Exploration of Resnet Variants in High Spatial Resolution Domain Adaptation: From air-to-space imagery

S P Widodo[1,2,*], **N Rachmawati**[1,2]

[1] BPS-Statistics Indonesia, Jl. Dr. Sutomo 6-8, Jakarta, Indonesia
[2] Universitas Indonesia, Depok, Indonesia

*Corresponding author's email: sulisetyo.widodo@bps.go.id

Abstract. Land cover is nowadays mapped mostly from airborne and space-borne data. Because of the difference in sensors, large spectral differences and inconsistent spatial resolution may arise between these two data sources. Consequently, the same object may exhibit completely different features. In this case, models trained from annotated airborne and ineffective when applied to space-borne data. Cross-Sensor Land-COVER (LoveCS) shows good results in overcoming this problem. LoveCS leverages small-scale aerial image annotations to promote land cover mapping on large-scale spacecraft. LoveCS uses ResNet50 as its encoder. In recent years, many studies have tried to develop other variants of ResNet, such as ResNeXt, ResNeSt, Res2Net, and Res2NeXt. These variants turned out to give better results in a variety of tasks compared to ResNet. Therefore, in this study we modified the LoveCS encoder by replacing ResNet50 with ResNet variants such as ResNeXt, ResNeSt, Res2Net, and Res2NeXt in an effort to improve LoveCS accuracy. Our evaluation shows that Res2Net50 as an encoder improves the performance of LoveCS. The average F1 increases by 1.38%, OA by 1.96%, and Kappa by 2.75% from the baseline method.

1. Introduction

Land cover information is very important for resource allocation and sustainable development. Deep learning has also shown promising results in land cover mapping with high spatial resolution (HSR). In terms of large-scale mapping, spaceborne (spaceborne) data has more advantages because it covers a large area. Unfortunately, more data regarding information (land cover) is obtained from the air (airborne) than spaceborne. This is due to the convenience and flexibility of aerial photogrammetry, which makes airborne data easy to obtain [1] [2].

Unfortunately, when mapping the land cover from airborne data to spaceborne data, Figure 1. This sensor difference shows a problem, namely there is a large spatial resolution inconsistency and spectral difference. With this problem, the same object can show completely different features [3]. This causes the model trained from annotated airborne to always lose its effectiveness when applied to images from spaceborne [4].

