Multimodal Semantic Segmentation Model for Remote Sensing Images

Abstract

The conceptual division of remote sensing pictures is essential to remote sensing technology. However, predictions are hard to make because the main groups of these remote-sensing pictures are very complicated. Also, the things shown in shots from space are more involved, and many things in different groups are mixed. Because of this, it is hard to optimize based on the feature area. This study introduces a new non-supervised semantic segmentation method based on Mean Teacher (MT). This method is meant to make models more stable and feature-based class naming better. We also change things at the feature level. When we learn about features, we also use contrastive learning to ensure that things don't change when features change. The ISPRS Potsdam dataset and the challenging iSAID dataset have been used in many tests.

Keywords: remote sensing; semantic segmentation; semi-supervised learning; consistency regularization; feature perturbation; contrastive learning

1. Introduction

The accuracy of pictures taken by remote sensing has been slowly rising over the past few years [1]. It's easier to see the features of things in photos, but this worsens the differences between classes and the similarities between classes in picture data. This makes it harder to tell things apart in the spectral domain, which makes it hard to classify land use [2]. Figuring out what parts of a picture are meaningful based on their shape, colour, and surrounding information is called semantic segmentation. Then, these traits are used to put each pixel in the picture into a category. So, semantic segmentation is used to process high-resolution photos from space that contain more spatial information. There are now a lot of great semantic segmentation algorithms out there, like the FCN, Unet, Segnet, and Deeplab series [3]. These methods have done better than usual machine learning techniques when classifying land use.

Usually, there are two steps in semantic segmentation: the encoder and the decoder. A lot of researchers have worked to make the encoder work better. Hu et al. made the channel focus device called the squeeze-and-excitation module (SEM) [4]. Based on the world average value of features, this system clearly shows how features depend on each other.

After that, the links are used to scale traits in a way that isn't a straight line. To make the encoder work better, it helps the useful features stand out and hides the less useful ones. Woo et al. [5] made the convolutional block attention module (CBAM). With this method, the best and average values of traits worldwide are picked as the places to start making them better. It is used to improve the accuracy of channel focus in semantic segmentation. After thinking that feature statistics could be improved, Qin looked into how channel focus mechanisms choose features based on feature statistics. He got feature statistics by cutting down on data through frequency analysis. This is how he came up with the idea of frequency channel attention (FCA) [6]. This method did 1.8% better than the

SEM on the ImageNet dataset for the Top-1 accuracy test. Fu et al. [7] developed the dual attention mechanism (DAM), a fresh way to make things that focus attention. This kind of network sees each prominent feature as a reaction to a unique part of the original picture. It then only improves the high-level features after taking these into account. Because the highest-level features have small feature sizes, the highest-level features decide the feature weights.

As people think of new ways to pay attention to channels, remote sensing is used to sort land use from space. The SEM was added to remote sensing semantic segmentation by Panboonyuen et al. [8] to make it more accurate. They added more steps to the encoder than they already had [9]. This made it even more accurate. Abdollahi et al. added the SEM to the Unet network, which helped the decoder in some ways [10]. Lan et al. used the Deeplab concept and made the ASPA method [11] even better by adding the SEM to handle high-level features. They did what Yang et al. said and added the CBAM to the Unet network. They then used it in both the encoder and the decoder.

Many use channel attention ways to separate things that mean different things using distant sense. Yet, remote sensing often uses channel focus methods [12]. It doesn't look like pictures from space are getting better. Remote sensing images are different from photos taken in nature. Natural pictures only show information in bands that the human eye can see. However, IR images often show information. On the other hand, the visible light band and the near-infrared band carry their data and are not strongly connected. [13] Channel attention methods that are used most often link feature values like average and maximum to feature weights. When land is used for different types of things, like vegetation, open land, and artificial surfaces, the NIR band can have higher average values because it reflects more light. Since the average number in the NIR band is the largest, it might seem like this band needs more attention. This difference could make it hard for the channel focus system to figure out how to evaluate weights. Furthermore, keeping an eye on the highest numbers of each band could cause too much focus on one class at the expense of others.

Improved feature extraction has finally done what it was meant to do. The piece ends with analyzing how well the channel focus process works and how well the feature weights are set in terms of how easy they are to understand. [14,15] The FEM is an excellent way to figure out which features in remote sensing pictures are the most important. Visualizing feature weights and downscaling features can also ensure that feature weights are correct and improve the channel attention process.

2. Related Works

Semantic and mixed semantic segmentation studies will be critical for this part.

