Progress report I: Land Cover Change Detection using Neural Network for Satellite Images

Ashkan Bozorgzad (ab5243), Hari Prasad Renganathan (hr2514), Karveandhan Palanisamy (kp2941), Masataka Koga (mk4528), Yewen Zhou (yz4175), Yuki Ikeda (yi2220)

10/22/2022

Problem Definition and Progress Overview

This project is sponsored by JPMorgan Chase, an investment banking company. The goal of the project is to create high-resolution (1m / pixel) land cover change maps of a study area, the state of Maryland, USA, given multi-resolution imagery and label data. This project aims to provide an example of situations commonly found worldwide. In the field of earth observation, new images produce faster than high-quality, high-resolution labels. However, old and low-resolution labels are available, for example, 30m National Land Cover Database (NLCD) in the United States or 500 m MODIS land cover available worldwide. Therefore, it is significant to investigate how machine learning can be used to build a model that predicts high-resolution change without having a lot of higher-resolution change data. According to past studies [1], [2], and [3], weakly supervised segmentation and automatic super-resolution labeling are possible. These studies constructed high-resolution label predictors for high-resolution input imagery using regional supervision. They labeled a large block of land with four target classes (Water, Tree Canopy, Low Vegetation, and Impervious).

After creating a high-resolution model to predict land cover, predicting land cover change is a straightforward step. The model can be applied to two different years (images), and the super-resolved label can be compared to estimate changes in the four classes. The evaluation metric is the average intersection over union (IoU) between the predicted and ground truth labels for eight classes (loss and gain of the four target classes, except for no change). We can calculate the IoU as follows:

$$\frac{\#\text{ pixels labeled c in the model's prediction } \textbf{and} \text{ in the ground truth}}{\#\text{ pixels labeled c in the model's prediction } \textbf{or} \text{ in the ground truth}}.$$

In this report, we first describe the data and the type of changes that are being scored. Then we discuss the base model we used for this project. Finally, we discuss possible approaches and models we can use to address the problem of this project.

Dataset and Data Visualization

All data can be downloaded from the addresses listed at Data Link. The input images comprise nine layers covering the state of Maryland in the United States (~35.000 Km^2). All layers are upsampled from their native resolutions to 1m / pixel and provided as 2250 aligned tiles of dimensions not exceeding 4000 × 4000. The layers are as follows [4]:

1. NAIP (2 layers): 1m-resolution 4-band (red, green, blue, and near-infrared) aerial imagery from the US Department of Agriculture’s National Agriculture Imagery Program (NAIP) from two points in time: 2013 and 2017.
2. Landsat (5 layers): 30m-resolution 9-band image from the Landsat-8 satellite from five-time points: 2013, 2014, 2015, 2016, and 2017. Each of these images is a median composite from all cloud and cloud-shadow-masked surface-reflectance scenes intersecting Maryland.
3. NLCD (2 layers): 30m-resolution coarse land cover labels from the US Geological Survey’s National Land Cover Database in 15 classes (see Table 1) from two times: 2013 and 2016 (for labeling 2017 images). These labels were created in a semi-automatic way, with Landsat imagery as the principal input.

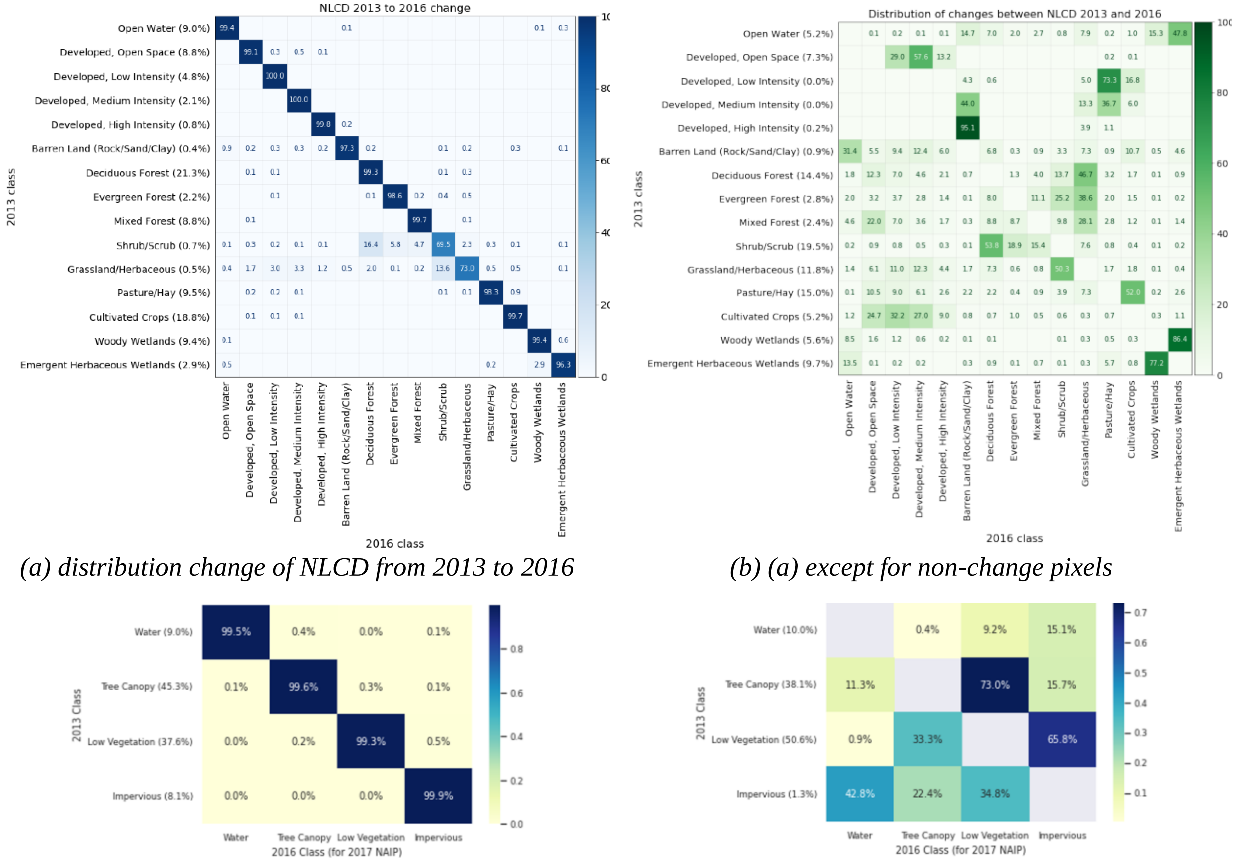
Table 1 shows the 15 classes the NLCD labels have and the four target class names we assigned to them according to [1].

