American Academic Scientific Research Journal for Engineering, Technology, and Sciences
ISSN (Print) 2313-4410, ISSN (Online) 2313-4402
https://asrjetsjournal.org/index.php/American_Scientific_Journal/index

Automatic Segmentation of Land Cover in Satellite Images

Ahmet Alp Kındıroğlu[a*], Metehan Yalçın[b], Mahiye Uluyağmur Öztürk[c], Ufuk Uyan[d], Furkan Burak Bağcı[e]

[a,b,c,d,e]*Huawei Turkey R&D Center, Istanbul, Turkey*

Abstract

Semantic segmentation problems such as landcover segmentation rely on large amounts of annotated images to excel. Without such data for target regions, transfer learning methods are widely used to incorporate knowledge from other areas and domains to improve performance. In this study, we analyze the performance of landcover segmentation models trained on low-resolution images with insufficient data for the targeted region or zoom level. In order to boost performance on target data, we experiment with models trained with unsupervised, semi-supervised, and supervised transfer learning approaches, including satellite images from public datasets and other unlabeled sources. According to experimental results, transfer learning improves segmentation performance by 3.4% MIoU (mean intersection over union) in rural regions and 12.9% MIoU in urban regions. We observed that transfer learning is more effective when two datasets share a comparable zoom level and are labeled with identical rules; otherwise, semi-supervised learning is more effective using unlabeled data. Pseudo labeling based unsupervised domain adaptation method improved building detection performance in urban cities. In addition, experiments showed that HRNet outperformed building segmentation approaches in multi-class segmentation.

Keywords: Satellite images; Semantic segmentation; Transfer learning; Semi-supervised learning.

1. Introduction

Land cover analysis maps provide spatial data on many types of physical surface cover, such as a village, forests, croplands, lakes, and roads. Thanks to the information obtained from land cover maps, many necessary applications, such as city and regional planning, agricultural planning, communication, and transportation infrastructure planning, can be easily made. Methods such as recording with vehicles with three-dimensional scanners and sensors, combining images from piloted and unpiloted aerial vehicles, and using satellite images are actively used for the detailed extraction of these maps.

Received: 6/26/2023
Accepted: 7/30/2023
Published: 8/11/2023

---

* Corresponding author.

The most detailed and high-quality annotations for land cover segmentation are collected through ground-level surveys and scanning. However, obtaining data with this method is challenging and expensive, especially in many areas where vehicle access is not possible, such as rural areas, coasts, islands, and mines. For this reason, efforts to obtain earth-use maps from satellite images cheaply in recent years have gained speed.

Machine learning methods use the category of each pixel of images as label information to solve segmentation problems in satellite imagery. Thus, a pixel-based classification model can be trained with the training set obtained, and prediction can be performed on the test data. However, there must be enough data with quality labels in the training set for this process to be carried out successfully. In addition, it is crucial for training data and annotations to have content similar to the target problem data to achieve good performance [2].

Since the data sets collected for remote sensing problems depend on the variety of satellite and autonomous aerial sensors, the standardization of annotations, and so on, it is quite challenging to find enough data to be used for the desired problem. Utilizing different public datasets will reduce dataset dependency. Training the model over different examples with datasets from multiple sources effectively achieves a higher capacity and a more comprehensive general model.

In this study, we aimed to produce low-cost land cover maps to be used in studies to improve the telecommunication infrastructure. The main difficulty in obtaining these maps is the difficulty of labeling and recognizing the resulting labels and images due to the low resolution of satellite images of these regions. For this purpose, we propose methods such as semi-supervised learning, transfer learning, and combined learning to improve prediction success in low-resolution target data using city images and detailed images from other databases. In this method, we developed a model that extracts a land cover map using 20-meter resolution satellite images. We performed detailed hyperparameter analysis on this model with different architectures. Then we trained the model for a 20-meter resolution dataset using one and two-meter resolution datasets with semi-supervised learning and transfer learning approaches. We presented an analysis of these methods in our work.

2. Related Works

Segmentation is an image processing problem in which objects in the input image are estimated at the pixel level. Semantic segmentation aims to assign the same label to all objects in the same class. In particular, studies in this field have gained momentum with the emergence of deep neural networks and annotated large datasets. Fully Convolutional neural networks [3] (FCN) was the first method to apply deep neural networks to the semantic segmentation problem effectively. In the FCN method, the outputs obtained from the neural network layers of different sizes were collected, with the result obtained from the lowest layer, and quality image outputs of the object edges were obtained.

U-Net [4], which has a similar structure, is one of the most well-known models in image segmentation. The U-Net consists of two parts, the encoder and the decoder. While the encoder part extracts spatial-invariant features in the traditional CNN structure, The decoder part samples these features to the output image the same size as

the input image. The difference between U-Net from FCN is that it uses a multi-channel structure in the decoder section and delivers the semantic information to high-resolution layers. Unet++ [5] method uses skip connections from shallow layers to deep layers that combine low and high-level features for better predictions.

