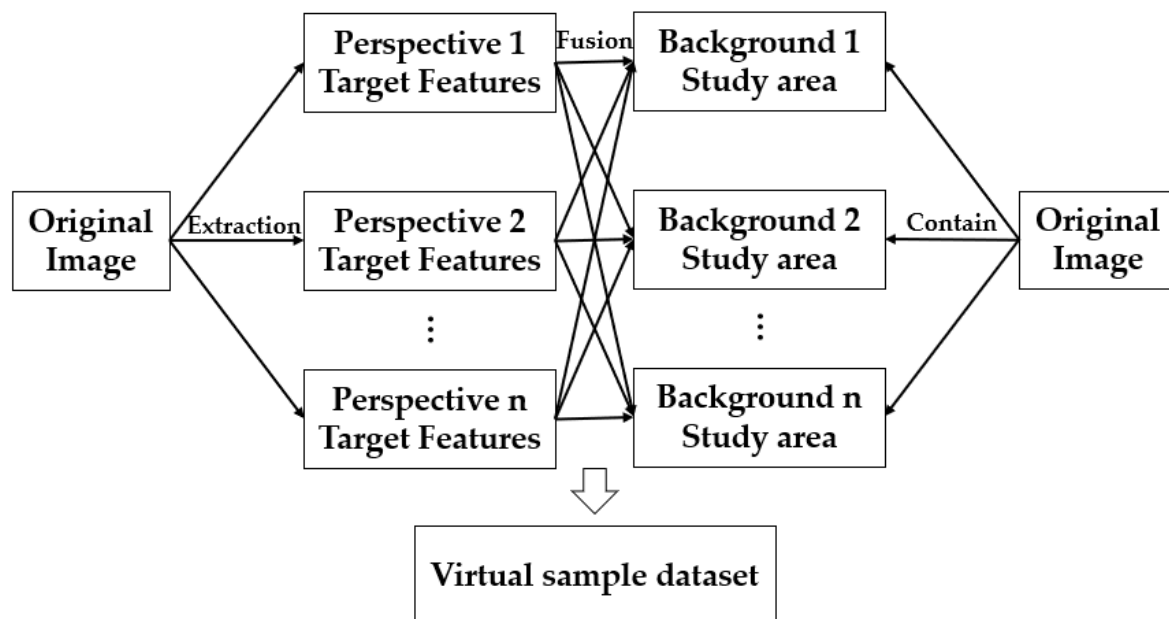


Figure 4. Workflow diagram of the virtual sample data augmentation method.

In Figure 5, the original image with the road recognition target is extracted from the existing dataset. Image extraction techniques are used to extract the roads in the original image for backup. A single image with a single background was selected as the original background image in the existing dataset. In Figure 5, the background image used is bare ground, but of course the background image could be grass or farmland. However, the background image must not be of water, as it would be a common-sense mistake to mix the road with the water. It is common sense that there are no roads over water, only bridges. Finally, the roads extracted from the original image and the background image selected from the original image are merged to form a new virtual sample image. Similarly, any virtual sample required for the study can be created following the steps above.

