Pertanika Proceedings 1 (5): 159 - 167 (2025)

PERTANIKA JOURNALS

PERTANIKA PROCEEDINGS

Journal homepage: http://www.pertanika.upm.edu.my/

Multimodal Semantic Segmentation Model for Remote Sensing Image

Ashokkumar N[1]*, Shaik Javid Basha[2], Nagarajan. P[3], Kavitha. T[4], Shaik Mohammad Eliyas[2] and K. Abdul Rahman[2]

[1]Department of Electronics and Communication Engineering, Mohan Babu University, Tirupathi, Andhra Pradesh 517102, India
[2]Department of Electronics and Communication Engineering, Santhiram Engineering College, Nandyal, Andhra Pradesh 518501, India
[3]Department of Electronics and Communication Engineering, SRM Institute of Science and Technology, Vadapalani Campus Chennai, Tamil Nadu 603203, India
[4]Department of Electronics and Communication Engineering, Vel Tech Rangarajan Dr Sagunthala R&D Institute of Science and Technology, Chennai, Tamil Nadu 600062, India

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 do not change when features change. The ISPRS Potsdam dataset and the challenging iSAID dataset have been used in many tests.

Keywords: Contrastive learning, consistency regularization, feature perturbation, remote sensing, semantic segmentation, semi-supervised learning

ARTICLE INFO

Article history:
Received: 29 August 2025
Published: 31 October 2025

DOI: https://doi.org/10.47836/pp.1.5.028

E-mail addresses:
ashoknoc@gmail.com (Ashokkumar N)
javidbasha1104@gmail.com (Shaik Javid Basha)
nagarajan.pandiyan@gmail.com (Nagarajan. P)
drkavitha@veltech.edu.in (Kavitha. T)
eliyasshaik1998@gmail.com (Shaik Mohammad Eliyas)
rahman394@gmail.com (K. Abdul Rahman)
* Corresponding author

INTRODUCTION

The accuracy of pictures taken by remote sensing has been slowly rising over the past few years (Kussul et al., 2017). It is 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. Figuring out the parts of a picture that 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. There are now a lot of great semantic segmentation algorithms out there, like the FCN, Unet, Segnet, and Deeplab series (Huang & Zhang, 2012). A lot of researchers have worked to make the encoder work better by making the channel focus device called the squeeze-and-excitation module (SEM) (Khatami et al., 2016).

Based on the world's 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 is not a straight line. To make the encoder work better, it helps the useful features stand out and hides the less useful ones. Long et al. (2015) 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. Badrinarayanan et al. (2015) 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).

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 (Ronneberger et al., 2015). Channel attention methods that are used most often link feature values like average and maximum to feature weights. Since the average number in the NIR band is the largest, it might seem like this band needs more attention.

RELATED WORKS

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

Grouping Based on Meaning

The Fully Convolution Network (FCN) made a significant impact in the field of semantic segmentation. Pyramid Scene Parsing Network (PSPnet) is new because it uses a pyramid pooling, a way to combine feature maps from different sizes to create more accurate models of the features (Qin et al., 2023). Swin-Unet is a pure Transformer-based model that was made to separate parts of medical images (Ronneberger et al., 2015). Its skip links and encoder-decoder structure make it easy to get environmental traits and integrating them.

Segmentation of Multimodal Semantic Data

RGB pictures gather information by combining data from various sensor types. The main goal of this technology is to combine different modes of communication together to make semantic division more accurate and reliable. This method shows that thermal image data

can help with mean segmentation tasks. There are two main ways to build multimodal semantic segmentation models: Firstly , by combining data from different modes as the model's sources (Zhang et al. 2023). 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.

MATERIALS AND METHODS

Methods

The Mean Teacher format is an excellent semi-supervised learning method. Its main goal is to make the model more reliable and less sensitive to the small changes in the input data. Along with the encoder, we added a feature representation head to help contrastive learning for better task performance. Figure 1 shows a semi-supervised learning framework with student-teacher network architecture for image classification using labeled and unlabeled data with VAT-based feature augmentation.