Deeplab [6], with its atrous convolutional layer, extended the receptive field of the models, allowing them to train better quality segmentation models with fewer parameters. In the following years, many improvements were made to this model; DeeplabV2 [7], which can segment objects at different scales more consistently with ASPP (Atrous Spatial Pyramid Pooling), DeepLabV3 [8] with improvements on AC (Atrous Convolution) units, DeeplabV3+ [9] which developed an efficient decoder module to improve segmentation results along object boundaries on to DeeplabV3 model were developed.

Identifying and assigning meaning to pixels in satellite images by semantic segmentation is an actively studied research area [10]. The complex nature of the backgrounds in satellite images and the high in-class variability of objects (such as buildings, cities, and roads) are among the main challenges of the segmentation problem. However, there has been much progress in this regard with obtaining high-resolution satellite images and developing deep learning-based semantic segmentation methods.

Many U-Net-based models [11, 12] have been proposed for the urban image segmentation problem. The DeepResUnet [13] proposed by Yi and colleagues for building segmentation in urban areas has fewer parameters and performs better than the original U-Net, albeit with a longer inference time. Combining DeepLabV3+ [9] with object-based image analysis, Du and colleagues [14] achieved successful results in ultra-high-resolution satellite images. In that study, classification results obtained using a DeepLabv3+ network on the spectral image and a random forest classifier on the features obtained by image analysis were converted to final estimates with a conditional random field. In many recent studies, transformer-based models [15, 17] have been successfully applied to the semantic segmentation problem.

Recently Wang and colleagues [1] created HRNet, outperforming semantic segmentation architecture. They maintained classical encoder-decoder architecture to increase high-resolution representations. The classical encoder-decoder approach first encodes the input image as a low-resolution representation, then the decoder part process this low-resolution representation into a high-resolution representation. HRNet contributes parallel information exchange between these two parts, not in series as in the classical approach.

Geonrw [18], Inria [19], and Deepglobe [20] are examples of remote sensing datasets created to make land cover maps. Geonrw is a data set of RGB and SAR remote sensing images mostly taken from urban areas by synthesizing with a generative adversarial network. In addition to having ten image segmentation classes, Geonrw is suitable for altitude estimation or semantic image synthesis applications from aerial photographs.

Transfer learning refers to methods developed to use labeled data or models for one or more classes on a different but related task. The studies carried out in this field are summarized in the literature review by Wang and colleagues[21]. Supervised learning transfer is used if the feature space and the intended task are the same but differ between the feature distributions [22]. Suppose the target and source features differ or the data

annotations suit different tasks. The semi-supervised [23] or unsupervised transfer learning [24] methods are used on the target task.

Studies on supervised learning transfer focus on field adaptation methods such as feature space alignment, adding loss functions that provide space adaptation to the model error function, or adding capacity to the model architecture to store shared and discrete model parameters between target and source tasks separately. With the popularization of deep neural networks, the most popular transfer learning method, which is widely used, is a fine-tuning method that uses the weights of a trained model at the initial weights of a new model [25]. Studies on supervised transfer learning are categorized into diversity-based, contrast-based, and reconstruction-based methods by Zhang [22]. Examples of the most popular difference-based methods are those that transfer between tasks using class information [26], methods that measure statistical distribution difference with distance calculations [27] such as MMD (maximum mean discrepancy), and various transferable methods such as domain-specific layers and normalizations [28] in the model architecture.

Semi-supervised transfer learning is used successfully where large amounts of unlabeled or labeled data are available with a different distribution [23]. The methods used in this field can be examined in three groups methods that perform continuity regulation, methods that produce a common pseudo task or pseudo-label based methods, and generative methods. In continuity regulation-based methods, two different basic methods can be examined. The Minimum Class Confusion (MCC) method [26] is a method that increases the stability of the model over the unlabeled data by changing the model weights to make more stable predictions if the prediction of the model over the unlabeled data is not stable. Another approach uses methods such as randaugment [29], which make small and insignificant label changes. Minimizing the impact of these changes on class predictions is a popular method among semi-supervised transfer learning methods. Unlike these methods, an example of false label generation-based methods can be given as the noisy student method [30]. Unsupervised domain adaptation methods improve model performance on target domain data by utilizing source domain unlabeled data. Lee *and colleagues* [35] proposed a domain transfer method that utilizes source domain data by predicting each image label with maximizing prediction results. In this paper, we will explore the pseudo-labeling based domain adaptation method compared with MCC based method.

3. Methodology

3.1. Semantic Segmentation

The extraction of land cover maps from satellite images is defined as a semantic segmentation problem. This study used four semantic segmentation architectures: HRNet, Unet++, DeeplabV3+ and BisenetV1 architectures, and ResNet and Efficient-Net encoders. The flowchart of our semantic segmentation method is shown in Figure 1.