2.1. Grouping based on meaning

The Fully Convolution Network (FCN) [16] made a significant impact in the area of semantic segmentation. Pyramid Scene Parsing Network (PSPnet) [17] is new because it uses a pyramid

Pooling is a way to combine feature maps from different sizes so that they have more accurate models of the features. By mixing skip links and expanded convolutions, Deeplabv3 makes semantic segmentation even more precise. The HRNet model can accurately divide images at high resolutions using parallel convolution routes at different resolutions.

Pyramid Scene Parsing Network (PSPnet) [17] is new because it uses pyramid pooling to combine feature maps from different sizes to have more accurate feature models. By mixing skip links and expanded convolutions, Deeplabv3 makes semantic segmentation even more precise. The HRNet model can accurately divide images at high resolutions using parallel convolution routes at different resolutions.

Swin-Unet [17,18] is a pure Transformer model that was made to separate parts of medical images. Its skip links and encoder-decoder structure make it easy to get environmental traits and put them together. [19] Furthermore, RGB cameras can be influenced by lighting conditions or show picture fuzz when moving quickly, making semantic segmentation less accurate. So, more studies need to be done on combining different data types and using other monitoring technologies to make semantic segmentation more reliable and precise.

2.2. Segmentation of Multimodal Semantic Data

RGB pictures gather information by combining data from various sensor types. The main goal of this technology is to use how different modes of communication work well together to make semantic division more accurate and reliable. [20,21] The new co-attention feature in CANet, on the other hand, changes RGB and depth information to work together.

This method shows that thermal image data can help mean segmentation tasks. Researchers are also looking into improving the semantic segmentation performance in self-driving cars by combining light flow data with RGB data. [22] For RGB-LiDAR, researchers have devised several ways to integrate data from these two types of sensors. Builds more robust cross-modal feature representations by using connections between modes. CMNeXt [23] speeds up semantic segmentation by adding to the model in a way that isn't smooth.

There are two main ways to build multimodal semantic segmentation models: Firstly, by combining data from different modes as the model's sources [24]. Conversely, this method has big problems because it can only be made for one mode. In the second method, features are extracted independently for each modality. This means that different backbones are needed for various types of modality feature extraction jobs. [25] While this method works well for cross-modal semantic segmentation, it's hard to add more modality types because of how hard it is to make complex feature extraction modules for each one.

3. Materials and Methods

3.1. Methods

The Mean Teacher format is an excellent semi-supervised learning method. Its main goal is to make the model more reliable so that it is less affected by small changes in the input data. We changed some things about both the student and teacher networks by adding new modules that change stuff after the encoder. Along with the encoder, we added a feature representation head to help contrastive learning move forward and do our job and data better. Figure 1 clarifies how the network and model flow are set up.

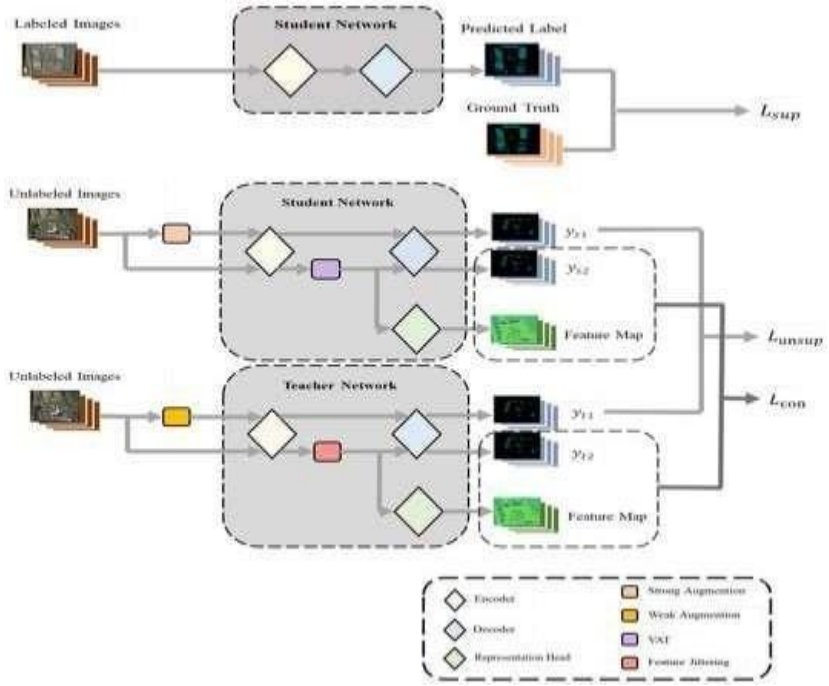


Figure 1. Proposed approach framework overview.