Table 1 Correspondence between NLCD classes and the four target classes (cited from [4]) and assigned classes.

NLCD class name		Label Color	Target class	Approximate class frequencies			
				W	TC	LV	I
11	Open Water	🟦	Water	98%	2%	0%	0%
21	Developed, Open Space	🟩	Low Vegetation	0%	39%	49%	12%
22	Developed, Low Intensity	🟥	Impervious	0%	31%	34%	35%
23	Developed, Medium Intensity	🟥	Impervious	1%	13%	22%	64%
24	Developed, High Intensity	🟥	Impervious	0%	3%	7%	90%
31	Barren Land (Rock/Sand/Clay)	🟥	Impervious	5%	13%	43%	40%
41	Deciduous Forest	🟩	Tree Canopy	0%	93%	5%	0%
42	Evergreen Forest	🟩	Tree Canopy	0%	95%	4%	0%
43	Mixed Forest	🟩	Tree Canopy	0%	92%	7%	0%
52	Shrub/Scrub	🟩	Tree Canopy	0%	58%	38%	4%
71	Grassland/ Herbaceous	🟩	Low Vegetation	1%	23%	54%	22%
81	Pasture/Hay	🟩	Low Vegetation	0%	12%	83%	3%
82	Cultivated Crops	🟩	Low Vegetation	0%	5%	92%	1%
90	Woody Wetlands	🟩	Tree Canopy	0%	94%	5%	0%
95	Emergent Herbaceous Wetlands	🟩	Tree Canopy	8%	86%	5%	0%

In graphs (a) and (b) in Figure 1 cited from [1], the distribution of the 15 classes in the NLCD labels in 2013 (left axis) and its change between 2013 and 2016 can be seen. The numbers of elements in the matrices denote the distribution of 2016 classes for pixels of each 2013 class in percentage. (b) shows the distributions for the pixels limited to those where their labels changed between 2013 and 2016. From graph (a), we can see that Deciduous Forest (21.3%) and Cultivated Crops (18.8%) accounted for the broadest part of Maryland in 2013, while Developed, High Intensity (0.8%), Shrub/Scrub (0.7%), Grassland/Herbaceous (0.5%), and Barren Land (0.4%) the minor part. Moreover, 30.5% of Shrub/Scrub and 27.0% of Grassland/Herbaceous in 2013 became other classes by 2017. In graph (b), 60% of changes was composed of Shrub/Scrub (19.5%), Pasture/Hay (15.0%), Deciduous Forest (14.4%), and Grassland/Herbaceous (11.8%) in 2013. Also, some changes, such as Developed, High Intensity to Barren Land (95.1%) and Woody Wetlands to Emergent Herbaceous Wetlands (86.4%), seem more common than others.

The distribution changes for the four target classes can be seen in graphs (c) and (d). The distribution in 2013 (left axis) in (c) shows that most part in Maryland is Tree Canopy (45.3%) or Low Vegetation (37.6%), and there was little change in land cover during the time because more than 99% of each class remained the same category. As for graph (d), the change of labels from Impervious (1.3%) in 2013 appears the least. Losses of Tree Canopy into Low Vegetation (73.0%) and those of Low Vegetation into Impervious (50.6%) are distinctive points.