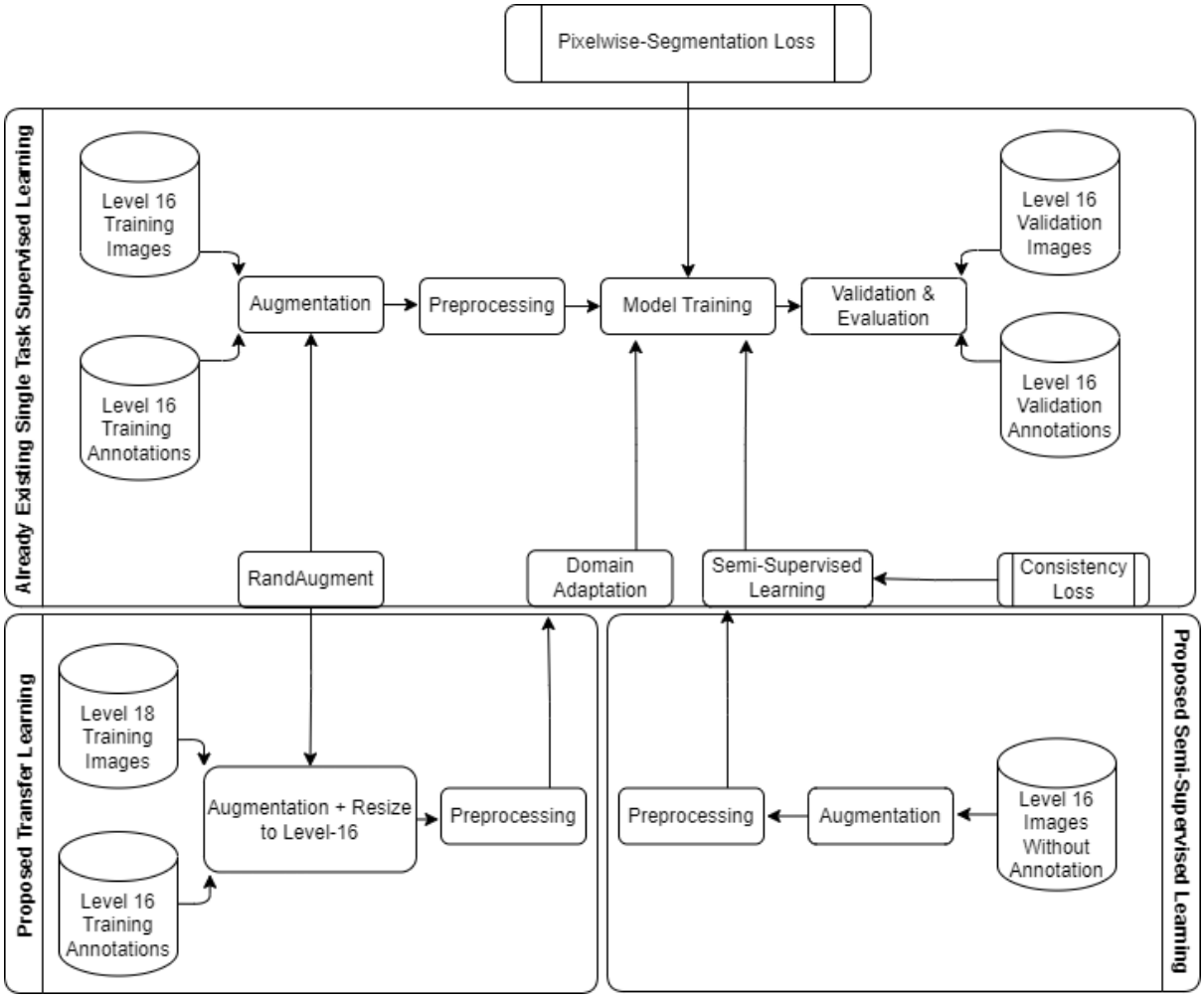

Figure 1: Semantic Segmentation Scheme.

A large part of the satellite images given as input to the spatial segmentation methods that we use consists of 5000x6000x3 channel RGB color pictures and labels of these pictures in the same size. In the first step of our training and testing method, a hyper-parameter optimization study was performed on these images for the training. As a result, all experiments were performed on 512x512 randomly cropped images taken over 1000x1000 images extracted in grid form from large images. In addition, a test is performed using the 512x512 sliding window method on 5000x6000 test images. Data augmentation is applied to cropped images to increase data diversity and make the trained models more robust to light, color, displacement, and rotation effects. The randaugment [20] method is used for data augmentation. In this method, increment methods such as translation, resizing, affine rotation, affine shearing, reversing the image, and random horizontal or vertical rotation were chosen according to our problem. Two data augmentation methods are randomly selected and applied to the image during the training. This method uses augmentation methods such as translation, resizing, affine rotation, affine shearing, reversing the image, and random horizontal or vertical rotation according to our problem. The degree of rotation is determined by taking ten equal sections between 0 and 150. As the training step progresses, the rotation limit is increased, and random selection is made from this section again. Following the data augmentation step, the images in the dataset are normalized with the mean and standard deviation values of the ImageNet dataset.