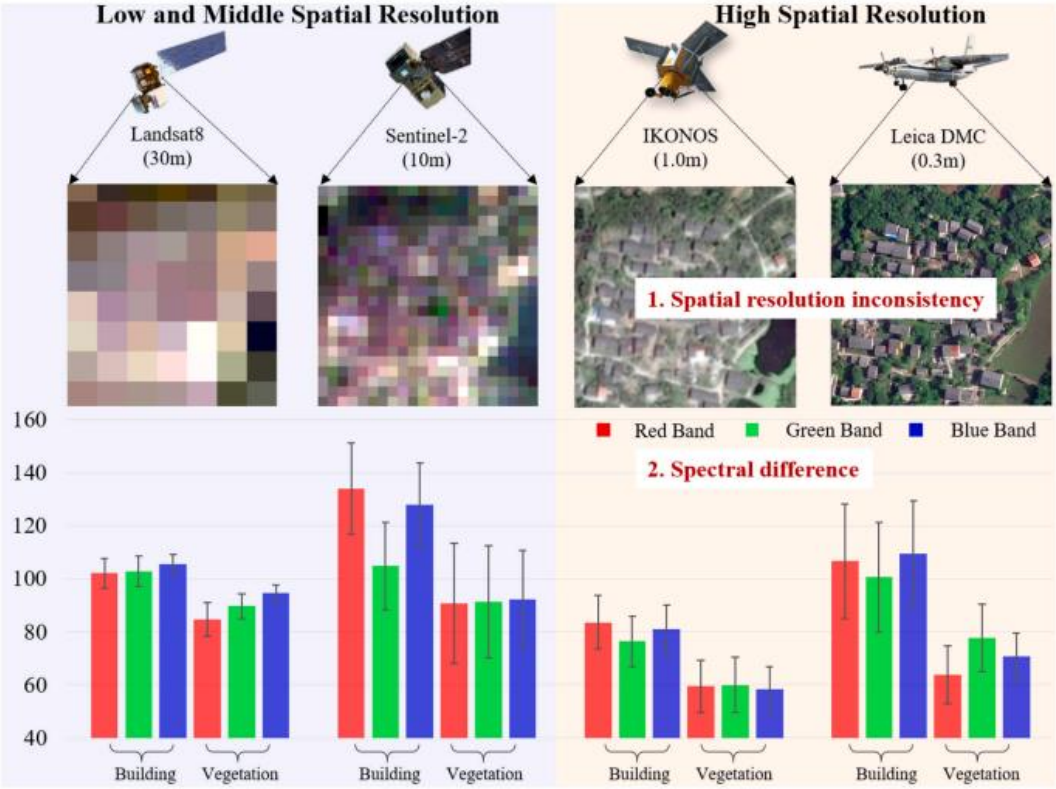

Figure 1. Spatial resolution and spectral differences between different sensors, figure from [5].

A study from [5] proposed a framework called Cross-Sensor Land-COVER (LoveCS). LoveCS is an Unsupervised domain adaptation (UDA) method that introduces a cross-sensor encoder (CSEncoder) and a dense multi-scale decoder (DMSDecoder). This study shows good results in overcoming spatial resolution and spectral differences by utilizing small-scale airborne image annotations to promote land cover mapping on large-scale spaceborne [5].

The LoveCS framework uses the ResNet model built by [6] as the backbone in the CSEncoder. Then gradually research on ResNet turned out to be very developed. [7] proposed ResNeXt, a ResNet-based model that uses grouped convolutions. ResNeXt shows that increasing cardinality will lead to better classification accuracy. In addition, [8] also proposed another variant of ResNet named ResNeSt. This model universally improves the representation of the features studied. The goal is none other than to improve performance across image classification, object detection, instance segmentation, and semantic segmentation. Then Gao et al. [9] proposed two other ResNet variants, namely Res2Net and Res2NeXt. These two models divide the input feature maps into several groups. Res2Net and Res2NeXt consistently perform well compared to state-of-the-art, including ResNet [6], ResNeXt [7], DLA, etc.
In this study, there are several research questions (RQ) that we have investigated:

- RQ1: How to optimize the LoveCs framework?
- RQ2: How to implement Resnext, ResNeSt, Res2Net, and Res2NeXt as a replacement for CSEncoder in LoveCs?

Then our contributions to this research are: (1) we modified the LoveCS framework by changing the CSEncoder section as an effort to improve the accuracy of LoveCS; (2) we did a comparison of the accuracy of the modified results to find the LoveCS scheme that had the best accuracy results; (3) we also offer a modified LoveCS scheme to utilize small-scale airborne images with labels as a source for classifying unlabeled large-scale spaceborne images.

2. Literature Review

2.1. Unsupervised domain adaptation (UDA) In Remote Sensing

UDA, short for Unsupervised Domain Adaptation, is a subclass of transfer learning that is used to transfer a learned model in a labeled source domain to an unlabeled target domain. If classified, there are two types of UDA that exist, namely adversarial training, and self training [5]. Adversarial training is Unsupervised Domain Adaptation which uses a discriminator to predict the domain label and a feature extractor to generate invariant features which are later used to confuse the discriminator. Then self-training is Unsupervised Domain Adaptation which alternately assigns pseudo-labels to unlabeled data. After that this method will use pseudo samples to refine the existing model flow [5].

There are several recent studies regarding UDA in remote sensing. In recent years, several domain adaptive segmentation methods have been proposed, such as the study by Wittich et al. [10]. Then there is also research from Wang et al. [11] who have built a land cover dataset for Domain Adaptive (LoveDA) semantic segmentation. In experimentation, this data set was able to provide a common benchmark for extending existing UDA algorithms and advancing urban-rural domain adaptation.

In addition, the study proposed by Tong et al. [4] tried to adopt a self-training strategy within the land cover mapping framework. It turns out that this strategy can improve performance significantly within the land cover mapping framework. Iqbal and Ali [12] also proposed a weakly supervised domain adaptation network based on adversarial learning for the constructed region segmentation. Then, Lu et al. [13] have proposed the Global-local Adversarial Learning (GOAL) method to focus more on difficult road samples when segmenting roads across domains.