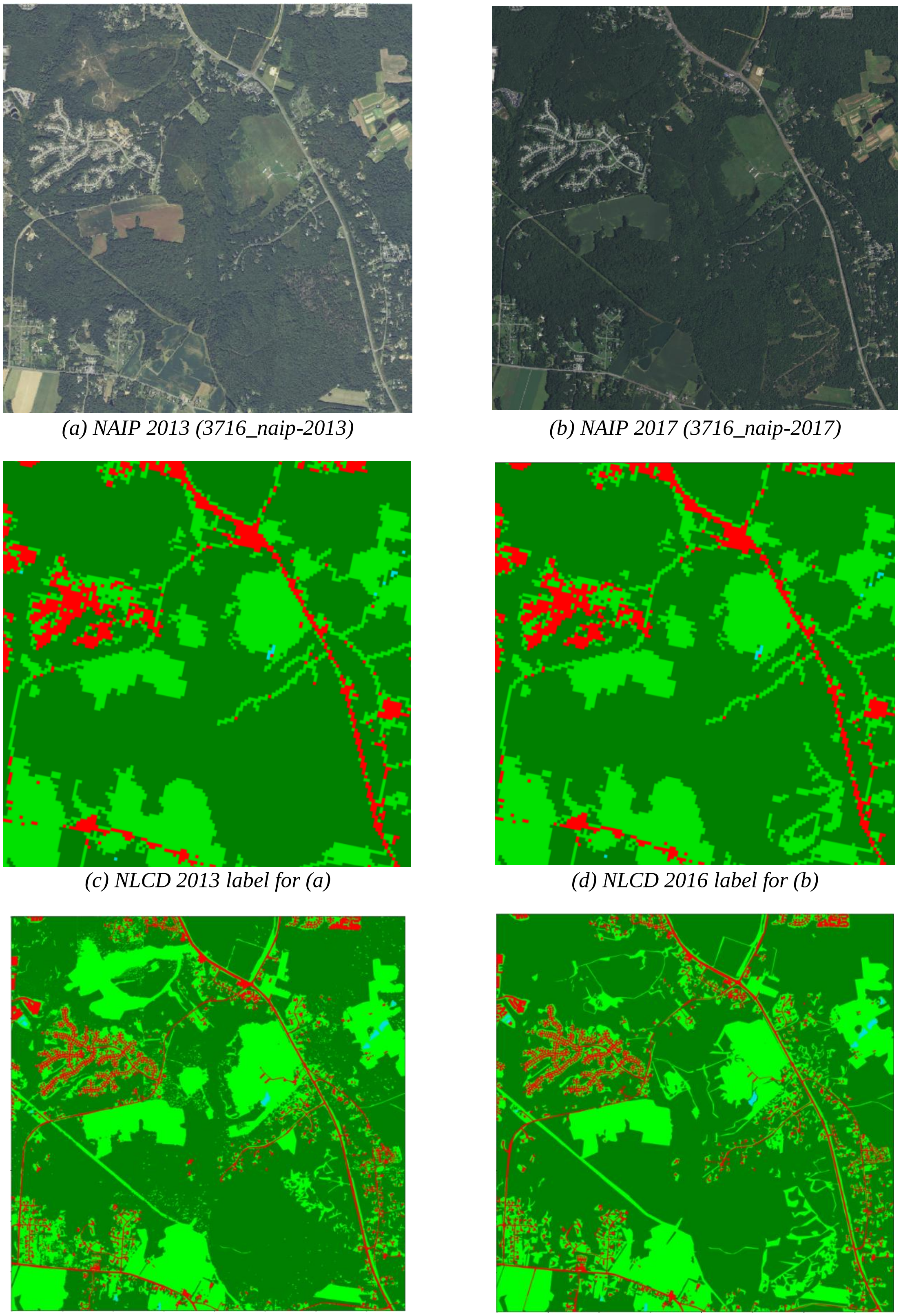
(a) distribution change of NLCD from 2013 to 2016

(b) (a) except for non-change pixels

(c) (a) for the four target classes

(d) (b) for the four target classes

Figure 1 Distribution of changes between NLCD 2013 and 2016
((a) and (b) were cited from [1], and (c) and (d) were calculated by us using the numbers in (a) and (b))



(a) NAIP 2013 (3716_naip-2013)

(b) NAIP 2017 (3716_naip-2017)

(c) NLCD 2013 label for (a)

(d) NLCD 2016 label for (b)

(e) high-resolution label for (a)

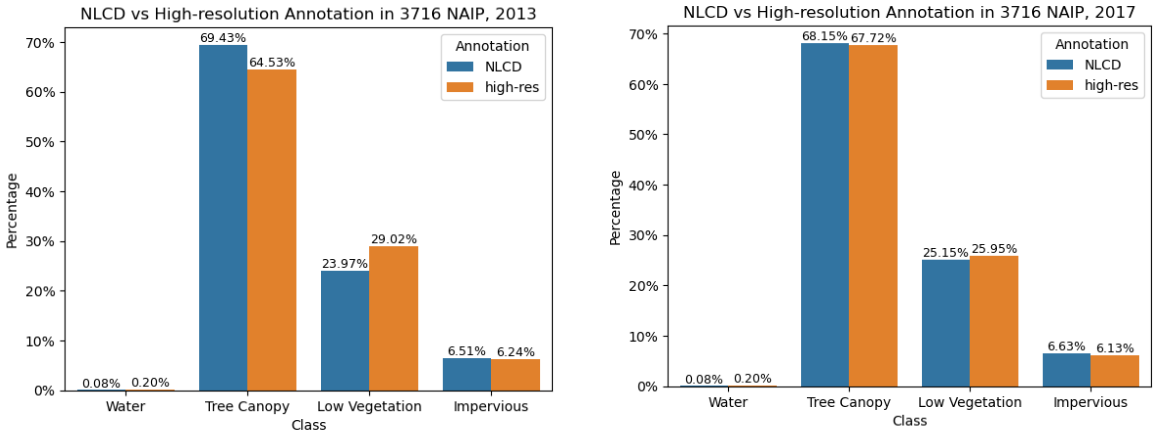
(f) high-resolution label for (b)

Figure 2 Sample data and their labels

Examples of the input NAIP images (3716_naip-2013 and 3716_naip-2017), their NLCD labels, and their high-resolution labels we manually annotated are shown in Figure 2. As for 3716_naip-2013, 3716 denotes the ID for the place, and 2013 shows the year when the image was shot. The most conspicuous change that happened in the period in this area would be the gain of Tree Canopy in the upper left and its loss in the bottom right. Unfortunately, NLCD labels do not capture the former difference. Also, since they are too coarse, especially concerning Low Vegetation and Impervious, the road of Low Vegetation in the middle left side is not recognized.