The Unet++ architecture, which is the first of the methods used during the training, consists of encoder and decoder structures similar to the letter U. Unet++ [5] added an array of nested convolution blocks prior to merging to the skip connection where the same size feature maps between encoder and decoder are transferred to reduce semantic information loss due to the separation of the encoder and decoder paths in the UNet structure, and updated these jumps to be more intense. The second architecture we use, the DeeplabV3+ method, uses methods such as atrous convolution and feature pyramiding compared to the Unet++ method. For this reason, it has advantages, particularly in the effect of objects being recognized by a wider area of influence on the image or in the recognition of objects of different sizes. Our experiments use Resnet50, Efficient-Net B3, and Efficient- Net B5 architectures as encoders. The Efficientnet architecture [32] is divided into two parts, convolution, point, and depth convolutions, thus reducing the computational cost. It was chosen primarily because of the ease with which the model complexity can be studied parametrically. HRNet architecture changes classical encoder-decoder architecture. This model uses s parallel encoder decoder path that shares representations parallelly using multi-scale parallel convolutions. HRNet model achieves the best building segmentation results in our experiments in urban cities.

3.2. Transfer Learning with Minimum Class Confusion

One of the main problems we encountered during the semantic segmentation training for satellite images is that the labels used in training need to be of more quantity and quality. In this study, the combined training method in which data from different sources are used at the same time in training in order to learn data from different sources together. Semi-supervised transfer teaching and learning methods using the Minimum Class Confusion (MCC) method [26] are used. In the standard training methodology used for all three methods, the data from the target and source datasets are sampled to each form half of the mini-batch during the deep neural network training. In the co-learning method, the total loss function is calculated over the signs y_t and y_s of the target data x_t and the source data x_s, as seen in Equation 1**.** The class label L_{SN} represents the loss function in the formula, and *f(x,w)* represents the model output obtained with *x* data. Moreover, w constitutes the learnable model parameters.

$$L_\varepsilon\big(y, f(x,w)\big) = L_{SN}\big(y_t, \mathrm{f}(x_t, \mathrm{w})\big) + L_{SN}\big(y_s, \mathrm{f}(x_s, \mathrm{w})\big) \quad (1)$$

The Minimum Class Confusion technique is a method that aims to minimize the classification error in the source data by converging the binary class confusion values of the estimates. This method inputs the output of a mini-batch of logits from the model and multiplies it by its inverse to obtain a correlation matrix that converges to the class confusion matrix. The process of unknown re-weighting is applied to normalize probabilities and to reflect more weight of important samples. Following this, a category normalization is performed in mini-batches to minimize the impact of the number of classes available on weights. In the matrix obtained at the end of these operations, classes showing high confusion with each other give higher correlation results. Since the values on the diagonal of the resulting matrix will represent the confusion of each class with itself, it will enable the model, which uses the loss function that aims to maximize these values or minimize the remaining values, to make more stable predictions on unlabeled data. In the MCC method in semi-supervised training, the MCC loss function L_{MCC} is used for the source data, and the class label loss function is used for the current target data, as

seen in Equation 2.

$$L_{\varepsilon}\big(y, f(x,w)\big) = L_{SN}\big(y_t, \mathrm{f}(x_t, \mathrm{w})\big) + L_{MCC}\big(y_s, \mathrm{f}(x_s, \mathrm{w})\big) \quad (2)$$

In the MCC-based transfer learning, the MCC loss function for the source data and the class label loss function for the whole data set are used, as shown in Equation 3.

$$L_{\varepsilon}\big(y, f(x,w)\big) = L_{SN}\big(y, \mathrm{f}(x, \mathrm{w})\big) + L_{MCC}\big(y_s, \mathrm{f}(x_s, \mathrm{w})\big) \quad (3)$$

3.3. Transfer Learning with Pseudo Labeling

We consider predicting regions from several regions from different map services problematic to our models. To deal with this problem, we enlarged our dataset with unlabeled images. Pseudo-labeling [35] is a semi-supervised learning method that models training performed simultaneously with labeled and unlabeled data. Class labels that maximize prediction probabilities for the unlabeled training supervised part are selected.

4. Experimental Results

4.1. Datasets

In this study, the Huawei Land Cover datasets HWLC16 and HWLC18 were created by Huawei, and the publicly available Geonrw and DeepGlobe datasets were used. The HWLC16 and HWLC18 datasets consist of satellite images taken from satellites by Huawei. The resulting images are at level 16 and level 18 according to OpenStreetmaps zoom levels [33]. The HWLC16 dataset consists of images with a resolution of 2.5 meters per pixel obtained from rural areas. This dataset contains 315 5000*6000 training and 110 test images with the exact resolution, and the labels have a lower resolution than the images. Each image has 21 classes of semantic segmentation labels with a resolution of 20 meters per pixel. Examples, class names, and colorings of this dataset are given in Figure \ref{fig:legend}. This dataset was collected from rural areas where high-resolution satellite imagery is difficult and costly to obtain. The images used as test sets are divided into rural and urban regions, the set on the city region is called *urban_test* in the tables, and the set on the rural region is called *rural_test*.