But unfortunately, from all the recent research on UDA above, there is no research that considers cross-sensor mapping with problems of inconsistency of spatial resolution and spectral differences. Good news also comes from research conducted by Wang et al. [5]. This researcher proposes a framework called LoveCS. This LoveCS framework proposes a scheme to overcome the problems associated with spatial resolution inconsistencies and spectral differences by leveraging small-scale airborne image annotations for large-scale spaceborne land cover mapping.

Figure 2 illustrates the existing system architecture in LoveCS. There are two main sections in LoveCS, namely Cross Sensor Network (CSN) and Multi Scale Pseudo Label (MSPL). CSN consists of CSEncoder and DMSDecoder. CSEncoder implements ResNet50 as a backbone. Then DMSDecoder is used to learn domain-invariant discriminatory features and semantics. Then for the optimization of the model adopting domain self-training adaptation with MSPL which can later overcome the problem of spatial resolution inconsistency. A comprehensive Unsupervised Domain Adaptation Experiment conducted with data from three different cities in China proved the excellent performance and generalizability of LoveCS [5].

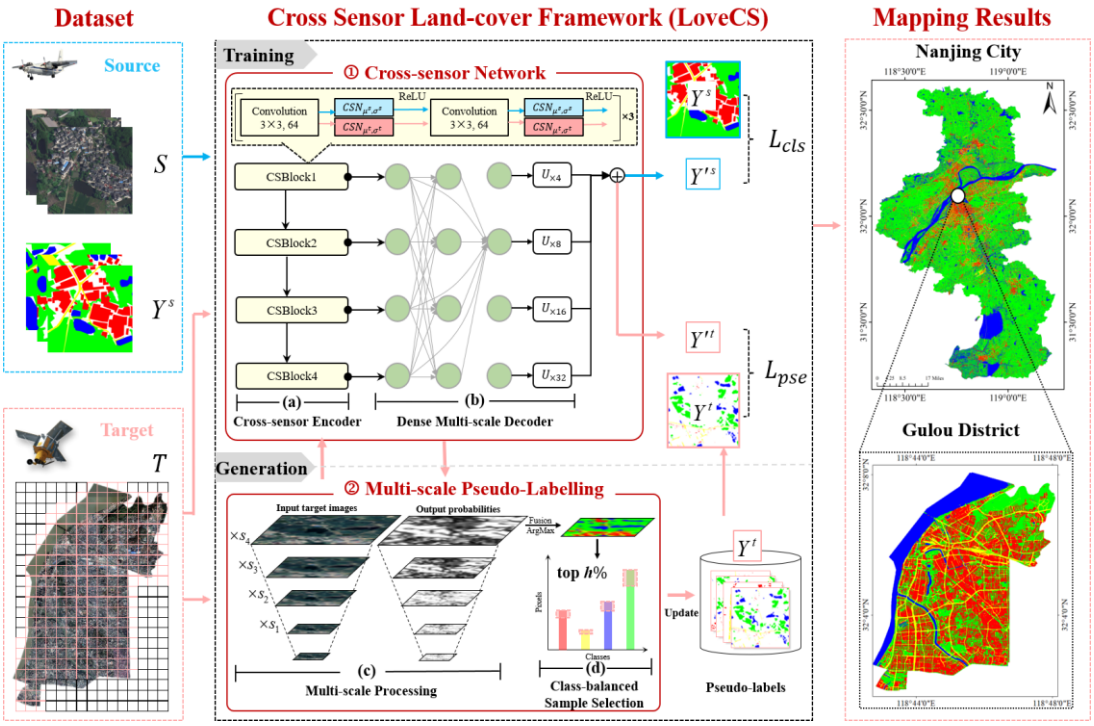

Figure 2. LoveCS framework proposed by [5].

The whole process in LoveCS can be seen in Figure 3, namely:

- (1) Initialization model: the cross-sensor network is trained on the data source only (S).
- (2) Multiscale pseudo-labeling: based on the model trained, Yt pseudo-labels are then generated via MSPL.
- (3) Training Model: training is carried out in.
- Steps (2) and (3) then carried out alternately until the end of the training.
- (4) Inference model: after training, the model will predict the target image based on the target. Then MSPL is performed without class-balanced sample selection.