The training data for a year consist of 2250 NAIP aligned tiles (1m /pixel, high resolution) and NLCD (30m/pixel, low resolution). For the test data, 50 NAIP aligned tiles (1m / pixel, high resolution) selected from the 2250 training tiles are available, but their high-resolution land cover label is not available. We classified all pixels of 3716_naip-2013 and 3716_naip-2017 into the four target classes to create high-resolution labeling for the images to evaluate models using manual and semi-automatic tools in GroundWork. It took more than 24 hours to complete labeling only for the two NAIP images. Nevertheless, some parts of the high-resolution manual labeling seemed not to be very accurate. Therefore, we will modify them and add the labels based on the analysis of what kind of annotation rules should be followed.

For previous images, 3716_naip-2013 and 3716_naip-2017, we computed the area percentage of each label category and showed the differences between NLCD and high-resolution label images in Figure 3. Most categories for both labels have almost equal percentages. However, since high-resolution labeling is more accurate, some gaps appear. The most significant differences are the Tree Canopy and Low Vegetation percentages in 2013. These must be caused by the area and road of Low Vegetation mistaken for Tree Canopy in NLCD labels, as shown in the upper left and the middle left side in (a), (c), and (e) of Figure 2.



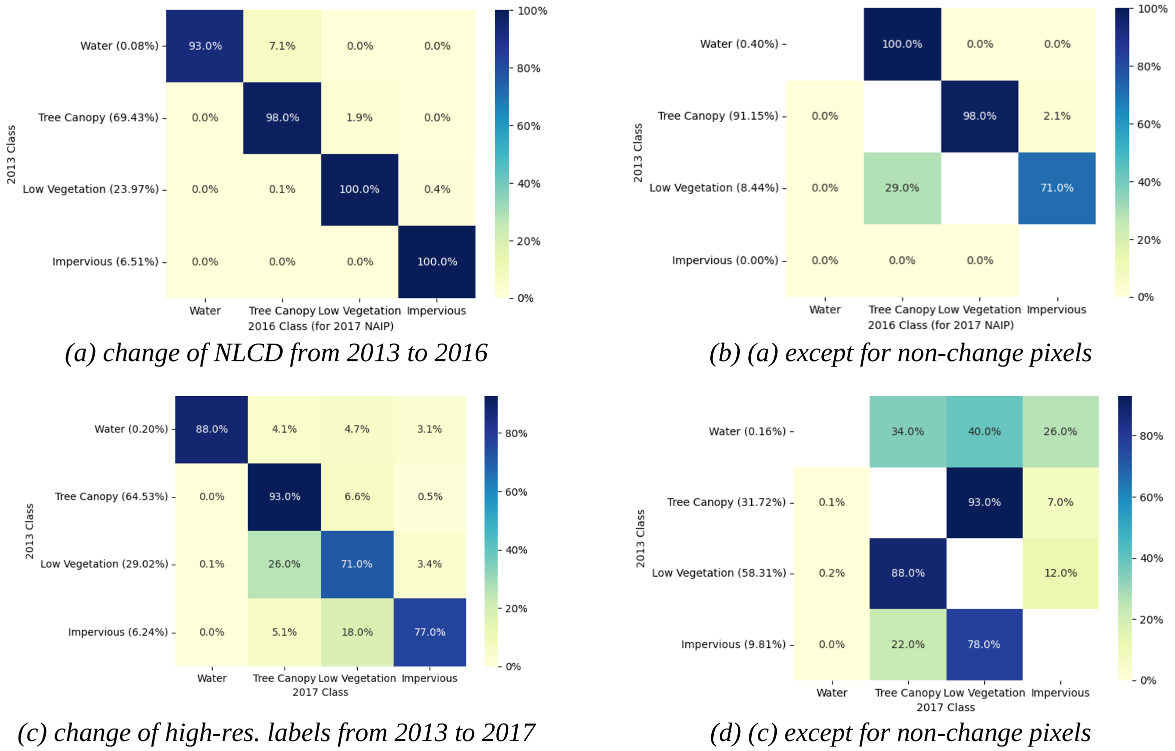
(a) comparison in 2013

(b) comparison in 2017

Figure 3 Area comparison between low-resolution (NLCD) and high-resolution labels

Figure 4 shows distribution changes of NLCD and high-resolution labels for 3716_naip-2013 and 3716_naip-2017. From graphs (a) and (c), we can see that more Low Vegetation (29.0%) and Impervious (23.0%) were lost between 2013 and 2017 than the volume NLCD implied.

Then, the most significant loss in 2013 was not Tree Canopy (91.15%) as (b) indicates, but Low Vegetation (58.31%) shown in (d). Most losses in Low Vegetation led to gains in Tree Canopy (88.0%). We can guess that the unrecognized area and road of Low Vegetation in NLCD in 2013 mentioned above generated these gaps. Furthermore, the change volume of Impervious was not 0.00% but 9.81% and was removed or covered by Low Vegetation mainly (78.0%) in 2017. Water was filled by Tree Canopy, Low Vegetation, and Impervious in nearly equal proportion. These differences would be a pile of minor ones due to the roughness of NLCD labels.