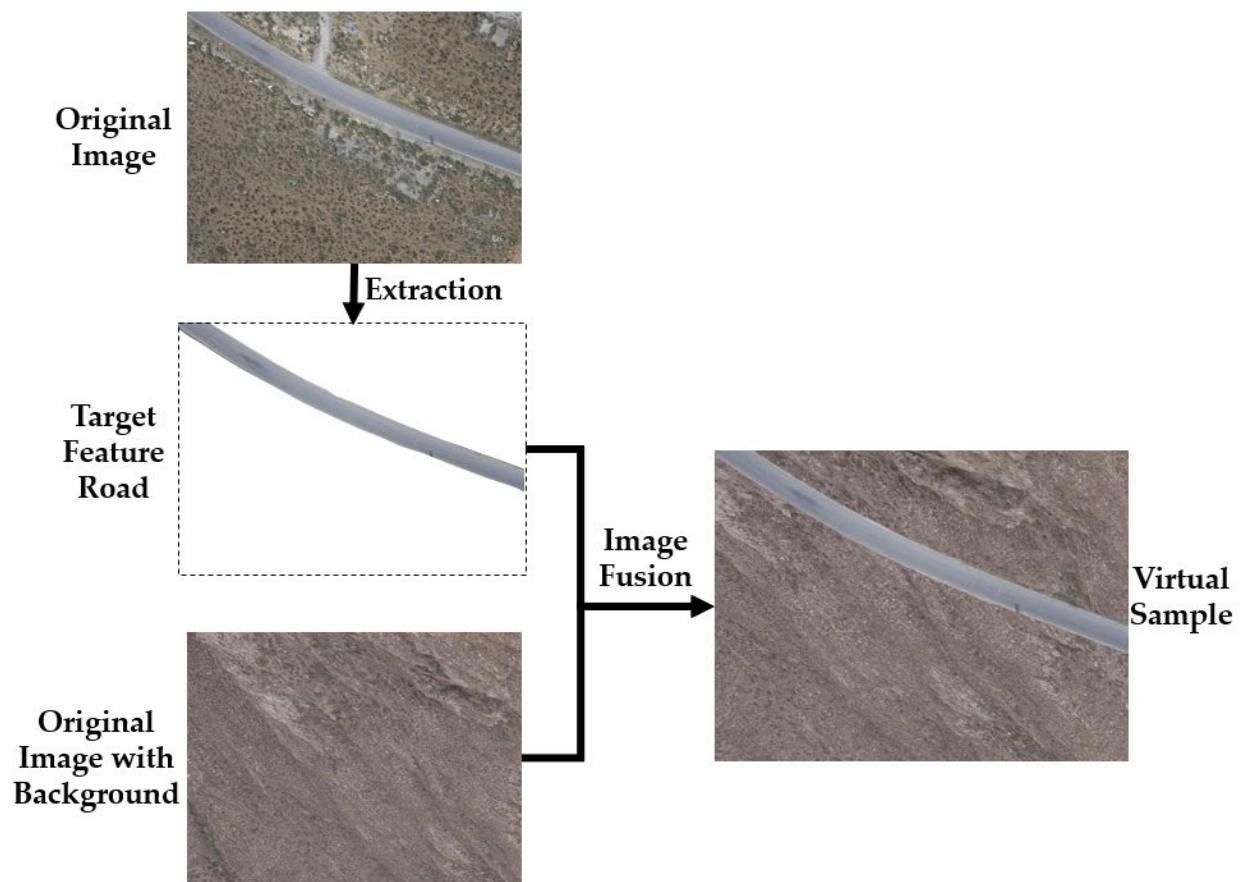


Figure 5. A virtual sample generation process.

2.4. Transfer Learning

As discussed in the previous section, transfer learning can serve as a data augmentation technique [25]. Transfer learning is a machine learning technique that involves transferring knowledge from one domain to another. This approach allows the target domain to achieve improved learning outcomes. In other words, transfer learning is the use of strategies that have been used to solve a problem by transferring existing experiences.

There are many application areas for semantic segmentation but there are many limitations due to problems, such as sparse training samples and diverse sample features. These limitations are the main problems in this study. At this point, transfer learning is needed to address these problems. In this study, a deep learning model trained on the VOC2007 dataset is used as a pre-trained model. Finally, the pre-trained model is fine-tuned on the training set.