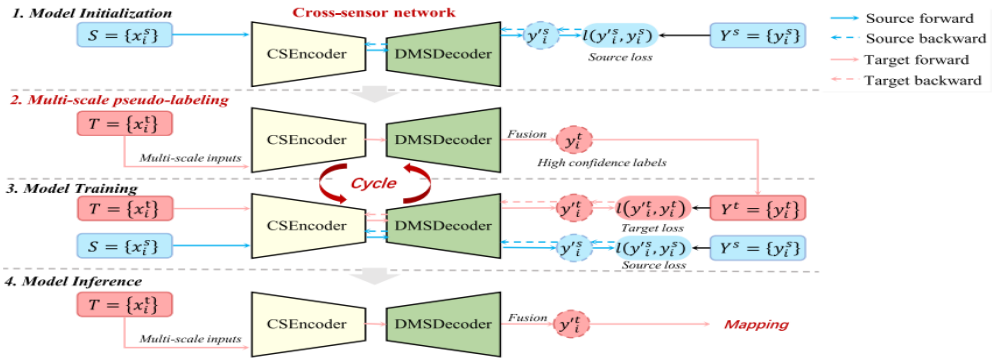

Figure 3. The whole process in LoveCS proposed by [5].

In this research, we focus on research on Unsupervised Domain Adaptation (UDA) in remote sensing, especially on how to optimize the LoveCS framework to get better performance.

2.2. ResNet variant

Deep Neural Networks or DNN is usually very difficult to do when facing the training process. Therefore [6] introduced ResNet, a residual learning framework that aims to facilitate the network training process whose substance is much deeper. This ResNet is then used within the LoveCS framework as the backbone in CSEncoder. The ResNet model has evolved gradually. [7] introduced another variation of ResNet, namely ResNeXt. This model incorporates ResNet's iterative layer strategy then introduces a simple and extensible way to implement split, transform, and merge strategies. The number of paths in a ResNeXt block is defined as the cardinality or referred to as the C symbol.

ResNet [6] and ResNeXt have different widths. Layer-1 in ResNet has one convolution layer with a width of 64, while layer-1 in ResNext has 32 different convolution layers with a width of 4 (width 32*4 or with C=32 and d=4). Even though the overall width is larger in ResNeXt, both architectures share the same number of parameters which is around 70k. ResNeXt showed better results than state-of-the-art, namely ResNet, Inception-v3/v4, and Inception-ResNet-v2 on the ImageNet-5K dataset and the COCO detection dataset [7].

[8] introduced another variation of ResNet [6], namely ResNeSt. This study introduces a split-attention-block consisting of a group of feature maps and a split-attention operation. This research presents a modulated architecture that applies channel-based attention to various network branches. The goal is to increase the model's success in capturing interactions across features and learning from multiple representations. This design results in a simple, unified computation block that can be parameterized using only a few variables.

In an experiment conducted by [8], ResNeSt improves the representation of learned features so that it can improve performance across image classification, object detection, instance segmentation, and semantic segmentation. Compared to ResNet [6], ResNeSt is able to increase the Mean Average Precision (MAP) by about 3% on Faster-RCNN and Cascade-RCNN. The experimental results in this study show that ResNeSt has good generalization abilities. For example, ResNeSt50 outperforms ResNet101 for the Faster-RCNN and Cascade-RCNN detection models using significantly fewer parameters [8].

Subsequent research on variations of ResNet [6] is Res2Net. [9] proposed a new building block for CNN, namely Res2Net. The Res2Net model developed in this study seeks to increase the ability of multiscale representation at a more granular level. This multiscale representation refers to the multiple receptive fields available at a more granular level. To achieve this goal, this study replaces a 3x3 filter of n channels, with a set of smaller filter groups, each with w channels (without loss of similarity, where $n = s \times w$). s is the number of scale dimensions.

The proposed Res2Net blocks can be plugged into advanced CNN backbone models, for example, ResNet [6], ResNeXt [7], and DLA. Res2Net combined with ResNeXt makes it the Res2NeXt model. This research evaluates the Res2Net block on all of these models and shows consistent performance improvements over the base model on widely used data sets, eg, ImageNet [9].

In this research, we try to modify the CSEncoder backbone in the LoveCS framework. CSEncoder which previously used ResNet will be replaced by using other ResNet variations, namely ResNeXt [7], ResNeSt [8], Res2Net [9], and Res2NeXt [9].