This method has two networks: one for students and one for teachers. Both networks are put together in the same way. Things work better because of three loss functions. The number comprises the predicted labels of the annotated images and the ground truth. The number consists of the predicted labels of two enhanced unannotated images. The number consists of the predicted labels of two images where contrastive learning has changed the features. Labels on a dataset: ={(,)}| | =1 This file does not have any labels: ={( )}| | =1 This way of learning is based on the idea that named pictures are not the only ones in the training set, or | |≦| |. Students will have their network, and teachers will have their network. The student and teacher networks are built similarly but have different parameters. The teacher network's parameters are the exponential moving average (EMA) of the student network's parameters. Here is the formula for updating :

$$\theta_t^e = \gamma\theta_t^{e-1} + (1-\gamma)\,\theta_s^e \tag{1}$$

Feature Sampling with Entropy Threshold Assist for Learning from Differences: In this work, the contrastive learning method is used to help make the best use of the feature area. We use entropy as an extra way to choose questions positive and harmful keys for contrastive.

learning. We get rid of keys that are more right when we set an entropy limit. To learn by comparison and work better, do these things.

3.1.1. Mean Teacher Model with a Disturbed Feature

A tool for changing the features is added at the end of the encoder process. The features that come straight from the encoder and decoder might not work. They need to be broader. An extra picture head is made to draw out and contrast traits, which helps tell them apart. The picture head r from [15] is used for this. When the student network is trained again, the loss changes how it is set up.

We randomly pick pictures every time we train, so there are always the same number of named images (̀) and nameless images ( ), with | |=| |. The student network gets each named picture to guess what it is. Next, we check the directed loss against the actual labels.

$$L_{sup} = \frac{1}{|N_l|} \sum_{(x_i^l, y_i^l) \in N_l} l_{ce}\left(y_{st}^l, y_i^l\right) \tag{2}$$

$$y_{st}^l = O\left(S\left(f \circ h\left(x_i^l; \theta_s\right)\right)\right) \tag{3}$$

燮 ̀(•) (•) stands for the cross-entropy loss and means the marks that were

added by hand. (•)The softmax function is shown by (•), and the one-hot encoding form is shown by (•). You can change the level of the picture with strong image enhancement. On the other hand, the process without image enhancement changes the features in a way that makes more sense and is easier to work with. In this way, we simultaneously add changes at both the picture and feature levels, improving the model in more than one way. After going through different enhancement processes, the pictures are sent to the student network, which creates two forecast labels: 1 1 and 2 2. After they go through the encoder, a VAT adjustment is added to pictures that haven't been enhanced. This is what it means:

$$F_{s2}^u = h\left(x^u\right) + \delta \tag{4}$$

$$\delta = \arg \max_{\|\delta\| \le \epsilon} D_{KL}\left[S\left(f\left(h\left(x^u\right); \theta_s\right)\right) \| S\left(f\left(h\left(x^u\right) + \delta; \theta_s\right)\right)\right] \tag{5}$$

$$y_{s2}^u = O(S\left(f\left(F_{s2}^u\right)\right) \tag{6}$$

The above method shows the student network's change to Virtual Adversarial Training (VAT). ((ℏ( );)) ((ℎ( );)) is the softmax chance of the label that the picture should have given itself if it hadn't been changed in any way using the student

network. ((h( )+ ;)) ((h( )+ ;)) is the biggest possible chance that a label will be made after the image that wasn't improved with VAT disturbance is improved. The Kullback-Leibler divergence, shown by [•], is a way to find the difference between two sets of odds.

The forecast of weak perturbation makes the estimates of solid perturbation work better. In the training network, a feature jittering method is selected, which has less effect on the features. The prediction map 2= ((2)) 2 = ((2)) is made after the feature changes are added and the data is sent through the decoder.

$$L_{unsup} = \frac{1}{|N_u|} \sum_{(x_i^u) \in N_u} l_{ce}\left(y_{s1t}^u, y_{t1t}^u\right) \tag{7}$$

3.1.2. Contrastive Learning with Entropy Threshold Helped Sampling of Features

It was first used for picture-sorting tasks. The goal of contrastive learning is to find the question, the positive and negative keys. The negative key and the question are to be compared to learn how they are alike and different. When semantic segmentation is used with this method, the sample spread goes from a picture to a pixel. It's essential to pick suitable samples for the positive and harmful keys to contrastive learning. Selecting the right harmful keys is critical to get the most out of contrastive learning. It's simple to choose the question and the correct answer. To answer this case problem, we will discuss the idea of entropy.