2.5. Model Evaluation

In this paper, indices, such as IoU, Recall, Precision, MIoU, MPA and Accuracy were used as evaluation criteria to assess the SSTFM. When evaluating the results of semantic segmentation, the predicted results are commonly categorized into four sections: True Positive (TP), False Positive (FP), True Negative (TN) and False Negative (FN). The calculation equation for Precision is Formula (1); the calculation equation for Accuracy is Formula (2); and the calculation equation for Recall is Formula (3):

$$\text{Precision} = \frac{TP}{TP + FP} \quad (1)$$

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (2)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (3)$$

The Intersection over Union (IoU) metric is the ratio of the intersection between two sets and has been widely adopted as a standard measure for evaluating semantic segmentation tasks. The calculation equation of IoU is Formula (4):

$$\text{IoU} = \frac{\text{target} \cap \text{prediction}}{\text{target} \cup \text{prediction}} \quad (4)$$

Assume that there are $k + 1$ classes, where k represents the range from L_0 to L_k , and 1 represents the inclusion of an empty class or background. P_{ij} denotes the number of pixels that would have belonged to class i but are predicted to be class j . P_{ii} denotes the true number. In contrast, P_{ij} and P_{ji} are interpreted as false positive and false negative, respectively, although both are the sum of false the positives and false negatives. The Pixel Accuracy (PA) is the simplest evaluation metric, as it represents the percentage of accurately classified pixels out of the total number of pixels in an image. The Mean Pixel Accuracy (MPA) is a more refined version of the PA, as it calculates the proportion of the correctly classified pixels for each class, and then calculates the average across all the classes. The Mean Intersection over Union (MIoU) is a widely accepted metric for evaluating the effectiveness of semantic segmentation models. The calculation equations for PA, MPA and MIoU are Formulas (5)–(7):

$$\text{PA} = \frac{\sum_{i=0}^k P_{ii}}{\sum_{i=0}^k \sum_{j=0}^k P_{ij}} \quad (5)$$

$$\text{MPA} = \frac{1}{k+1} \sum_{i=0}^k \frac{P_{ii}}{\sum_{j=0}^k P_{ij}} \quad (6)$$

$$\text{MIoU} = \frac{1}{k+1} \sum_{i=0}^k \frac{P_{ii}}{\sum_{j=0}^k P_{ij} + \sum_{j=0}^k P_{ji} - P_{ii}} \quad (7)$$

3. Results

This section focuses primarily on the experimental setup, the results obtained and the subsequent discussion and analysis of the results.

3.1. Experimental Environment

All the experiments in this study were performed on a GPU with an Intel(R) Core(TM) i7—10700K CPU @ 3.80 GHz and a CPU with an NVIDIA GeForce RTX 2080 SUPER. The details of the environment configuration are shown in Table 1.

Table 1. Environment configuration.

Environment	Version
Operating system	Windows 10 (64-bit)
Framework	Pytorch 1.2.0
Memory	16G
GPU	NVIDIA GeForce RTX 2080 SUPER
CPU	Intel(R) Core(TM) i7—10700K CPU @ 3.80 GHz

In the training phase, the total training set was 2700 images. The total verification set was 300 images. A total of 135,000 iterations of the semantic segmentation model were trained. The maximum learning rate of the model is set to 1×10^{-4} , and the minimum learning rate is set to $(1 \times 10^{-4}) \times 0.01$. Then, the decrease in the learning rate takes the cosine type. The batch size is set to 2. The optimizer type is Adam. The momentum is 0.9. The number of classes is set to 10 (arbor, shrub, grass_land, farmland, bare_ground,

water, road, construction_land, slag_dump and background). The backbone network selects vgg. The training is divided into two phases: the freeze phase and the thaw phase. Freezing training can speed up training, the freeze_layers is set to 17. The freeze phase is set to meet the training needs of an insufficient machine performance. In our experiments, Freeze_Epoch = 50 and UnFreeze_Epoch = 100. According to the performance of the graphics card of the machine, the batch_size in the experiment is set to two, with five epochs save the weight once. The model evaluation takes a long time. To ensure the speed of training, five epochs are evaluated once.