3. Research Method

3.1. Dataset

We used the Land-COVER Domain Adaptive semantic segmentation (LoveDA) dataset [11] which is suitable for semantic assignment of land cover and unsupervised domain adaptation (UDA). LoveDA contains 5,987 HSR images with 166,768 annotated objects from three different cities. LoveDA covers two domains (urban and rural) which pose great challenges due to: 1) multi-scale objects; 2) complex background samples; and 3) inconsistent class distribution. The LoveDA dataset is then divided into

three parts, namely the training dataset, the validation dataset, and the test dataset. The training dataset consists of 1,366 images. The validation dataset consists of 677 images. Then the test dataset consists of 677 images.

3.2. Baseline

This study will use the LoveDA dataset [11]. Then, in order to fairly compare the accuracy results, we also retrained the model on LoveCS [5] using the LoveDA dataset [11]. Furthermore, the evaluation results will be used as a benchmark in this study.

3.3. Proposed Model

This study focuses on modifying LoveCS [5] in the CSEncoder section by replacing ResNet50 [6] with ResNext [7], ResNeSt [8], Res2Net [9], and Res2NeXt [9]. Each variant will be trained independently.

3.4. Evaluation matrix

We will also use the same matrix as that used in LoveCS [5] so that evaluation results can be measured fairly. The matrix used is:

- Average F1: Harmonic average of manufacturer's accuracy (PA) and user's accuracy (UA) as in equation (3.1).
- Overall Accuracy: the ratio of correctly classified pixels to all pixels in the entire test set as in equation (3.2).
- Kappa: a statistically based confusion matrix that measures overall classification accuracy as in equation (3.3).

$$F_1 = \frac{\sum_{i=1}^{n} x_{ii}}{\sum_{i=1}^{n} x_{ii} + \frac{1}{2}\left(\sum_{i=1}^{n}\sum_{j\neq 1}^{n} x_{ij} + \sum_{i=1}^{n}\sum_{j\neq 1}^{n} x_{ji}\right)} \tag{3.1}$$

$$OA = \frac{\sum_{i=1}^{n} x_{ii}}{\sum_{i=1}^{n}\sum_{j\neq 1}^{n} x_{ij}} \tag{3.2}$$

$$Kappa = \frac{M\sum_{i=1}^{n} x_{ii} - \sum_{i=1}^{n}\left(\sum_{j=1}^{n} x_{ij} \times \sum_{j=1}^{n} x_{ji}\right)}{M^2 - \sum_{i=1}^{n}\left(\sum_{j=1}^{n} x_{ij} \times \sum_{j=1}^{n} x_{ji}\right)} \tag{3.3}$$

3.5. Experimental Setup

In general we use the same settings and values as LoveCS [5]. Parameters used include batch size, number of workers, learning rate, iteration, warm-up step, pseudo generation per unit, and weight. However, due to limited resources, we reduced the number of workers to 0. We also changed the ResNet model to ResNeXt50 32x4d, ResNeSt50d, Res2Net50 26w 4s, and Res2NeXt50.

Table 1. Experimental Setup

Parameter	Setup
BATCH SIZE (Train Val)	8
BATCH SIZE (Test)	1
NUM WORKERS	0
LEARNING RATE	1E-2
ITERASI	15000
WARMUP STEP	10000
GENERATE PSEDO EVERY	2000
WEIGHTS	IMAGENET
RESNET VARIANT	ResNeXt50 32x4d; ResNeSt50d; Res2Net50 26w 4s; Res2NeXt50

3.6. Results

The results of the LoveCS training with the ResNet variation as the encoder can be seen visually in Figure 4. Figure 4 shows two sample segments of each ResNet variant. If observed by naked eye, we will find it difficult to find which one is good among these variants. However, based on table 1, in general the ResNet variant provides increased accuracy from the baseline method except for ResNeSt50d which actually provides the lowest accuracy value of the three evaluation metrics. ResNeSt50d is at 74.19% for the F1 average, 77.75% for the OA, and 64.31% for the Kappa. Actually, from the number of parameters, Res2NeXt50 has the fewest parameters but it doesn't seem able to describe the best accuracy. Likewise, ResNeSt50d which has the highest number of parameters also cannot provide the best accuracy value. This shows that the number of parameters does not determine the accuracy of LoveCS. The best accuracy was actually obtained by Res2Net50 26w 4s where the average F1 value was 80.01%, OA 83.68% , and Kappa 72.90%. In addition to providing the best accuracy, Res2Net50 26w 4s actually has relatively the same training time and a number of parameters as the others.