Entropy is a way to measure how unsure or random a set of data or a chance distribution is. You can use the entropy of the chance distribution of each image to figure out how sure or unsure the forecast is. What does a low Softmax probability entropy number mean? It means that the model is very sure it knows which group the pixel belongs to. We can make it more likely that the negative key we picked is not a fake positive one. This helps to make contrastive learning work better. This is how contrastive learning works, as shown in Figure 2.

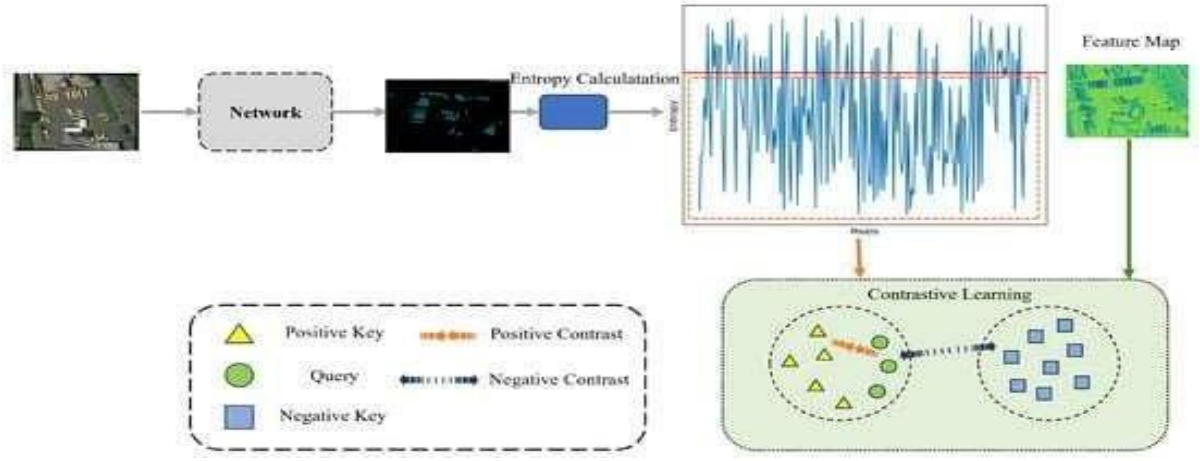


Figure 2. Contrastive learning with entropy.

A level of entropy is chosen, as the red line in the picture shows. For contrastive learning, the pixels and their traits less than the entropy cutoff are selected as the most important ones.

This paper uses contrastive learning loss, which looks like this:

$$L_{con} = -\frac{1}{C \times M} \sum_{c \in C} \sum_{l \in M} log \frac{e^{\left(z_{ci} z_{ci}^{+} / \tau\right)}}{e^{\left(z_{ci} z_{ci}^{+} / \tau\right)} + \sum_{j \in J} e^{\left(z_{ci} z_{cij}^{-} / \tau\right)}} \tag{8}$$

$$E_{ij} = -\sum_{c \subset C} S_{ij}(c) \, log S_{ij}(c) \tag{9}$$

The best guess for the jth point in the ith picture being of class c is. We set a limit and the critical value to help us pick the negative key. Minor changes in the teacher network and significant changes in the student network occur when you change the traits. This is why, most of the time, we use the questions the teaching network comes up with. These are some ways to use the word "query."

$$Q_c^s = \mathbb{1}\left(\hat{y}_{ij} = c\right)\left(r \circ \left(h\left(x_{ij}\right) + \delta\right)\right) \tag{10}$$

$$Q_c^t = \mathbb{1}\left(\hat{y}_{ij} = c\right)\left(r \circ \left(h\left(x_{ij}\right)\right)\right) \tag{11}$$

As long as ❑ Thank you very much. Eighty per cent of our questions come from samples in the teacher network. The other twenty per cent come from samples in the student network. Once the question and positive key have been chosen, the harmful keys ( ) are selected randomly from the rest of the features and must also meet the entropy threshold condition, explained below.:

$$N_c \sim \text{Uniform}\left(z \setminus Q_c, P_c\right) \quad and \quad E_{ij} < \alpha \tag{12}$$

The contrastive loss can be found once the final key values have been chosen. Finally, the model in this work has the following loss update:

3.2. Datasets

3.2.1. USAID

The USAID dataset is used in this study to see how well our suggested method for semantic segmentation works. In the said collection, there are 2806 high-resolution flyover shots. To simplify the tests, the dataset is split into a training set with 1411 pictures and a test set with 458 images. A way we use to make the data better is to randomly cut the images to 512x512 pixels while they are being trained.

