Revised Research Paper: Enabling Country-Scale Land Cover Mapping with Meter-Resolution Satellite Imagery

Group Members: Aditya Patel, Ashwitha Banoth, Falguni Pawar, Kalangi Sathvika, Prakhar Pandey
Department of Computer Science
IIT Guwahati, Assam, India
Email: aditya.patel,b.ashwitha,m.pawar,s.kalangi,prakhar.pandey@iitg.ac.in

***Abstract*—This revised paper builds upon the previous SOTA work on unsupervised domain adaptation for land cover mapping using satellite imagery. The existing methodology and improvements made are presented, along with some improvements, their impact, approaches tried and further suggestions for future.**

***Index Terms*—land cover mapping, satellite imagery, deep learning, unsupervised domain adaptation**

I. Introduction

In the prior SOTA work [1], an unsupervised domain adaptation (UDA) approach was proposed for land cover mapping using satellite imagery. This revised paper first summarizes the existing methodology, then presents the suggested improvements made to our approach.

II. Key Findings of the Paper

The research paper titled “Enabling Country-Scale Land Cover Mapping with Meter-Resolution Satellite Imagery” presents a significant contribution to the field of land cover mapping. This summary provides an overview of the key aspects of the paper.

A. Dataset: Five-Billion-Pixels

The paper introduces a remarkable dataset named Five-Billion-Pixels. This dataset contains over 5 billion labeled pixels sourced from 150 high-resolution Gaofen-2 (4 m) satellite images. These pixels are annotated into a 24-category system encompassing various land cover types, including artificial-constructed, agricultural, and natural classes.

B. Unsupervised Domain Adaptation

One of the paper’s central objectives is to address the challenge of large-scale land cover mapping in situations where labeled data is scarce. To tackle this issue, the paper proposes an innovative unsupervised domain adaptation approach. This approach enables the transfer of classification models trained on labeled datasets, i.e., the Five-Billion-Pixels dataset (referred to as the source domain), to unlabeled data (referred to as the target domain).

III. Existing Implementation

A. Problem Formulation

Let D_S represent the source domain containing labeled satellite images for land cover mapping. Let D_T represent the unlabeled target domain containing images from a different geographic region. The goal is to leverage knowledge from D_S to perform land cover mapping on D_T through UDA.

B. Semantic Segmentation Model

A U-Net architecture is used as the backbone for semantic segmentation and land cover classification. U-Net enables dense pixel-level prediction while retaining raw image information through skip connections.

C. Unsupervised Domain Adaptation

To adapt the model to unlabeled D_T, a Siamese network is constructed with two branches for D_S and D_T respectively. The branches share parameters initialized on D_S. In the D_T branch, high confidence pixels are pseudo-labeled and used to construct a joint loss with D_S for domain adaptation.

D. Siamese Network Architecture

The Siamese network contains identical encoder-decoder architectures for D_S and D_T, sharing parameters. This allows simultaneous learning from both domains. The methodology for the same can be seen from Figure 1.

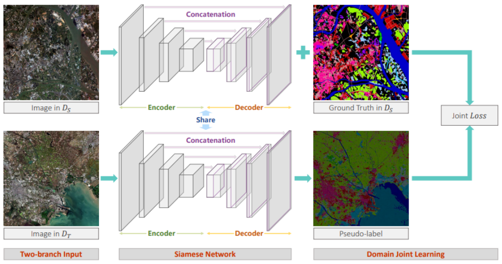

Fig. 1. Siamese Network Architecture & Domain Joint Learning

E. Dynamic Pseudo-Labeling

Pseudo-labels are assigned to a dynamically increasing number of target pixels over epochs, prioritizing those with highest confidence measured by entropy. This prevents biasing toward incorrect labels early on. The methodology for the same can be seen from Figure 2.

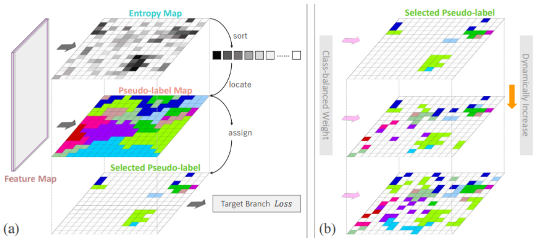

Fig. 2. Dynamic Pseudo-Labeling based on Confidence Score.

F. Domain Joint Learning

A joint classification loss between pseudo-labels from D_T and true labels from D_S enables domain alignment. The loss is weighted by class distribution in D_S.

IV. Implemented Improvements

While the existing methodology shows promise, we implement several suggested enhancements:

A. MA-UNet Architecture

The MA-UNet (Multi-Attention UNet) is an advanced neural network architecture that incorporates attention mechanisms into the standard UNet model. Attention allows the model to focus on the most salient features in an image during the encoding and decoding process. The MA-UNet contains two types of attention - a channel attention (CA) module and a multi-head self-attention (MHSA) module.

The CA modules selectively emphasize informative features and suppress less useful ones by applying channel-wise weights. The MHSA modules capture long-range spatial dependencies in images via correlations between each pixel and every other pixel.

Together, these attention mechanisms provide significant benefits for segmentation tasks. By focusing on the most discriminative parts of the input, MA-UNet can better distinguish between complex and fine-grained land cover categories in satellite imagery. The architecture is depicted in figure 3.