Table 2. Evaluation Result

Encoder	Parameter	Training time (hh:mm:ss)	F1 per class (%)				F1 mean (%)	OA (%)	Kappa (%)
			Pervious	Building	Road	Water			
ResNet50 [Baseline]	28.481 M	08:47:48	85.90	82.38	75.16	71.07	78.63	81.72	70.15
ResNeXt50 32x4d	27.953 M	10:05:49	87.28	85.37	71.31	74.90	79.71	83.32	72.65
ResNeSt50d	30.407 M	09:51:41	83.12	80.64	70.05	62.93	74.19	77.75	64.31
Res2Net50 26w 4s	28.623 M	09:39:50	87.63	82.63	72.65	77.15	80.01	83.68	72.90
Res2NeXt50	27.595 M	09:25:52	85.36	83.76	76.00	66.12	77.81	80.88	68.46

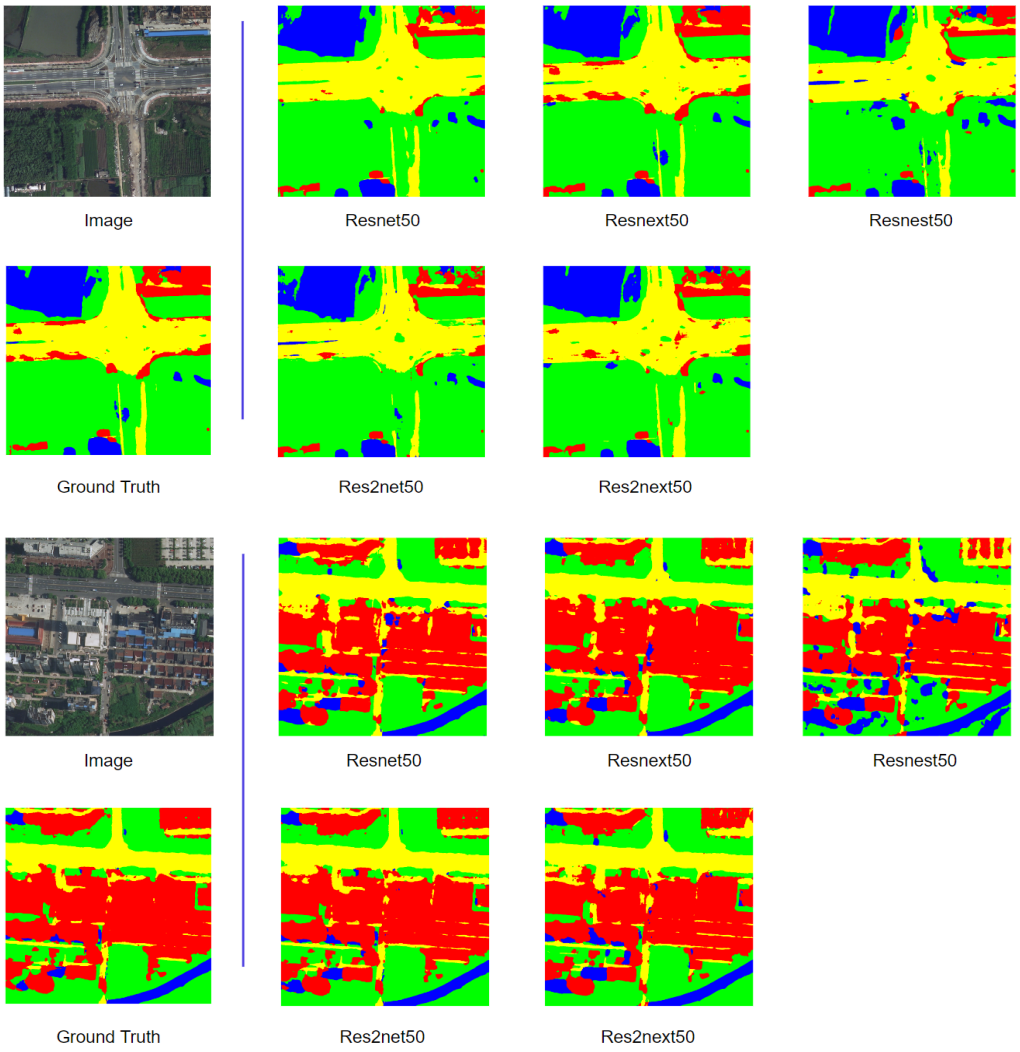

Figure 4. Two segmentation samples of each ResNet encoder variant.