3.2.2. Potsdam

The Potsdam collection can also be used. The books in the Potsdam library are from around Germany around Potsdam. In semantic segmentation for remote sensing, we always use this dataset. The International Society for Photogrammetry and Remote Sensing (ISPRS) and the German Society for Photogrammetry, Remote Sensing, and Geoinformation (DGPF)

They worked together to make it. High-resolution pictures from above with an UltraCamXp large-format digital overhead camera make up the set. Each picture is 5 cm away from the next. The images in this set are all bigger than the pictures in the Vaihingen set. They each show more about the land and cover an area of about 1000x1000 metres.

3.3. Evaluation Metrics

The mean intersection over union ( ) is used to measure this. It is a famous variable used for semantic segmentation tasks. It measures how much-projected segments and their ground facts match. Giving you a complete picture of how accurate the model is. People who work with semantic segmentation often talk about the mean ( ) found for all the classes in the dataset.

$$IoU = \frac{TP}{TP + FP + FN}$$

3.4. Implementation Detail

The project used Deeplab V3+, and ResNet-101 was the core network. Pics are randomly cut to 512x512 sizes before they are sent to the training network. You can train for 200 times in a single batch. The speed of learning has been slowed down by 0.9 of a second. The stochastic gradient descent (SGD) engine was used. The learning rate starts at 0.01 but drops over time to 0.0005. The collection has groups of 1/2, 1/4, 1/8, and 1/16 named pictures. The model is trained with the rest of the images that don't have names. For the extra drop's weight, it is set to 0.4. When the number of 1 is changed, the EMA smoothing factor is set to 0.99.

4. Experiments

4.1. Data

The slide window cropping method cuts the remote sensing picture into 300 × 300 deep learning samples. Houses, roads, woods, and lakes are some features that can be found in remotely sensed images.

There are two sets of data: the training set and the validation set. The ratio of the training set to the validation set is 4:1. QGIS is used for the picture labels. Different grey colours are used

to show various parts of the picture. You can tell cells of the same type apart because they all have the same grey value.

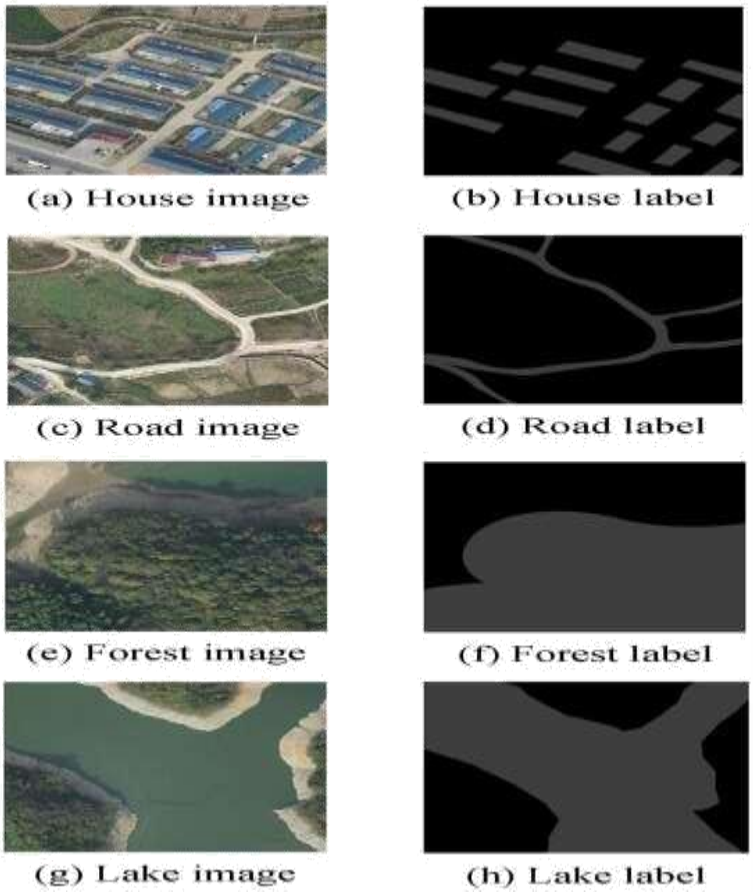


Figure 3. Images and labels from remote sensing.

Randomly moving the training data across, down, and diagonally, along with applying the right amount of linear stretching. To make the validation dataset look like the variable domain, it is randomly stretched by 0.8%, 1%, 1.5%, and 2%.