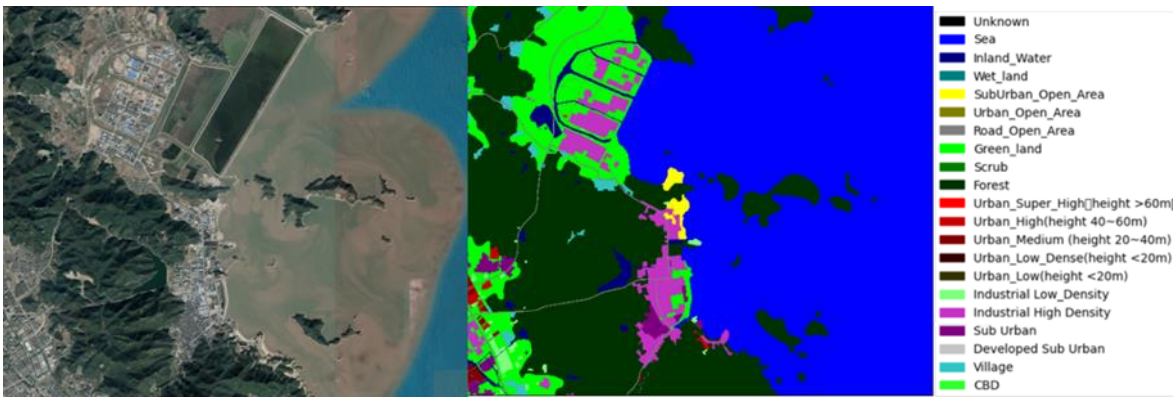

Figure 2: Example RGB image and 21-class semantic segmentation labels of dataset HWLC16.

The second dataset, the HWLC18 dataset, consists of level 18 satellite images, a more detailed zoom level than the HWLC16. Each image comprises RGB color images with a resolution of 0.5 meters, while the labels have a quality of detail of 2 meters. In the study, 413 pieces of 5000*6000 pictures are utilized. The HWLC18 dataset consists of images of metropolitan cities and their surroundings, not rural areas, unlike the level 16 HWLC16 dataset. The labels have much more detailed and high quality. However, due to the difficulty and cost of labeling at this level of detail, the total covered area is approximately 15 times smaller than the HWLC16 database. Apart from these two datasets, Geonrw [34] and DeepGlobe [20] datasets are also used in the study. The Geonrw database consists of 7783 images with a resolution of 1 meter per pixel, in 1000*1000 dimensions, and was created by combining images taken from aircraft of various cities in Germany and includes ten different classes. The training set of the DeepGlobe dataset contains 1000*1000 7227 images, and each pixel corresponds to 0.05 meters resolution. This dataset has seven different classes. All supervised learning methods used in this study are made by reducing them to 5 classes, which are the subclasses of the data used above. These classes are in order: [‘unknown’, 'urban', **'open_area'**, 'water', 'forest’]

4.2. Implementation Details, Evaluation Criteria, and Experiments

In this study, all the experiments are performed using Pytorch. The model structures and weights in the SMP library [36] are taken as initial values. In the experiments, Nvidia 2080TI graphics cards are used for training with Automatic Mixed Precision. The model performance is calculated with the intersection over union (IoU) value calculated as described in Equation 4 for all classes in each image in the experiments. The average IoU value for each image is found by averaging the IoU scores of all other classes except the unknown class of that image. The total performance measurement is made on the dataset by taking the average of these values for all test pictures.

$$IOU = \frac{TP}{TP+FP+TN} \quad (4)$$

The first experiment was conducted to determine the most appropriate semantic segmentation architecture to be used to extract land cover maps in this study. In this context, DeeplabV3+, Unet++, and BisenetV1 algorithms are tested with the Resnet50, EfficientnetB3, and EfficientnetB5 architectures, and their model capacities and suitability for the problem are compared. The test results are shown in Table 1.

Table 1: The results of the experiments performed on the HWLC16 dataset with the architectures of different models.