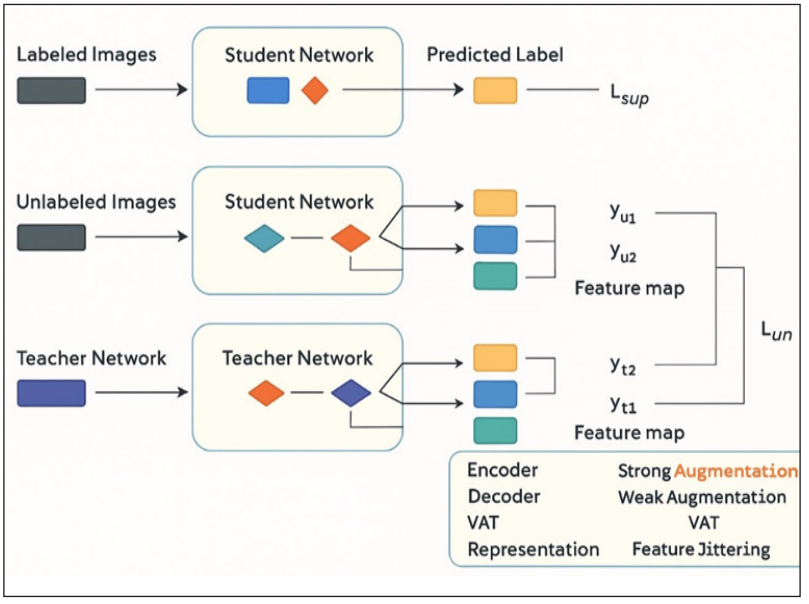

Figure 1. Proposed approach framework overview (Yang et al., 2022)

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. 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:

$$\theta_1^e = \gamma_1^{\theta^{e-1}} + (1-\gamma)\theta_s^e$$

Feature Sampling with Entropy Threshold Assist for Learning from Differences: In this work, the contrastive learning method is used for a better utilization of the feature area. We use entropy as an extra way to choose questions with positive and negative keys for contrastive learning. The keys that are more right are eliminated by applying an entropy limit.

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 extract and contrast traits, which helps tell them apart. The student network received labelled picture to guess what it is. Next, we check the directed loss against the actual labels.

$$L_{sup} = \frac{1}{|N_t|}\sum_{(x_r^1 y_i^1)\in N_l} I_\alpha(y_{st}^1, y_l^1)$$

$$y_{st}^1 = o(s(f^{\circ}h(x_t^l; \theta_s))$$

The method shows the student network's adaptation to Virtual Adversarial Training (VAT). Here isthe intended formulation: Original prediction: p(y|x; θ) is the softmax probability distribution of the label that the image should have, given by the student network without any perturbations. A weak augmentation method is applied that has minimal effect on the feature representations. Perturbed prediction: After feature perturbations are added and data passes through the decoder, the prediction becomes p(y|x + r; θ), where r represents the adversarial perturbation. VAT objective: The method aims to minimize the KL divergence between original and perturbed predictions: KL(p(y|x; θ) || p(y|x + r; θ))

Contrastive Learning with an Entropy Threshold Helped the Sampling of Features

It was first used for picture-sorting tasks. The goal of contrastive learning is to identify the question, the positive and negative keys. The negative key is compared to the to learn how they are alike and different. The entropy of the probability distribution for each image is used to figure out how sure or unsure the forecast. 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 cut off are selected as the most important ones (Figure 2).

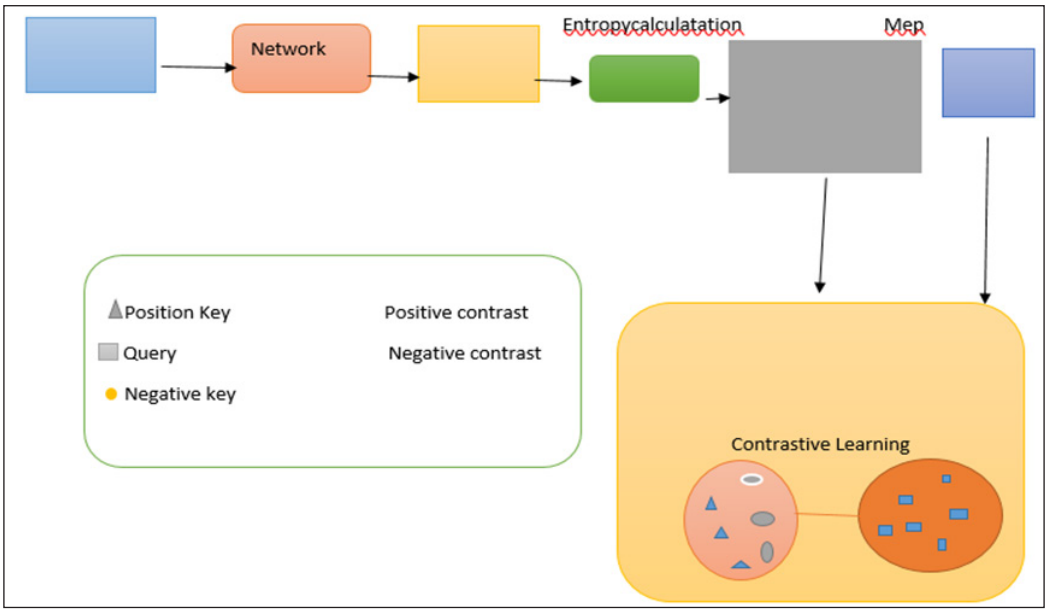

Figure 2. Contrastive learning with entropy (Khatami at al 2016)