(a) change of NLCD from 2013 to 2016

(b) (a) except for non-change pixels

(c) change of high-res. labels from 2013 to 2017

(d) (c) except for non-change pixels

Figure 4 Comparison of distribution change of NLCD and high-res. labels for 3716_naip-2013 and 3716_naip-2017

Methodology

NLCD Difference Baseline

We studied the NLCD difference algorithm as our first baseline algorithm. It takes only input from the 2013 and 2016 NLCD layers. They were used for each pixel, and high-resolution classes were assigned according to the target class of Table 1. Table 2 shows the IoU scores in the eight categories of gains and losses and the average IoU resulting from the NLCD difference algorithm. They were calculated by regarding two high-resolution labels for NAIP images in 2013 and 2017 in a particular area of Maryland (3716_naip-2013 and 3716_naip-2017) as ground truth labels. The average IoU score is 0.025, so it seems too poor. However, the average IoU score (0.139) and other scores of the NLCD difference algorithm reported by [1] using all high-resolution test labels that they created, “NLCD diff for all test labels [1]” in Table 2, are much higher than our scores. Their model’s predictions and ours must be the same because the

NLCD difference algorithm does not require training and we used the same inputs, NLCD labels, to estimate classes for each pixel. Since the test labels were unavailable and making all high-resolution labels took too much time, we evaluated using the two high-resolution labels. These differences in IoU scores between [1]'s and ours mean that the two labels we selected were furthermore demanding than the others, and we need to estimate the models' performances more than the calculated value. Also, in the following work, it would be better that we avoid overfitting models to these high-resolution labels and consider re-selecting images with adequate difficulty levels to test.

Table 2 IoU scores for our baseline models. −C and +C denote the loss and gain of class C, respectively.

Algorithm	-W	-TC	-LV	-I	+W	+TC	+LV	+I	Avg.
NLCD diff	0.000	0.106	0.002	0.000	0.000	0.000	0.092	0.003	0.025
1-layer FCN	0.022	0.063	0.080	0.032	0.010	0.085	0.064	0.034	0.049
(NLCD diff for all test labels [1])	0.148	0.167	0.282	0.014	0.031	0.001	0.106	0.362	0.139

One Layer FCN Baseline (Multiclass Logistic Regression)

We constructed a fully convolutional network (FCN) with a single convolutional layer as our second baseline model and set its architecture as follows: the number of input channels = 4 for (Red, Green, Blue, and Near-Infrared), the number of output classes = 5 (Water, Tree Canopy, Low Vegetation, Impervious, and None), filter size = 3, stride = 1, and padding = 1. The None output class is for the case there are pixels we cannot classify into the four target classes, and we do not evaluate its IoU scores. We added padding to the images to make sure the model trained on the same size images before training because the size of images is different in some images. Since we used cross-entropy as the loss function, the model architecture is the same as the multiclass logistic regression with 36 features (= 4 pixel colors * 9 surrounding pixels) to estimate the label of each pixel. After randomly choosing 128 pairs of NAIP images and NLCD labels out of 4500 pairs either in 2013 or 2017, we trained the model for 20 epochs in a batch size = 4 utilizing Adam with an initial learning rate = 0.001. Due to the total dataset size for this project, around 254 GB, and our computational resource limitations, we trained our model only on the part of the images. We predicted labels for one pair of our test images (3716_naip-2013 and 3716_naip-2017) and compared them with the high-resolution labels produced by GroudWork. The IoU scores of this model are written in Table 2. The average IoU is two times better than our NLCD difference algorithm. Especially, it can estimate gains and losses of Water (0.022 and 0.010) and Impervious (0.032 and 0.034), loss of Low Vegetation (0.080), and gain of Tree Canopy (0.085) more accurately, whereas the others less accurately.

As can be seen in Figure 5, the model is poor at predicting most categories other than the Impervious category. This might be because of the simplicity of the model. Adding more layers to the model can create more capacity to learn and improve accuracy. In addition, the training and test images for this model were randomly selected. The more variable and diverse images would lead to better models. On the other hand, its predictions about the Impervious labels and the Low Vegetation labels expressing the road in the middle left side are superior to NLCD labels. Therefore, ensemble learning with the shallow network only for relatively thin areas and paths can be one candidate for the following models. Also, training two models separately each for one year using only images and labels in the year may work better because the prediction for 2013 appears better and because the training images are possibly biased in terms of their years.