3.2. Experimental Results

3.2.1. Results of the Introduction of Data Augmentation Methods

The first series of experiments tested the effectiveness of the virtual sample data augmentation methods and transfer learning. Firstly, this experiment was based on 1500 images used as a training set (T1) to obtain the model SSTFM-1. The SSTFM-1 segmented the nine target categories and the accuracy of the model was 76.95%. Among them, water had the best precision of 89.7% and the slag dump had the worst precision of 0. The low precision rate can be attributed to the inadequate number of training samples and the imbalanced distribution of the training categories. In this study, 500 training samples of farmlands and 1000 training samples of slag dumps were generated by the virtual sample data augmentation method and added to T1 to obtain the training set T2. The SSTFM-2 was obtained by training the model on the training set T2. The accuracy of the model was 85.22% after the semantic segmentation of the nine types of objects. Among them, farmland had the best precision of 97.33% and the construction land had the worst precision of 77.68%. The descriptions of the four SSTFM models are shown in Table 2. The detailed results are shown in Tables 3 and 4.

Table 2. The descriptions of the four SSTFM models.

	U-Net	DeepLab v3+	PSPNet	Data Augmentation	Transfer Learning	Dataset
SSTFM-1	✓					T1
SSTFM-2	✓			✓		T2
SSTFM-3	✓			✓	✓	T2
SSTFM-4		✓		✓	✓	T2
SSTFM-5			✓	✓	✓	T2

Table 3. A comparison of the values of each type of indicator in the model.

		Arbor	Shrub	Grassland	Farmland	Bare_Ground	Water	Road	Construction_Land	Slag_Dump
IoU (%)	SSTFM-1	63.75	65.73	58.25	46.9	60.26	68.99	52.7	74.16	0
	SSTFM-2	82.96	67.15	70.65	93.06	70.29	73.45	67.65	63.96	68.9
	SSTFM-3	83.83	69.53	72.69	93.88	73.63	76.95	70.65	69.09	71.35
	SSTFM-4	72.44	61.83	63.87	87.48	60.23	41.12	51.82	45.09	64.18
	SSTFM-5	83.03	65.23	68.44	88.20	67.37	50.94	51.81	60.41	67.32
Recall (%)	SSTFM-1	73.91	88.51	65.88	65.73	73.21	74.93	64.34	85.89	0
	SSTFM-2	90.01	80.95	83.88	95.5	85.7	83.4	79.62	78.36	78.54
	SSTFM-3	90.66	83.07	84.56	96.13	87.83	86.8	80.24	82.51	80
	SSTFM-4	81.99	79.77	74.1	92.93	85.92	48.1	61.07	57.96	74.09
	SSTFM-5	86.65	77.09	87.97	96.07	81.83	61.04	66.17	71.34	78.60
Precision (%)	SSTFM-1	82.26	71.86	83.41	62.08	77.31	89.7	74.45	84.46	0
	SSTFM-2	91.37	79.75	81.75	97.33	79.63	86.03	81.81	77.68	84.87
	SSTFM-3	91.75	81	83.81	97.56	82	87.15	85.52	80.94	86.85
	SSTFM-4	86.14	73.32	82.24	93.72	66.82	73.9	77.39	67.01	82.76
	SSTFM-5	95.21	80.92	75.51	91.51	79.22	75.49	70.47	79.76	82.42
MIoU (%)	SSTFM-1	64	66	58	47	60	69	53	74	0
	SSTFM-2	83	67	71	93	70	73	68	64	69
	SSTFM-3	84	70	73	94	74	77	71	69	71
	SSTFM-4	72	62	64	87	60	41	52	45	64
	SSTFM-5	83	65	68	88	67	51	52	60	67

Table 4. A comparison of the evaluation metrics of the model.