4.2. Environment and Parameter Configuration

Table 1 shows how the system is set up for AS-Unet+

+. Table 1. Configuring the environment.

Name	Version
GDAL	3.3.3
segmentation-models-PyTorch	0.3.2
CUDA	11.1.0

Table 2 shows the training values that were used in the test.

Table 2. Set parameters.

Parameter	Value
Batch size	16
Initial learning rate	0.0001
Learning Momentum	0.9

4.3. Evaluation Indicator

The measures used for review are IoU, MIoU, Precision, and Recall. A confused matrix, like in Table 3, determines the rating signs.

Table 3. Classification outcome confusion matrix.

Reality	Predicted Results	
	Positive	Negative
Positive		
Negative		

Precision is the percentage of accurately predicted images in a particular group. Here's how to figure it out:

$$\text{Precision} = \frac{TP}{TP + FP}$$

This is how the method for figuring out recall works: recall is the percentage of all the pixels in a particular group that the network correctly identified.

$$\text{Recall} = \frac{TP}{TP + FN}$$

If the expected result and the actual number of each class are compared, the result is called IoU. The MIoU can be found by adding the average number to the crossing ratio of each class. The projected number is more likely to be correct if the MIoU is bigger. Here's how to figure out the IoU and MIoU:

$$\text{IoU} = \frac{TP}{TP+FP+FN},$$
$$\text{MIoU} = \frac{\sum \text{IoU}}{n},$$

where n is the number of groups that can separate images.

4.4. Experimental Results

The attempt with ablation

This is how we checked how well the two parts of the ASPP and SE model worked: Different networks made the predicted division plans for lakes, trees, roads, and houses, as seen in Figure 4.

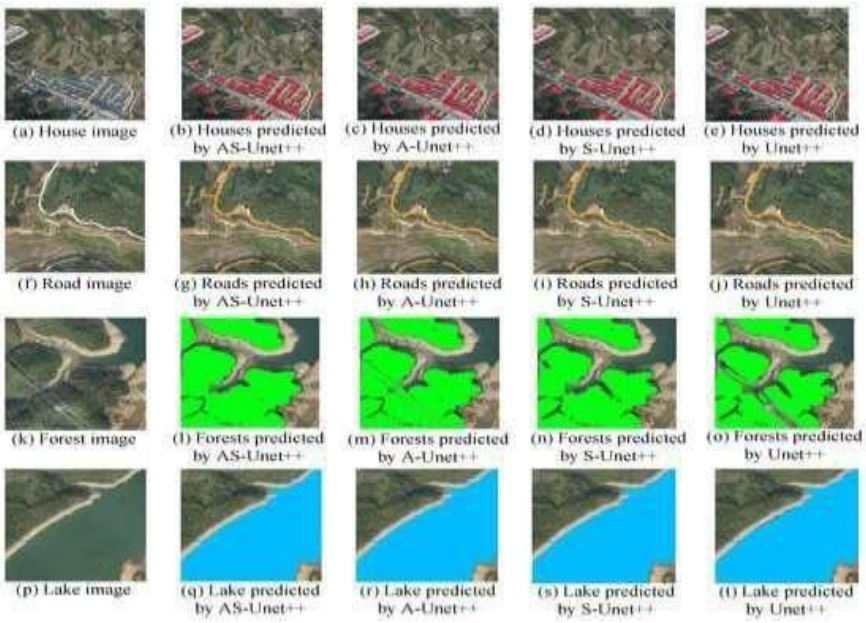


Figure 4. Ablation experiment segmentation predictions.