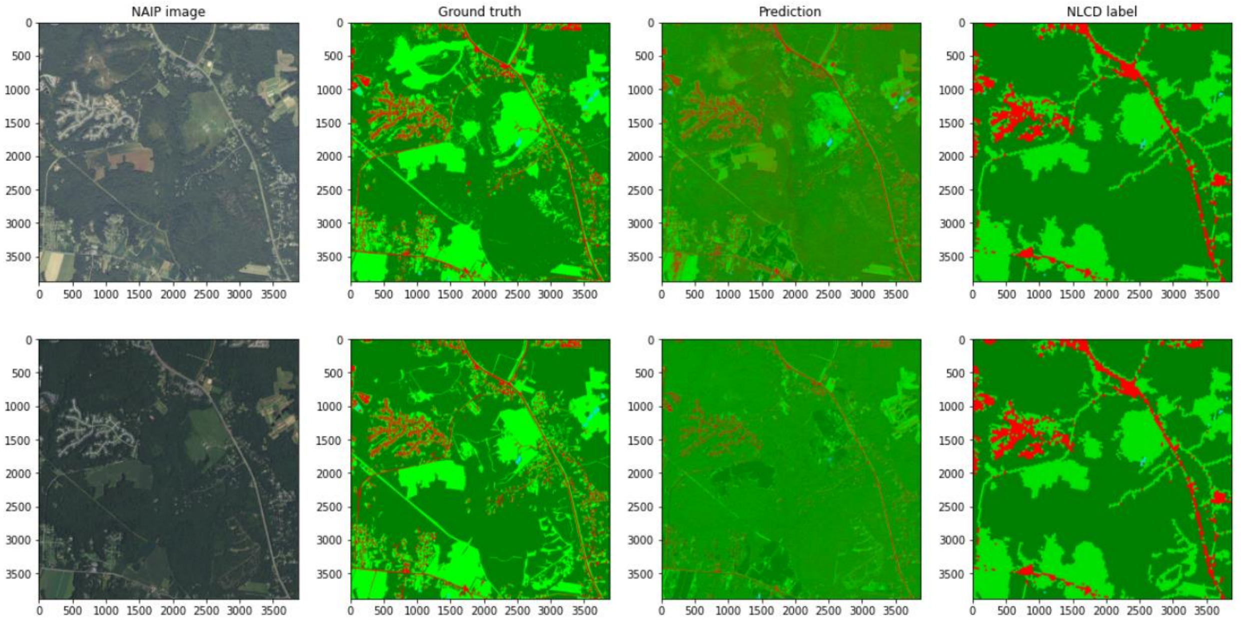
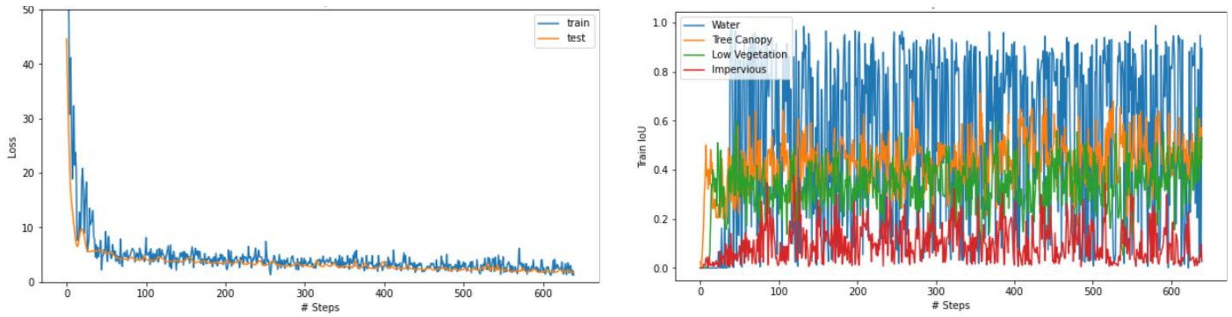


Figure 5 Comparison between model prediction and NAIP images, ground truth labels, and NLCD labels

Figure 6 shows the training and test losses and the training IoU scores for the target four classes on each step, not for the gains and losses of the four classes. Even after the training losses seem to converge on around 50 steps, the IoU scores fluctuate significantly. One reason for this must be the shortage of pixels to evaluate each class, especially Water. Therefore, we need to obtain more test labels to assess models accurately and to ascertain the primary factor to improve them. Also, in addition to trying to add layers to capture the architecture of the true predictor, experimenting with more patterns of the combination of optimization methods, initial learning rates, and model architectures perhaps ameliorate this volatility, observing and comparing the convergence of loss and IoU scores.



(a) the training and test losses

(b) the training IoU scores for the target four classes

Figure 6 The training loss and training IoU scores of the 1-layer FCN baseline model on each step

Our Current Candidate Approach

According to past experiences and studies [5], [6], and [7], the potential models that can address high-resolution labeling include the following model architectures:

1. Convolutional neural networks: Convolutional neural networks have several benefits that make them useful for many different applications. CNNs don't require manual feature engineering: they can grasp relevant features during training. CNN, due to the procedure of convolution, is much more computationally efficient than regular neural networks. They usually show higher accuracy than non-convolutional NNs, especially when there is a lot of data involved. Convolutional neural networks are often used for image classification. By recognizing valuable features, CNN can identify different objects in images. CNN can be also used in agriculture. The networks receive images from satellites like LSAT and can use this information to classify lands based on their level of cultivation. Consequently, this data can be used for making predictions about the fertility level of the grounds or developing a strategy for the optimal use of farmland. In our project, we can use a fully convolutional network (FCN), a CNN model without Dense layers, instead of traditional CNN architecture for image classifications. FCN is suitable for the semantic segmentation task of classifying the object class for each pixel within an image, unlike traditional CNN models. We already tried an FCN with a single layer, but the adequately deep architectures would catch the more intricate relationship between features and target labels. Also, other hyper-parameters that we have not explored yet, including kernel size and optimization methods, would affect IoU scores significantly. There would be two types of basic FCN models we can try first:
    a. Two separate for two years (FCN / tile): We will train different FCNs for each of the validation tiles (One for NAIP 2013 and NLCD 2013 and the other for NAIP 2017 and NLCD 2016). Then the high-resolution labeling of NAIP 2013 and NAIP 2017 imagery will be predicted independently. By comparing the labeling of two different years, the landcover change can be calculated.
    b. One on both years (FCN / all): We will train one FCN model for the entire dataset. This may give an advantage due to the much larger training data offering more potential for generalization. At the same time, the model may require more computational resources to fit all of the relationships that appeared in only 2013, 2017, and both years using deep architectures. The model can independently predict the NAIP 2013 and NAIP 2017 images. Landcover changes can be calculated by comparing the high-resolution label images of two different years in the same location.
2. U-Net (all dataset): This method is similar to FCN / all. However, we will use a U-Net family architecture. U-Net is an FCN for image semantic segmentation and consists of encoder and decoder parts connected with skip connections. The encoder extract features of different spatial resolution, and the decoder uses them to define segmentation masks. We will first use a ResNet-18 encoder structure with the first three blocks for the downsampling path and convolutions with 128, 64, and 64 filters for the corresponding upsampling layers, like the model used in the past study [1].