	MIoU (%)	MPA (%)	Accuracy (%)
SSTFM-1	49.26	59.71	76.95
SSTFM-2	70.33	80.71	85.22
SSTFM-3	72.86	82.6	86.62
SSTFM-4	58.63	69.94	79.21
SSTFM-5	64.04	75.12	82.93

3.2.2. Results of the Introduction of Transfer Learning Methods

The pre-trained models for the SSTFM-1 and SSTFM-2 models are deep learning models trained on the VOC2007 dataset. Using the SSTFM-1 as the pre-training model, training was performed on the T2 training set to obtain the SSTFM-3 model. The accuracy of the model is 86.62%. The best precision is 97.56% for farmland and the worst precision is 80.94% for the construction land. The detailed results are shown in Tables 3 and 4. Table 3 shows a comparison of the values of each indicator type in the model, while Table 4 shows a comparison of the evaluation metrics of the model.

3.2.3. Results of the DeepLab v3+ Model

Currently, DeepLabV3+ [12] is considered to be the most effective network in the field of semantic segmentation. It has achieved remarkable success in accurately segmenting objects in complex images and has become the preferred choice for researchers and practitioners in this field. However, due to the complex structure of the DeepLab v3+ network and a large amount of extracted information, this network is time-consuming and has high equipment requirements, making it unsuitable for the application scenario in this study. The semantics of the images in this study are relatively simple and fixed in structure, so the skip connection (feature stitching) structure of the U-network works better to its advantage. Due to the difficulty in acquiring data from the small number of images in the study area, the network is easily overfitted when using the large DeepLabv3+. The advantage of the large networks is that they have a greater ability to represent images, whereas simpler, fewer images do not have as much content to represent. Therefore, the U-Net network, with its light and simple structure, has more room for operation. Therefore, in the second series of experiments, this study is compared with the DeepLab v3+ model. In this study, T2 is taken as the training set and the SSTFM-1 is taken as the pre-training model to obtain SSTFM-4. A comparison of the values of each type of indicator in the model is shown in Table 3. Table 4 shows a comparison of the evaluation metrics for the model. The results of the experiment indicate that the U-Net network performs better than the DeepLab v3+ network in the traffic facility construction scenario.

3.2.4. Results of the PSPNet Model

The PSPNet [13] architecture uses a pyramid pooling module to capture multi-scale contextual information from different regions of an input image, which is then concatenated with the feature maps extracted by a convolutional neural network (CNN). The pyramid pooling module divides the feature maps into multiple regions of different sizes and then performs a global pooling within each region, resulting in a set of fixed-length features that capture different scales of contextual information. The output of the PSPNet is a pixel-wise classification map that assigns a label to each pixel in the input image, indicating the semantic category to which it belongs. The PSPNet network is a model that is inferior to DeepLab v3+ in the field of semantic segmentation. Therefore, the experiment in this study compared the improved U-Net network with the model trained by the PSPNet network. In this study, T2 was used as the training set for training, and the SSTFM-1 was used as the pre-training model to obtain the SSTFM-5 model. The description of the SSTFM-5 model is shown in Table 2. The comparison of the different types of index values in the model is shown in Table 3. The comparison of the evaluation indicators of the models is shown in

Table 4. The experimental results show that the U-Net network outperforms the PSPNet network in traffic facility construction scenarios.

3.2.5. Presentation of Prediction Results

Among the four SSTFM models, the SSTFM-1 model performed the worst and the SSTFM-3 model performed the best. The comparison between the prediction results of the SSTFM-1 and SSTFM-3 models is obvious. The performance of the SSTFM-2 and SSTFM-4 models is between SSTFM-1 and SSTFM-3. When all four prediction results are displayed, the comparison of the model prediction results is not obvious. The prediction results of the SSTFM-1 and SSTFM-3 are shown in Figure 6. Among them, the arbor is light green; shrubs are dark red; grassland is green; farmland is blue; water is brown; the road is grey; bare_ground is pink; construction_land is light blue; and slag_dump is red. The first column from the left in the image represents the original; the middle column in the image represents the prediction results obtained from the SSTFM-1; on the far right is the prediction for the SSTFM-3. In Figure 6a, there is a single type of ground object, so there is no significant difference between the prediction results of the SSTFM-1 and SSTFM-3. In Figure 6b, there are fields, roads and bare land. In the prediction results of the SSTFM-1, farmland was predicted as a mixture of grassland and farmland due to the poor accuracy of farmland. In the SSTFM-3, the prediction of farmland is very accurate. This is also well illustrated in Figure 6d. In Figure 6c, the model predicted all shrubs as grassland, while in the SSTFM-3, some shrubs were predicted.