Model	**Encoder**	**Rural Area Dataset**					**City**
		mIoU	**Urban**	**Open Area**	**Forest**	**Water**	**Building**
BisenetV1	Resnet18	0.558	0.55	0.372	0.807	0.504	0.721
Unet++	Resnet50	0.5874	**0.6128**	0.3636	0.7564	0.6167	0.76
Unet++	EfficientnetB3	0.602	0.595	0.339	0.84	0.634	0.796
Unet++	EfficientnetB5	0.592	0.582	0.324	0.846	0.618	0.787
Deeplab+	Resnet50	0.458	0.471	0.249	0.704	0.331	0.703
Deeplab+	EfficientnetB3	0.5856	0.584	0.3375	0.8267	0.594	0.68
Deeplab+	EfficientnetB5	0.592	0.582	0.323	0.846	0.617	0.762
HRNet	EfficientnetB5	**0.603**	0.582	**0.521**	**0.863**	**0.702**	**0.863**

It is experienced that the model trained using the Unet++ architecture and the Efficient-Net encoder showed the most successful results in both the HWLC16 rural area test set and urban area dataset. The capacity of this model is higher than the BiseNetv1 model and smaller than the DeepLabV3+ model. When the outputs of different models are examined, the U-Net model can find sharper boundaries in segmenting small objects on the image, especially buildings, and inner-city neighborhoods. The failure to train more complex models such as DeeplabV3+ and EfficientnetB5 Encoder with the same success may be due to a lack of data, the excess of annotation errors in the dataset, or the fact that the training process is not long enough to find the best parameters.

Following the architectural analysis, the capacity and best performance of the model were measured in training using a single database. To improve this performance, combined training, semi-supervised learning, and transfer learning methods were applied to use information from other datasets with similar data, and their improvements were compared. The results of these experiments are given in Table II.

Table II: The table gives results on two test sets of the target HWLC16 dataset using different source data. The HWLC18->16 dataset contains images four times smaller than HWLC18.

Source Dataset	**Transfer Method**	**Rural Area**					**Urban Area**
		mIoU	**Building**	**Open Area**	**Forest**	**Water**	**Building**
-	-	0.602	0.595	0.339	0.84	0.634	0.796
HWLC18	**Combined Training**	0.617	0.567	0.391	0.87	0.641	0.812
HWLC18	MCC Semi-Supervised	0.632	0.614	0.405	0.849	0.66	0.869
HWLC18	MCC Transfer Learning	0.593	0.545	0.371	0.852	0.606	0.518
HWLC18->16	Combined Training	0.616	0.62	0.4	0.857	0.586	0.841
HWLC18->16	MCC Semi-Supervised	0.615	0.621	0.395	0.844	0.6	0.861
HWLC18->16	MCC Transfer Learning	0.636	0.621	0.422	0.861	0.64	0.901
Geonrw	Combined Training	0.62	0.616	0.372	0.826	0.666	0.842
Geonrw	MCC Semi-Supervised	0.573	0.546	0.269	0.856	0.622	0.925
Geonrw	MCC Transfer Learning	0.624	0.584	0.406	0.847	0.659	0.877
DeepGlobe	Combined Training	0.612	0.601	0.409	0.822	0.616	0.901
DeepGlobe	MCC Semi-Supervised	0.615	0.615	0.368	0.823	0.653	0.887
DeepGlobe	MCC Transfer Learning	0.533	0.583	0.121	0.837	0.593	0.923

The first row in Table II gives the results from the supervised learning problem without using any source data. The HWLC18 database, which is in the source dataset, is labeled similarly to the HWLC16. The most significant difference between the two datasets is the city/rural area differences they cover and the size of the images. As seen in experiments with HWLC18 dataset without resizing, semi-supervised learning provides the most remarkable performance increase using only images without labels. Combined training and transfer learning cannot increase prediction performance due to differences in the distribution of labels, reducing it from 60.2% to 59.3% in rural areas and from 79.6% to 51.8% in urban areas.

We see that the transfer learning performance exceeds both methods in the experiments where we equate the image and label sizes, the main difference between these two datasets, and show them in the table with HWLC18->16. In these experiments, the success reached the highest at 63.6%, especially in rural areas targeted by the HWLC16 dataset.

Transfer learning experiments on the Geonrw dataset showed us the effect of differences between labelings on transfer success from different performances in rural and urban regions. While the Geonrw database is labeled with similar annotation rules as HWLC16 in land cover classes, it shows significant differences in the labeling of building regions. The areas between buildings are labeled as buildings due to low resolution in HWLC16. Geonrw has more detailed annotations in areas such as industrial regions and airports. The differences between the class definitions caused the transfer to perform worse than semi-supervised learning, especially in the urban dataset. In contrast, semi-supervised learning performed about 5% below the combined training and transfer learning in rural areas.

Similarly, experiments with DeepGlobe dataset resulted in a 92.3% transfer learning success because the building labels in this dataset are similar to our problem. However, we suspect that the differences between the target areas and labelings of the datasets cause the targeted performance in rural areas to be lower than other datasets. We implemented the HRNet model to analyze building segmentation performance comparison with other models. Compared to other approaches, this model excelled in city regions, not rural area buildings. Table III shows that HRNet achieved nearly a 10% increment in IoU score in building segmentation. This model seems promising, and we will use this model for unsupervised domain adaptation model comparisons.