Goals and Next Steps

Currently, we have the training data and the base model and know our potential models for this project. For the next step, we need to implement the potential models and see if we can improve them. Then, we will evaluate it using test labeling data to find the room to develop. At the same time, labeling NAIP images for more high-resolution test datasets will be conducted. Finally, search to see if there are any other methods to reinforce the model's weak points, such as the combination of strategies the past studies listed in the References section adopted. Besides these fundamental steps to improve models, we will try to employ other datasets, such as Dynamic World V1 labels, to refine our models and arrange and upgrade our Python scripts to deliver to JP Morgan at the end of this project. Dynamic Worl V1 label is a 10m near-real-time Land Cover dataset in Google Earth Engine Data Catalog with class probabilities and label information for nine classes.

Contribution

- Ashkan Bozorgzad: Made baseline models and the progress report, cleaned and preprocessed data, and created high-resolution labels for the test data
- Hari Prasad Renganathan: EDA. Cleaned and preprocessed data and created high-resolution labels for the test data
- Karveandhan Palanisamy: EDA. Cleaned and preprocessed data and created high-resolution labels for the test data
- Masataka Koga: Team Captain: Set up logistics, planned schedule, and managed progress. EDA. Made baseline models and the progress report, refactorized codes, cleaned and preprocessed data, and created high-resolution labels for the test data
- Yewen Zhou: EDA (main contributor). Cleaned and preprocessed data and created high-resolution labels for the test data
- Yuki Ikeda: Created a tutorial video of GroudWork, made baseline models (main contributor), cleaned and preprocessed data, and created high-resolution labels for the test data

References

[1] K. Malkin, C. Robinson and N. Jojic, "High-resolution land cover change from low-resolution labels: Simple baselines for the 2021 IEEE GRSS Data Fusion Contest," arXiv preprint arXiv:2101.01154, 2021.

[2] Z. Li, F. Lu, H. Zhang, G. Yang and L. Zhang, "Change cross-detection based on label improvements and multi-model fusion for multi-temporal remote sensing images," *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, pp. 2054-2057, 2021.

[3] L. Tu, J. Li and X. Huang, "High-resolution land cover change detection using low-resolution labels via a semi-supervised deep learning approach - 2021 IEEE data fusion contest track MSD," *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, pp. 2058-2061, 2021.

[4] "2021 IEEE GRSS Data Fusion Contest: Track MSD," [Online]. Available: https://www.grss-ieee.org/community/technical-committees/2021-ieee-grss-data-fusion-contest-track-msd/.

[5] Q. Bao, Y. Liu, Z. Zhang, D. Chen, Y. Yang, L. Jiao and F. Liu, "MRTA: Multi-resolution training algorithm for multitemporal semantic change detection," *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, pp. 2062-2065, 2021.

[6] Z. Li, F. Lu, H. Zhang, L. Tu, J. Li, X. Huang, C. Robinson, N. Malkin, N. Jojic, P. Ghamisi, R. Hänsch and N. Yokoya, "The outcome of the 2021 IEEE GRSS Data Fusion Contest-Track MSD: multitemporal semantic change detection," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 15, pp. 1643-1655, 2022.

[7] Z. Zheng, Y. Liu, S. Tian, J. Wang, A. Ma and Y. Zhong, "Weakly supervised semantic change detection via label refinement framework," *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, pp. 2066-2069, 2021.