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

$$E_{tf} = -\sum_{cCC} S_{tf}\,(c) log S_{tf}\,(c)$$

The best guess for the *j*-th point in the *i*-th picture being of class *c* is. We set a limit and the critical value to help us pick the negative key. Minor changes occurred in the teacher network, and significant changes in the student network occurred when the traits are changed. For this reason, we use the questions the teaching network comes up with. These are some ways to use the word "query." 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:

$$N_c \sim Uniform(z \backslash Q_c, P_c) and\ E_{tf} < \alpha$$

Datasets

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. We make the data better by randomly cut the images to 512×512 pixels while they are being trained.

Evaluation Metrics

The Mean Intersection over Union (mIoU) is employed as the primary evaluation metric. IoU is widely used in semantic segmentation tasks as it quantifies the overlap between the predicted segmentation and the ground truth. It is calculated as the ratio of the intersection to the union of the predicted and actual regions. The mean IoU (mIoU) represents the average IoU across all classes in the dataset and provides a comprehensive measure of segmentation accuracy.

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

Implementation Detail

The project used Deeplab V3+, and ResNet-101 was the core network. Pics are randomly cut to 512×512 sizes before they are sent to the training network. 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.

EXPERIMENTS

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 picture labels. Different grey colors 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.

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%.

Environment and Parameter Configuration

Table 1 shows the training values that were used in the test.

AS-Unet and Unet are two networks. The three differences stayed relatively the same, as shown in Figure 3. During the

Table 1
Training parameters used in the experiments

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

training to recognize road parts, the AS-Unet++ network changed the least, just a little faster than the other two.

The performance of three distinct models (Model A, Model B, and Model C) across 100 training epochs is shown in each of the four graphs in Figure 4. The Intersection over Union (IoU) metric, a popular method for assessing the precision of semantic segmentation models, is used to gauge the performance. A home, a road, a forest, or a lake are the classes that each graph focuses on.

A similar pattern can be seen in all four graphs: the IoU score for each of the three models considerably improves as the number of epochs rises, suggesting that the models are becoming more adept at properly classifying the various classes. In contrast to Models A and B, Model C often begins with a lower IoU score, but it exhibits a quicker learning curve in the early epochs. The starting IoU values of Models A and B are typically greater, and their performance curves increase gradually. All three models converge to a high IoU score at the conclusion of the 100 epochs, indicating that they have achieved a similar degree of segmentation accuracy for each class. In particular, all three models achieve IoU ratings of about 0.70 for the dwelling and woodland classes. The final IoU for the road class is also around 0.70, although the scores for the lake class are a little lower, at approximately 0.75, but still converge. This suggests that despite variations in their learning behavior during the first training phase, all three models are efficient for this segmentation task and that their ultimate performance is quite comparable.

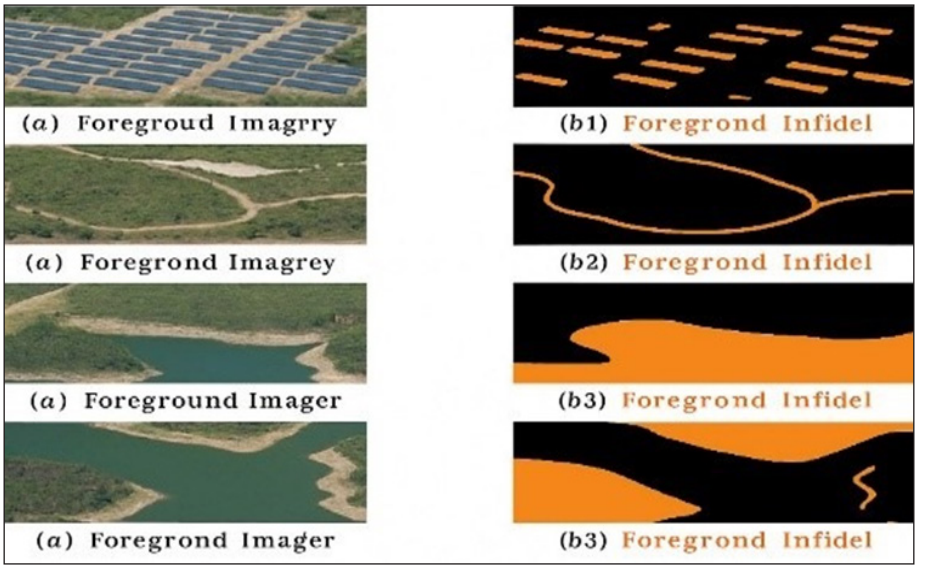

Figure 3. Images and labels from remote sensing (Qin et al., 2021)