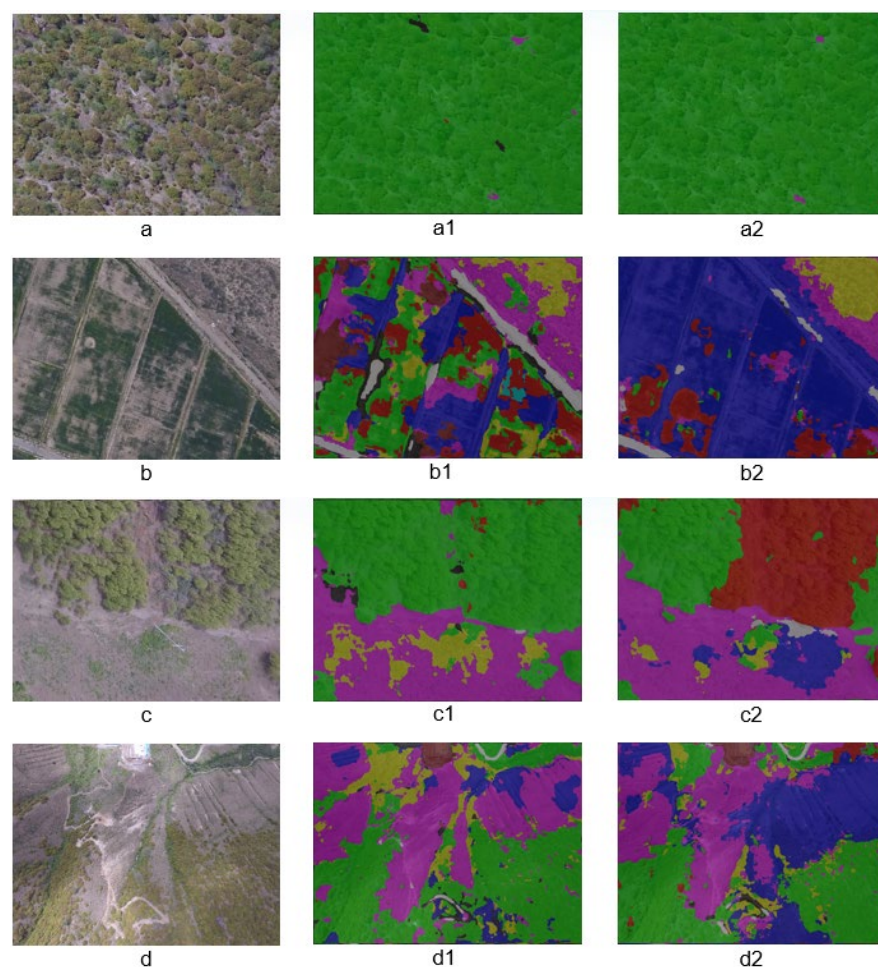


Figure 6. Prediction results of SSTFM-1 and SSTFM-3 ((a)–(d) are the original images. (a1)–(d1) are the prediction images of the SSTFM-1 model corresponding to the original image. (a2)–(d2) are the prediction images of the SSTFM-3 model corresponding to the original image).