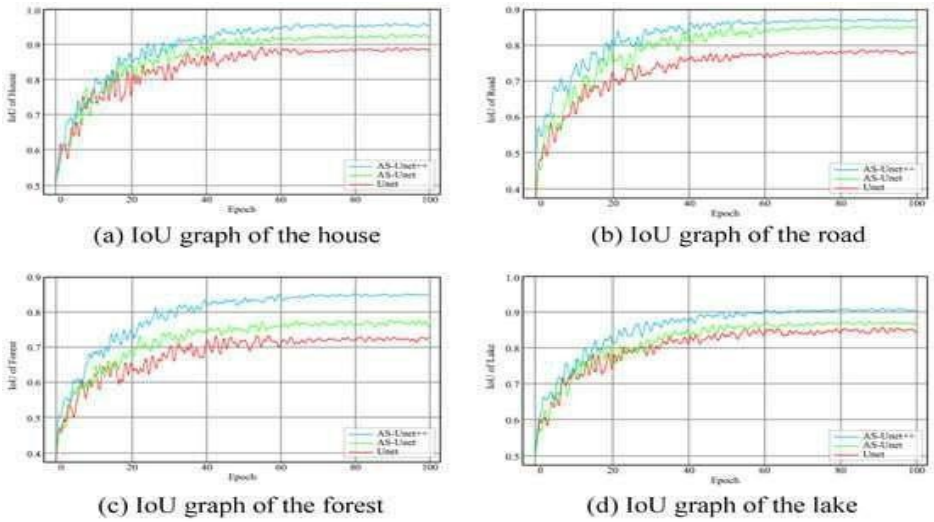
Some houses are missed by Unet++ when it comes to recognizing their parts because of differences in light and colour, and the way the edges of houses are segmented could be better for recognition. The result of the separation of house edges has yet to get better. Both missed detection and the edge segmentation effect improved in the AS-Unet++ network after adding both modules. Even though it had one more module than A-Unet++, it wasn't better at recognizing houses than that version. It was better at recognizing omissions in S-Unet++ with the SE model but did not need to criticize the edges of houses. Table 4 shows the Precision, Recall, and IoU of different networks used in the test sets to predict houses, roads, forests, and lakes.

Table 4. Ablation experiment network comparisons.

Elements	Valuation Indexs	AS-Unet++	A-Unet++	S-Unet++	Unet++
Forest	Recall	0.859	0.813	0.802	0.766
	IoU	0.854	0.802	0.787	0.759
Lake	Precision	0.907	0.882	0.857	0.852
	Recall	0.917	0.905	0.878	0.863

Elements	Valuation Indexs	AS-Unet++	A-Unet++	S-Unet++	Unet++
	IoU	0.912	0.894	0.869	0.858

Figure 5 shows the MIoU graphs in the three types of networks while they were trained with homes, roads, woods, and lakes.



(a) IoU graph of the house (b) IoU graph of the road

(c) IoU graph of the forest (d) IoU graph of the lake

Figure 5. The IoU graphs are used for training in AS-Unet++, Unet, and AS-Unet.

The AS-Unet++ proof set's MIoU reached 88.9% after it was trained. On the other hand, Unet and AS-Unet had an MIoU of 80.8% and 85.8% on the test set.

Table 5. Network comparisons on various verification sets.

Elements	Valuation Index	AS-Unet++	AS-Unet	Unit
	Recall	0.879	0.858	0.783
	IoU	0.874	0.852	0.781
Forest	Precision	0.847	0.779	0.702
	Recall	0.856	0.788	0.718
	IoU	0.851	0.784	0.711
Lake	Precision	0.902	0.857	0.842
	Recall	0.911	0.865	0.853
	IoU	0.905	0.862	0.848

AS-Unet and Unet are two networks. The three differences stayed relatively the same, as shown in Figure 6. During the training to recognize road parts, the AS-Unet++ network changed the least, just a little faster than the other two.

Table 6. Network performance on various test sets.

Elements	Valuation Indexes	AS-Unet++	AS-Unet	Unit
House	Recall	0.978	0.943	0.899
	IoU	0.971	0.937	0.896
	Precision	0.907	0.856	0.841
Lake	Recall	0.917	0.863	0.852
	IoU	0.912	0.859	0.846

Ninety-two per cent of the test set had MIoUs for AS-Unet++, eighty-five per cent for Unet, and eighty-five per cent for AS-Unet. Based on what you just read, AS-Unet++ does better on all test sets than Unet and AS-Unet.

5. Conclusions

In addition, AS-Unet++ can successfully lower the number of times devices are misidentified or missed. Even though the method in this work makes the segmentation more accurate, the generalization condition is still challenging to meet when dealing with complex and changing remote sensing pictures, like those that show elements in different lighting conditions or with complicated shapes. Bettering the model's ability to generalize and getting even better at classification should be the main goals of future work.