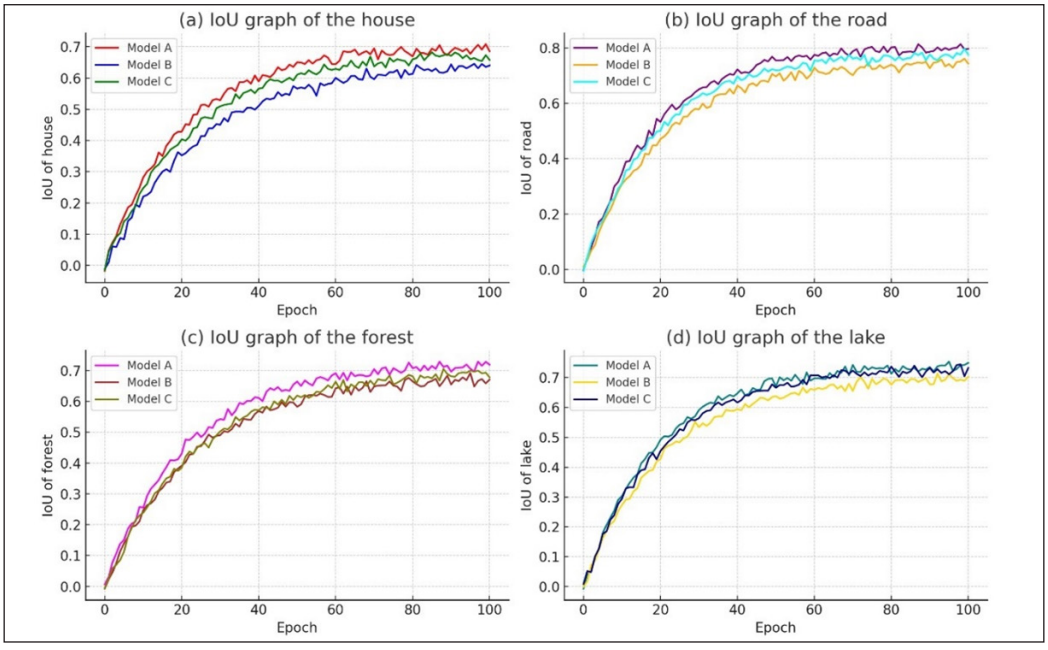

Figure 4. The IoU graphs are used for training in AS-Unet++, Unet, and AS-Unet

Ninety-two per cent of the tests had MIoUs for AS-Unet++, eighty-five per cent for Unet, and eighty-five per cent for AS-Unet. Based on these results, AS-Unet++ does better on all test sets than Unet and AS-Unet.

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 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.

ACKNOWLEDGEMENT

The authors would like to express sincere gratitude to everyone who supported us during the research process

REFERENCES

Badrinarayanan, V., Handa, A., & Cipolla, R. (2015). Segnet: A deep convolutional encoder-decoder architecture for robust semantic pixel-wise labelling. *arXiv preprint arXiv:1505.07293.* https://doi.org/10.48550/arXiv.1505.07293

Huang, X., & Zhang, L. (2012). An SVM ensemble approach combining spectral, structural, and semantic features for the classification of high-resolution remotely sensed imagery. *IEEE Transactions on Geoscience and Remote Sensing*, *51*(1), 257-272. https://doi.org/10.1109/TGRS.2012.2202912

Khatami, R., Mountrakis, G., & Stehman, S. V. (2016). A meta-analysis of remote sensing research on supervised pixel-based land-cover image classification processes: General guidelines for practitioners and future research. *Remote Sensing of Environment*, *177*, 89-100. https://doi.org/10.1016/j.rse.2016.02.028

Kussul, N., Lavreniuk, M., Skakun, S., & Shelestov, A. (2017). Deep learning classification of land cover and crop types using remote sensing data. *IEEE Geoscience and Remote Sensing Letters*, *14*(5), 778-782. https://doi.org/10.1109/LGRS.2017.2681128

Long, J., Shelhamer, E., & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 3431-3440). IEEE.

Qin, Z., Zhang, P., Wu, F., & Li, X. (2021). Fcanet: Frequency channel attention networks. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 783-792). IEEE.

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-assisted Intervention* (pp. 234-241). Springer. https://doi.org/10.1007/978-3-319-24574-4_28

Zhang, J., Liu, R., Shi, H., Yang, K., Reiß, S., Peng, K., Fu, H., Wang, K., & Stiefelhagen, R. (2023). Delivering arbitrary-modal semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 1136-1147). IEEE.