4. Conclusions

This paper proposes a multi-objective semantic segmentation algorithm based on an enhanced U-Net network to improve the recognition accuracy of diverse ground objects in the transportation facility construction area. Firstly, a sample dataset of transportation facility construction scenes based on remote sensing images is constructed. To address the problem of a limited number of image samples that contained target objects and an imbalanced distribution of the various training samples, this paper introduces a virtual data augmentation technique. This method aims to augment the number of training samples and balance the distribution of the various training samples. At the same time, the recognition accuracy of the model is improved by using transfer learning. The experimental results demonstrate the effectiveness of the proposed virtual data augmentation approach. The SSTFM enables a better semantic segmentation of multiple target features (arbor, shrub, grassland, farmland, bare land, water, road, construction land and slag dump). The semantic segmentation accuracy of each feature type is more than 80%. The highest semantic segmentation accuracy of the feature type was 97.56%. The prediction results of the model led to the creation of a target feature segmentation map of the transport facility construction area. This map serves as a basis for a risk assessment and monitoring during the construction of transport facilities.

In the future, this study will investigate unsupervised semantic segmentation methods for transport construction scenarios. In practice, training samples to label tags is a time-consuming project, and if accurate segmentation of multiple classes of target features can be achieved unsupervised, then labour costs can be greatly reduced. Finally, this study will develop a software platform that will automate the monitoring of transport construction sites while analysing the different potential risks present at the sites.

Author Contributions: Conceptualization, X.H. and X.L.; methodology, X.H.; validation, X.L., L.Y. and R.Y.; formal analysis, L.Y.; investigation, L.Z.; resources, X.H.; data curation, X.L.; writing—original draft preparation, X.H.; writing—review and editing, X.H. and X.L.; visualization, X.L.; supervision, R.Y.; project administration, X.H.; funding acquisition, X.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the Watershed Non-point Source Pollution Prevention and Control Technology and Application Demonstration Project (2021YFC3201505), Ecological protection and restoration of estuarine wetlands in the Yellow River Delta (2022-YRUC-01-0103). The National Key Research and Development Project (No. 2016YFC0502106), the Natural Science Foundation of China Research Grants (No. 41476161), and the Fundamental Research Funds for the Central Universities.

Data Availability Statement: The storage URL of the structured raw data to construct the knowledge graph is: <https://github.com/hao1661282457/SSTFM-images.git> (accessed on 2 February 2023).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Hao, S.; Zhou, Y.; Guo, Y. A Brief Survey on Semantic Segmentation with Deep Learning. *Neurocomputing* **2020**, *406*, 302–321. [CrossRef]
2. Thoma, M. A survey of semantic segmentation. *arXiv* **2016**, arXiv:1602.06541.
3. Guo, Y.; Liu, Y.; Georgiou, T.; Lew, M. A review of Semantic Segmentation Using Deep Neural Networks. *Int. J. Multimed. Inf. Retr.* **2018**, *7*, 87–93. [CrossRef]
4. Inglada, J. Automatic Recognition of Man-made Objects in High resolution Optical Remote Sensing Images by SVM Classification of Geometric Image Features. *ISPRS J. Photogramm. Remote Sens.* **2007**, *62*, 236–248. [CrossRef]
5. Kang, B.; Nguyen, T. Random Forest with Learned Representations for Semantic Segmentation. *IEEE Trans. Image Process.* **2019**, *28*, 3542–3555. [CrossRef] [PubMed]
6. Long, J.; Shelhamer, E.; Darrell, T. Fully Convolutional Networks for Semantic Segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
7. Badrinarayanan, V.; Kendall, A.; Cipolla, R. Segnet: A Deep Convolutional Encoder-decoder Architecture for Image Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 2481–2495. [CrossRef] [PubMed]

8. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, 5–9 October 2015; pp. 234–241.
9. Chen, L.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A. Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs. *arXiv* **2014**, arXiv:1412.7062.
10. Chen, L.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A. Deeplab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *40*, 834–848. [[CrossRef](#)] [[PubMed](#)]
11. Chen, L.; Papandreou, G.; Schroff, F.; Adam, H. Rethinking Atrous Convolution for Semantic Image Segmentation. *arXiv* **2017**, arXiv:1706.05587.
12. Chen, L.C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-decoder with Atrous Separable Convolution for Semantic Image Segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 801–818.
13. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid Scene Parsing Network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2881–2890.
14. Yin, X.H.; Wang, Y.C.; Li, D.Y. Suvery of Medical Image Segmentation Technology Based on U-Net Structure Improvement. *Ruan Jian Xue Bao/J. Softw.* **2021**, *32*, 519–550.
15. Nalepa, J.; Mrukwa, G.; Piechaczek, S.; Lorenzo, P.R.; Marcinkiewicz, M.; Bobek-Billewicz, B.; Wawrzyniak, P.; Ulrych, P.; Szymanek, J.; Cwiek, M.; et al. Data Augmentation via Image Registration. In Proceedings of the 2019 IEEE International Conference on Image Processing (ICIP), Taipei, Taiwan, 22–25 September 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 4250–4254.
16. Nishio, M.; Noguchi, S.; Fujimoto, K. Automatic Pancreas Segmentation Using Coarse-scaled 2d Model of Deep Learning: Usefulness of Data Augmentation and Deep U-net. *Appl. Sci.* **2020**, *10*, 3360. [[CrossRef](#)]
17. Jin, Y.W.; Jia, S.; Ashraf, A.B.; Hu, P. Integrative Data Augmentation with U-Net Segmentation Masks Improves Detection of Lymph Node Metastases in Breast Cancer Patients. *Cancers* **2020**, *12*, 2934. [[CrossRef](#)] [[PubMed](#)]
18. Uysal, E.S.; Bilici, M.Ş.; Zaza, B.S.; Özgenç, M.Y.; Boyar, O. Exploring the Limits of Data Augmentation for Retinal Vessel Segmentation. *arXiv* **2021**, arXiv:2105.09365.
19. Aboudi, F.; Drissi, C.; Kraiem, T. Efficient U-Net CNN with Data Augmentation for MRI Ischemic Stroke Brain Segmentation. In Proceedings of the 2022 8th International Conference on Control, Decision and Information Technologies (CoDIT), Istanbul, Turkey, 17–20 May 2022; IEEE: Piscataway, NJ, USA, 2022; Volume 1, pp. 724–728.
20. Sfakianakis, C.; Simantiris, G.; Tziritas, G. GUDU: Geometrically-constrained Ultrasound Data augmentation in U-Net for echocardiography semantic segmentation. *Biomed. Signal Process. Control* **2023**, *82*, 104557. [[CrossRef](#)]
21. Lilay, M.Y.; Taye, G.D. Semantic Segmentation Model for Land Cover Classification from Satellite Images in Gambella National Park, Ethiopia. *SN Appl. Sci.* **2023**, *5*, 76. [[CrossRef](#)]
22. Hashim, N.; Hamid, J.R.A. Multi-Level Image Segmentation for Urban Land-Cover Classifications. In *IOP Conference Series: Earth and Environmental Science*; IOP Publishing: Bristol, UK, 2021; Volume 767, p. 012024. [[CrossRef](#)]
23. Alshahrani, A.A.; Jaha, E.S. Locality-Sensitive Hashing of Soft Biometrics for Efficient Face Image Database Search and Retrieval. *Electronics* **2023**, *12*, 1360. [[CrossRef](#)]
24. Simard, P.Y.; LeCun, Y.A.; Denker, J.S.; Victorri, B. Transformation Invariance in Pattern Recognition—Tangent Distance and Tangent Propagation. In *Neural Networks: Tricks of the Trade*; Springer: Berlin/Heidelberg, Germany, 2012; pp. 235–269.
25. Hao, X.; Liu, L.; Yang, R.; Yin, L.; Zhang, L.; Li, X. A Review of Data Augmentation Methods of Remote Sensing Image Target Recognition. *Remote Sens.* **2023**, *15*, 827. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.