References

[1] Kussul, N.; Lavreniuk, M.; Skakun, S.; Shelestov, A. Deep learning classification of land cover and crop types using remote sensing data. IEEE Geosci. Remote Sens. Lett. 2017, 14, 778–782. [Google Scholar] [CrossRef]

[2] Huang, X.; Zhang, L. An SVM ensemble approach combining spectral, structural, and semantic features to classify high-resolution remotely sensed imagery. IEEE Trans. Geosci. Remote Sens. 2012, 51, 257–272. [Google Scholar] [CrossRef]

[3] Khatami, R.; Mountrakis, G.; Stehman, S.V. A meta-analysis of remote sensing research on supervised pixel-based land-cover image classification processes: General guidelines for practitioners and future research. Remote Sens. Environ. 2016, 177, 89–100. [Google Scholar] [CrossRef]

[4] 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. [Google Scholar]

[5] Badrinarayanan, V.; Handa, A.; Cipolla, R. Segnet: A deep convolutional encoder-decoder architecture for robust semantic pixel-wise labelling. arXiv 2015, arXiv:1505.07293. [Google Scholar]

[6] Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, 5–9 October 2015; pp. 234–241. [Google Scholar]

[7] Chen, L.-C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE Trans. Pattern Anal. Mach. Intell. 2017, 40, 834–848. [Google Scholar] [CrossRef] [PubMed]

[8] Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141. [Google Scholar]

[9] Woo, S.; Park, J.; Lee, J.-Y.; Kweon, I.S. Cbam: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19. [Google Scholar]

[10] Qin, Z.; Zhang, P.; Wu, F.; Li, X. Fcanet: Frequency channel attention networks. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 10–17 October 2021; pp. 783–792. [Google Scholar]

[11] Fu, J.; Liu, J.; Tian, H.; Li, Y.; Bao, Y.; Fang, Z.; Lu, H. Dual attention network for scene segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 3146–3154. [Google Scholar]

[12] Panboonyuen, T.; Jitkajornwanich, K.; Lawawirojwong, S.; Srestasathiern, P.; Vateekul, P. Semantic segmentation on remotely sensed images using an enhanced global convolutional network with channel attention and domain-specific transfer learning. Remote Sens. 2019, 11, 83. [Google Scholar] [CrossRef]

[13] Abdollahi, A.; Pradhan, B.; Shukla, N.; Chakraborty, S.; Alamri, A. Multi-object segmentation in complex urban scenes from high-resolution remote sensing data. Remote Sens. 2021, 13, 3710. [Google Scholar] [CrossRef]

[14] Lan, Z.; Huang, Q.; Chen, F.; Meng, Y. Aerial image semantic segmentation using spatial and channel attention. In Proceedings of the 2019 IEEE 4th International Conference on Image, Vision and Computing (ICIVC), Xiamen, China, 5–7 July 2019; pp. 316–320. [Google Scholar]

[15] Yang, J.; Zhao, L.; Dang, J.; Wang, Y.; Yue, B.; Gu, Z. A Semantic Segmentation Method for High-resolution Remote Sensing Images Based on Encoder-Decoder. In Proceedings of the 2022 Tenth International Conference on Advanced Cloud and Big Data (CBD), Guilin, China, 4–5 November 2022; pp. 98–103. [Google Scholar]

[16] Zhang, J.; Liu, R.; Shi, H.; Yang, K.; Reiß, S.; Peng, K.; Fu, H.; Wang, K.; Stiefelhagen, R. Delivering Arbitrary-Modal Semantic Segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 1136–1147. [Google Scholar]

[17] 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. [Google Scholar]

[18] Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5–9, 2015, Proceedings, Part III 18; Springer: Cham, Switzerland, 2015; pp. 234–241. [Google Scholar]

[19] 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. [Google Scholar]

[20] A.Rajasekar, B.R. Tapas Bapu, N.Ashokkumar, Optimization techniques and hybrid deep learning approaches for UV index predictions, Volume 1 Issue 1, April 2023.

[21] Chen, L.C.; Papandreou, G.; Schroff, F.; Adam, H. Rethinking atrous convolution for semantic image segmentation. arXiv 2017, arXiv:1706.05587. [Google Scholar]

[22] Sun, K.; Xiao, B.; Liu, D.; Wang, J. Deep, high-resolution representation learning for human pose estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 5693–5703. [Google Scholar]

[23] Borse, S.; Wang, Y.; Zhang, Y.; Porikli, F. Inverseform: A loss function for structured boundary-aware segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 5901–5911. [Google Scholar]

[24] Ding, H.; Jiang, X.; Liu, A.Q.; Thalmann, N.M.; Wang, G. Boundary-aware feature propagation for scene segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 6819–6829. [Google Scholar]

[25] Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16 × 16 words: Transformers for image recognition at scale. arXiv 2020, arXiv:2010.11929. [Google Scholar]

[26] Xie, E.; Wang, W.; Yu, Z.; Anandkumar, A.; Alvarez, J.M.; Luo, P. SegFormer: Simple and efficient design for semantic segmentation with transformers. Adv. Neural Inf. Process. Syst. 2021, 34, 12077–12090. [Google Scholar]