Table III: THE BUILDING SEGMENTATION RESULTS OF THE EXPERIMENTS PERFORMED ON THE HWLC16 DATASET WITH THE ARCHITECTURES OF DIFFERENT MODELS.

Model	**Encoder**	**Rural Area**	**City**
		Village	**Building**
BisenetV1	Resnet18	0.55	0.721
Unet++	Resnet50	0.613	0.76
Unet++	EfficientnetB3	0.595	0.796
Unet++	EfficientnetB5	0.582	0.787
Deeplab+	Resnet50	0.471	0.703
Deeplab+	EfficientnetB3	0.584	0.68
Deeplab+	EfficientnetB5	0.582	0.762
HRNet	EfficientnetB5	0.578	0.863

As in the previous experiments, we trained the HRNet model with the Efficientnet-B5 backbone with different supervision training types. Table IV shows the results of supervised training, pseudo-labeling-based unsupervised domain adaptation, and MCC-based unsupervised domain adaptation results. Pseudo-labeling-based transfer learning performed better than supervised and MCC-based unsupervised training. Also, this table states that both transfer learning experiments improved the results compared to supervised training.

Table IV: PSEUDO-LABELING TEST RESULTS.

Source Dataset	Target Dataset	Transfer Method	Number of Class	Rural Region	Urban Area
				Village	Building
-	HWLC18->16, HWLC16	Combined Training	4-Class	0.582	0.863
Unlabeled HW	HWLC18->16, HWLC16	Pseudo Labeling	4-Class	0.601	0.914
Unlabeled HW	HWLC18->16, HWLC16	Unsupervised MCC	4-Class	0.599	0.890

5. Conclusion

In this study, we present the results of our research on using semantic segmentation methods with various multi-source prediction methods for extracting land cover maps from satellite images on the HWLC16 and HWLC18 datasets.

In the experiments, the segmentation success of urban areas is 79.6 %, and in rural areas, 60.2 % was reached with the original resolution in the HWLC16 dataset. The addition of the high-resolution HWLC18 dataset, which has been taken from various sources and has the same annotation rules, through transfer learning increased segmentation performance to 90.1 % for urban areas and 63.6 % for rural areas in the same test set. Geonrw and DeepGlobe databases are used in training using a semi-supervised learning method because although these datasets have high resolution, they are annotated with different labeling rules and classes. The possible reason for this can be that transfer between these tasks leads to negative transfer. For these experiments, the urban area dataset performance increased to 92.3 %.

In future work, we aim to improve the diversification of transfer learning and semi-supervised learning methods and improve the minimum class confusion method designed for classification by subjecting it to appropriate normalizations per the segmentation problem. Indeed we will modify minimum class confusion loss to be able to change priority according to epochs. Additional experiments with HRNet showed us that multi-scale fusion between encoder and decoder architectures in an end-to-end manner improved semantic segmentation results dramatically. To conclude, using several transfer learning approaches, we will conduct more experiments with this architecture to improve semantic segmentation results in rural and urban areas.

References

[1] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang, W. Liu, and B. Xiao, “Deep high-resolution representation learning for visual recognition,” CoRR, vol. abs/1908.07919, 2019. [Online]. Available: http://arxiv.org/abs/1908.07919

[2] G. Druck, B. Settles, and A. McCallum, “Active learning by labeling features,” in Conference on Empirical Methods in Natural Language Processing, Aug. 2009, pp. 81–90.

[3] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 3431–3440.

[4] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in MICCAI2015. Springer, 2015.

[5] Z. Zhou, M. M. Rahman Siddiquee, N. Tajbakhsh, and J. Liang, “Unet++: A nested u-net architecture for medical image segmentation,” in Deep learning in medical image analysis and multimodal learning for clinical decision support. Springer, 2018, pp. 3–11.

[6] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, “Semantic image segmentation with deep convolutional nets and fully connected crfs,” arXiv preprint arXiv:1412.7062, 2014.

[7] L.-C. Chen, G. Papandreou, I. Kokkinos, and K. Murphy, “Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs,” IEEE transactions on

PAMI, vol. 40, no. 4, pp. 834–848, 2017.

[8] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," arXiv preprint arXiv:1706.05587, 2017.

[9] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder- decoder with atrous separable convolution for semantic image segmentation," in ECCV, 2018, pp. 801–818.

[10] B. Neupane, T. Horanont, and J. Aryal, "Deep learning-based semantic segmentation of urban features in satellite images: A review and meta analysis," Remote Sensing, vol. 13, no. 4, p. 808, 2021.

[11] Z. Guo, H. Shengoku, G. Wu, Q. Chen, W. Yuan, X. Shi, X. Shao, Y. Xu, and R. Shibasaki, "Semantic segmentation for urban planning maps based on u-net," in IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium. IEEE, 2018, pp. 6187–6190.