Appendices

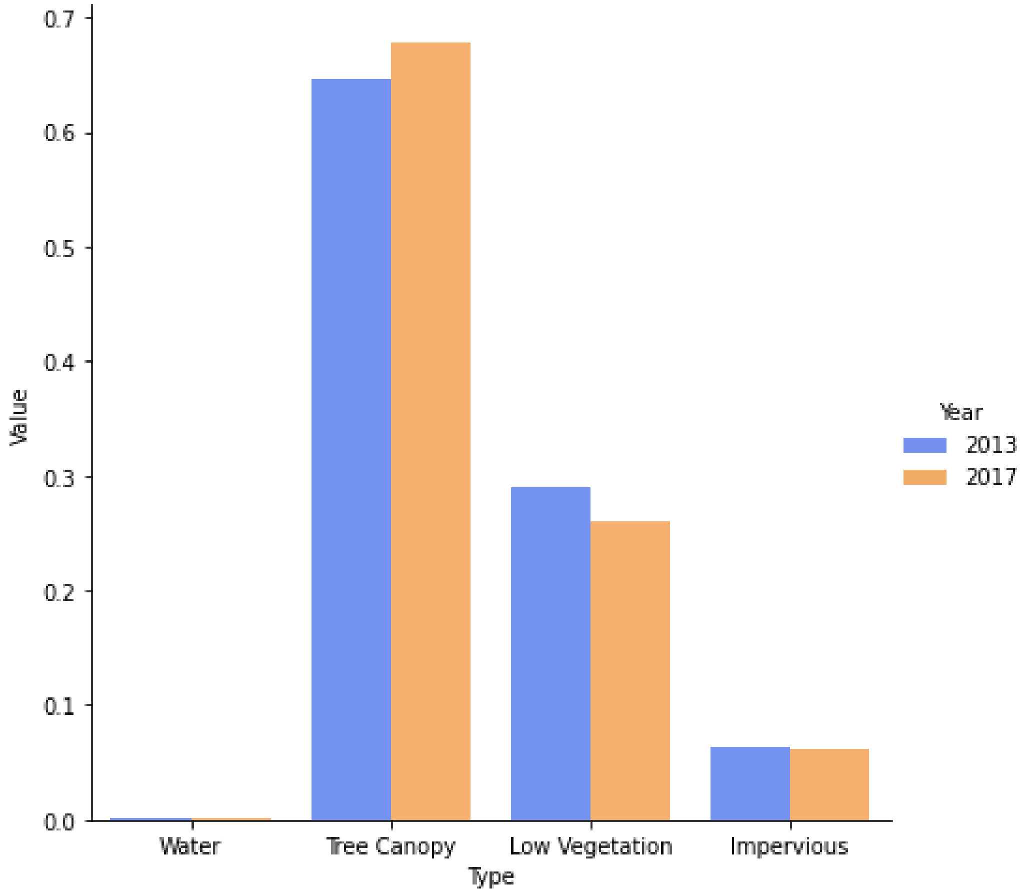


Figure 7 Area comparison in high-resolution labels between 2013 and 2017

Figure 7 shows the area comparison in the high-resolution labels in 2013 and 2017 in a particular zone, Maryland state (3716_naip-2013 and 3716_naip-2017). We can see an increase in the proportion of Tree Canopy and a decrease in Low Vegetation. The proportion of Water has been little and almost remains the same.

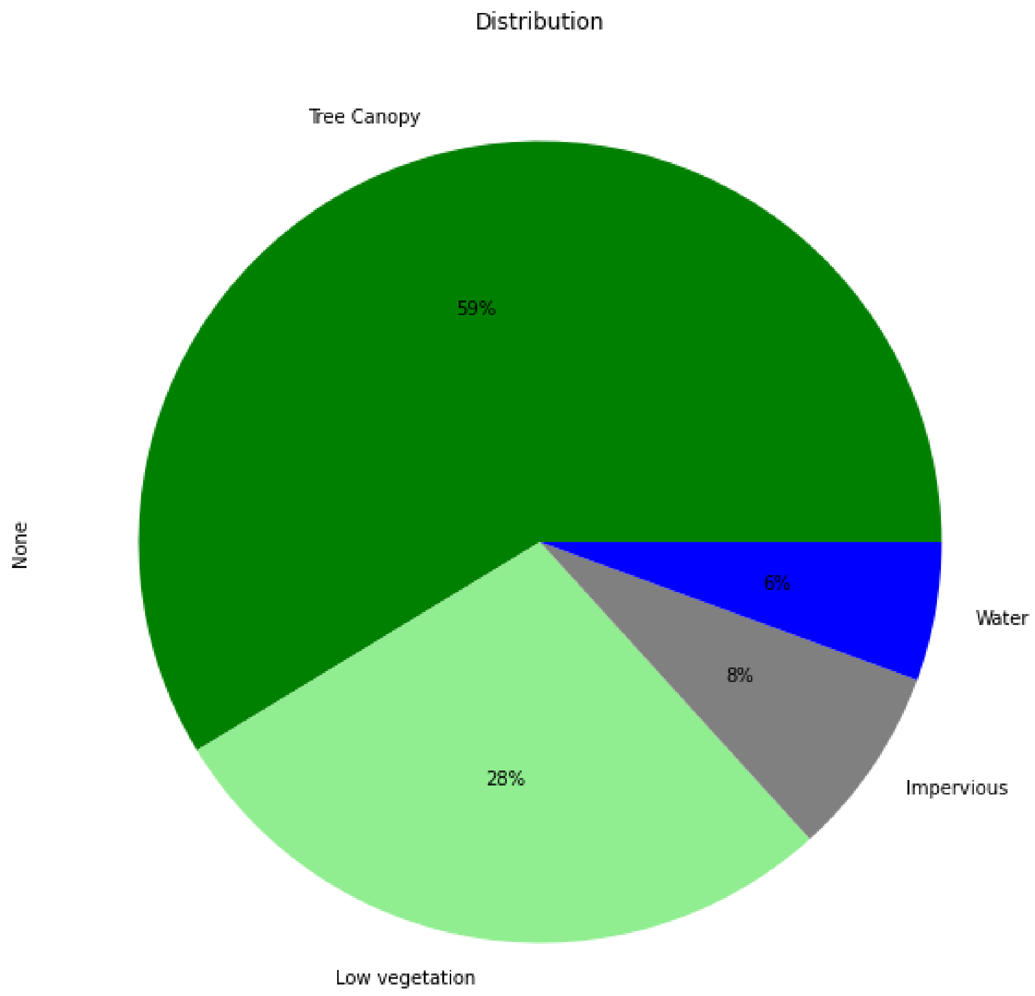


Figure 8 The distribution of the four target classes in an area other than 3716 NAIP

Figure 8 shows the distribution of the four target labels in an area of Maryland different from 3716 NAIP. This area includes the four target classes in a ratio similar to the average in (c), Figure 1 (in 2013, Water: 9.0%, Tree Canopy: 45.3%, Low Vegetation: 37.6%, and Impervious: 8.1%), in contrast to 3716 NAIP. Therefore, we can employ this area to tune the general purpose model and the other with distinctive distribution, like the one for 3716 NAIP, to sophisticate the model that specializes in discerning a specific type of area.