4. Conclusion

Replacing CSEncoder with another variant of ResNet shows that LoveCS performs better than the base method, namely ResNet50 on average F1, OA, and Kappa values. Res2Net50 has relatively the same parameters and training time as other encoder models but shows the best results where the average value of F1 is 80.01%, OA is 83.68%, and Kappa is 72.90%. The average F1 increased by 1.38% and OA by 1.96% and Kappa by 2.75% from the baseline method. Multiscale Representation Capability on Res2Net50 can improve encoder capability on LoveCS. Then as an effort to improve LoveCS performance and based on the results of the evaluation that has been done, we recommend Res2Net50 as a replacement for ResNet50 in the LoveCS encoder.

5. Future Work

For future research, we plan to conduct more extensive experiments. We want to try to optimize the LoveCS framework for ResNet variants with more layers like ResNet101, ResNeXt101, ResNeSt101, Res2Net101, Res2NeXt101, etc. Apart from that, it is also possible to optimize other parts of LoveCS such as Decoder or Multi Scale -Pseudo-Labeling.

References

[1] Zhang C, Pan X, Li H, Gardiner A, Sargent I, Hare J, and Atkinson P M 2018 A hybrid mlp-cnn classifier for very fine resolution remotely sensed image classification *ISPRS Journal of Photogrammetry and Remote Sensing* vol 140 pp 133–144

[2] Zhang X, Du S, and Wang Q 2018 Integrating bottom-up classification and top-down feedback for improving urban land-cover and functionalzone mapping *Remote Sensing of Environment* vol 212 pp 231–248

[3] Russwurm M, Wang S, Korner M, and Lobell D 2020 Meta-learning for few-shot land cover classification *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops* vol 212 pp 200–201

[4] Tong X -Y,. Xia G S, Lu Q, Shen H, Li S, You S, and Zhang L 2020 Land-cover classification with high-resolution remote sensing images using transferable deep models *Remote Sensing of Environment* vol. 237 p 111322

[5] Wang J, Ma A, Zhong Y, Zheng Z, and Zhang L 2022 Cross-sensor domain adaptation for high spatial resolution urban land-cover mapping: From airborne to spaceborne imagery *Remote Sensing of Environment* vol 277 p 113058

[6] He K, Zhang X, Ren S, and Sun J 2016 Deep residual learning for image recognition *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* vol 237 pp 770–778

[7] Xie S, Girshick R, Dollar P, Tu Z, and He K 2017 Aggregated residual transformations for deep neural networks *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*

[8] Zhang H, Wu C, Zhang Z, Zhu Y, Lin H, Zhang Z, Sun Y, and He T 2022 Resnest: Split-attention networks *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*

[9] Gao S -H, Cheng M -M, Zhao K, Zhang X -Y, Yang M -H, and Torr P 2021 Res2net: A new multi-scale backbone architecture *IEEE Transactions on Pattern Analysis and machine Intelligence* vol 43 no 2 pp 652–662

[10] Wittich D and Rottensteiner F 2021 Appearance based deep domain adaptation for the classification of aerial images *ISPRS Journal of Photogrammetry and Remote Sensing* vol 180 pp 82–102

[11] Wang J, Zheng Z, Ma A, Lu X, and Zhong Y 2021 LoveDA: A remote sensing land-cover dataset for domain adaptive semantic segmentation *Proc. of the Neural Information Processing Systems Track on Datasets and Benchmarks* vol 1

[12] Iqbal J and Ali M 2020 Weakly-supervised domain adaptation for built-up region segmentation in aerial and satellite imagery *ISPRS Journal of Photogrammetry and Remote Sensing* vol 167 pp 263–275

[13] Lu X, Zhong Y, Zheng Z, and Wang J 2021 Cross-domain road detection based on global-local adversarial learning framework from very high resolution satellite imagery *ISPRS Journal of Photogrammetry and Remote Sensing* vol 180 pp 296–312