[12] W. Li, C. He, J. Fang, J. Zheng, H. Fu, and L. Yu, "Semantic segmentation-based building footprint extraction using very high- resolution satellite images and multi-source gis data," Remote Sensing, vol. 11, no. 4, p. 403, 2019.

[13] Y. Yi, Z. Zhang, W. Zhang, and C. Zhang, "Semantic segmentation of urban buildings from vhr remote sensing imagery using a deep convolutional neural network," Remote sensing, vol. 11, no. 15, 2019.

[14] S. Du, S. Du, B. Liu, and X. Zhang, "Incorporating deeplabv3+ and object-based image analysis for semantic segmentation of very high resolution remote sensing images," International Journal of Digital Earth, vol. 14, no. 3, pp. 357–378, 2021.

[15] D. Hong, Z. Han, J. Yao, L. Gao, B. Zhang, A. Plaza, and J. Chanussot, "Spectralformer: Rethinking hyperspectral image classification with transformers," IEEE Transactions on Geoscience and Remote Sensing, vol. 60, pp. 1–15, 2021.

[16] Z. Xu, W. Zhang, Z. Yang, and J. Li, "Efficient transformer for remote sensing image segmentation," Remote Sensing, vol. 13, p. 3585, 2021.

[17] L. Scheibenreif, J. Hanna, M. Mommert, and D. Borth, "Self-supervised vision transformers for land-cover segmentation and classification," in CVPR, 2022, pp. 1422–1431

[18] G. Baier, A. Deschemps, M. Schmitt, and N. Yokoya, "Building a parallel universe—image synthesis from land cover maps and auxiliary raster data," CoRR, 2020.

[19] E. Maggiori, Y. Tarabalka, G. Charpiat, and P. Alliez, "The inria aerial image labeling benchmark," in IEEE International Geoscience and Remote Sensing Symposium (IGARSS). IEEE, 2017.

[20] I. Demir, K. Koperski, D. Lindenbaum, G. Pang, J. Huang, S. Basu, F. Hughes, D. Tuia, and R. Raskar, "Deepglobe 2018: A challenge to parse the earth through satellite images," in Proceedings of the IEEE CVPR Workshops, 2018, pp. 172–181.

[21] M. Wang and W. Deng, "Deep visual domain adaptation: A survey," Neurocomputing, vol. 312, pp. 135–153, 2018.

[22] J. Zhang, W. Li, P. Ogunbona, and D. Xu, "Recent advances in transfer learning for cross-dataset visual recognition: A problem-oriented perspective," ACM Computing Surveys, vol. 52, pp. 1–38, 2019.

[23] Y. Ouali, C. Hudelot, and M. Tami, "An overview of deep semi- supervised learning," arXiv preprint arXiv:2006.05278, 2020.

[24] G. Wilson and D. J. Cook, "A survey of unsupervised deep domain adaptation," ACM Transactions on

Intelligent Systems and Technology (TIST), vol. 11, no. 5, pp. 1–46, 2020.

[25] G. Hinton et al., "How to do backpropagation in a brain," in Invited talk at the NIPS'2007 deep learning workshop, vol. 656, 2007, pp. 1–16.

[26] Y. Jin, X. Wang, M. Long, and J. Wang, "Minimum class confusion for versatile domain adaptation," in European Conference on Computer Vision. Springer, 2020, pp. 464–480.

[27] M. Long, H. Zhu, J. Wang, and M. I. Jordan, "Deep transfer learning with joint adaptation networks," in ICML, 2017.

[28] Y. Li, N. Wang, J. Shi, J. Liu, and X. Hou, "Revisiting batch normalization for practical domain adaptation," arXiv:1603.04779, 2016.

[29] E. D. Cubuk, B. Zoph, J. Shlens, and Q. V. Le, "Randaugment: Practical automated data augmentation with a reduced search space," in CVPRW, 2020, pp. 702–703.

[30] Q. Xie, M.-T. Luong, E. Hovy, and Q. V. Le, "Self-training with noisy student improves imagenet classification," in CVPR, 2020.

[31] Lee, Dong-Hyun. "Pseudo-Label : The Simple and Efficient Semi- Supervised Learning Method for Deep Neural Networks." (2013).

[32] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in ICML. PMLR, 2019, pp. 6105–6114.

[33] Openstreetmap contributors, "Zoom levelsopenstreetmap," 2022, [Online; accessed 6-June-2022]. [Online]. Available: https://wiki.openstreetmap.org/wiki/Zoom levels

[34] G. Baier, A. Deschemps, M. Schmitt, and N. Yokoya, "Geonrw," 2020.

[35] Lee, Dong-Hyun. "Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks." In Workshop on challenges in representation learning, ICML, vol. 3, no. 2, p. 896. 2013.

[36] P. Yakubovskiy, "Segmentation models pytorch," 2020. [Online]. Available: https://github.com/qubvel/segmentation models.pytorch