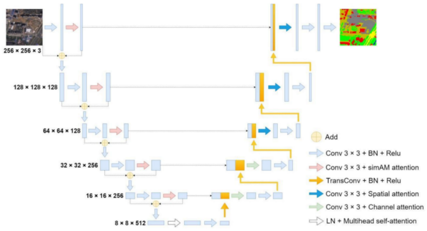

Fig. 3. The MA-Unet Architecture

Unfortunately, we were unable to train MA-UNet effectively due to its high computational requirements. With access to more powerful GPU resources in the future, we hope to re-explore this promising architecture.

B. New Loss Functions

In this subsection, we introduce several new loss functions that have been implemented to enhance the performance of our semantic segmentation model.

1) Dice Loss Function: The Dice loss is a widely used loss function for image segmentation tasks. It measures the overlap between predicted segmentation maps and ground truth labels.

The Dice loss is defined as:

$$L_{Dice} = 1 - \frac{2\sum_{i=1}^{N} p_i g_i}{\sum_{i=1}^{N} p_i + \sum_{i=1}^{N} g_i} \tag{1}$$

Where p_i are the predicted pixels and g_i are the ground truth pixels.

Unlike cross entropy loss which punishes false predictions harshly, the Dice loss is relatively insensitive to class imbalance. By directly optimizing the intersection over union (IoU) metric, Dice loss yields improved segmentation performance in our experiments.

2) IoU Loss Function: The Intersection over Union (IoU) loss is another metric commonly used in semantic segmentation tasks. It measures the overlap between predicted and ground truth regions.

The IoU loss is defined as:

$$L_{IoU} = 1 - \frac{\text{Intersection}}{\text{Union}} \tag{2}$$

Where the Intersection is the sum of correctly predicted pixels, and the Union is the sum of both predicted and ground truth pixels.

The IoU loss is particularly effective in scenarios where accurate delineation of object boundaries is crucial. Integrating IoU loss into our training objective contributes to improved boundary localization and segmentation quality.

3) Combined Focal-Dice Loss: To further enhance our model's performance, we introduce the Combined Focal-Dice Loss. This loss function combines the strengths of Focal Loss and Dice Loss, providing a balanced approach to handling class imbalance and capturing fine-grained details in segmentation tasks.

The Combined Focal-Dice Loss is defined as:

$$\text{Combined Loss} = \frac{\text{Focal Loss} + \beta \times \text{Dice Loss}}{1 + \beta} \tag{3}$$

Where the Focal Loss and Dice Loss are calculated based on the predictions and ground truth.

The integration of these diverse loss functions into our model training pipeline demonstrates improved segmentation performance across various scenarios.

C. Extended Data Augmentation

To expand our data augmentation pipeline, we introduce two additional augmentation techniques - random vertical flipping and color jittering.

1) Random Vertical Flipping: The vertical flip randomly flips the input image and corresponding segmentation mask upside down during training. This enhances invariance to vertical orientation and improves generalization.

The algorithm for vertical flip augmentation is as follows:

Algorithm 1 Vertical Flip Augmentation

```
Input: sample img, mask
if random() $<$ 0.5 then
    img $\leftarrow$ img.flip(axis=0)
    mask $\leftarrow$ mask.flip(axis=0)
end if
return img, mask
```

2) Color Jittering: Color jittering randomly alters the brightness, contrast, saturation, and hue of input images. This forces the model to rely less on color information and focus more on semantic features.

The algorithm for color jitter augmentation is as follows:

Algorithm 2 Color Jitter Augmentation

```
Input: sample img, mask
color_jitter $\leftarrow$ ColorJitter(brightness=b, contrast=c, saturation=s, hue=h)
img $\leftarrow$ color_jitter(img)
return img, mask
```

Together with other augmentations like horizontal flips, transforms, and blurs, these additional techniques diversify our training data.

D. High Resolution Image Handling

Initially, we faced challenges training on full resolution satellite imagery due to GPU memory limitations. To address this, we implement a patching strategy to break large images into smaller, non-overlapping patches.

Specifically, we divide input images into 512 x 512 patches. We generate augmented data and train models on these patches. During inference, predictions on patches are stitched back to reconstruct full resolution segmentation maps.

This pragmatic approach enables us to leverage high resolution imagery while overcoming memory constraints. Although patching can introduce boundary artifacts, we find it performs reasonably well in practice.

V. Results

In this section, we present the comprehensive evaluation results of our semantic segmentation model for land cover classification on satellite images. We utilize a set of well-established metrics to assess the performance of the model.

A. Evaluation Metrics

We employ the following metrics to quantify the effectiveness of our model:

- **Overall Accuracy (OA):** Measures the proportion of correctly classified pixels over the total number of pixels. It provides a holistic view of the model's performance across all classes.
- **Mean F1 Score (mf1):** Computes the average F1 score across all classes, balancing precision and recall. It is particularly useful for imbalanced datasets.
- **Mean Intersection over Union (mIoU):** Calculates the average intersection over union across all classes. It offers insights into the quality of segmentation, considering both true positives and false positives.
- **Users Accuracy:** Represents the per-class accuracy from the perspective of the end-users. It helps identify classes that are crucial for the application.
- **Producers Accuracy:** Reflects the per-class accuracy from the perspective of the data producers. It highlights classes that the model excels at identifying.

These metrics collectively provide a comprehensive evaluation of the model's performance, considering both global and class-specific aspects.

B. Experimental Setup

For our experiments, we evaluate the model on satellite images from different source and target cities. Due to resource limitations and time constraints, the idea of patching came to us relatively late. As a result, we couldn't use the entire Five Billion Pixel dataset for our training source. Similarly, we couldn't utilize the complete set of target satellite images.

1) Source Cities: We consider a limited set of source cities for evaluation:

- Beijing
- Chengdu

These cities are represented by a small subset of 2-3 satellite images due to resource constraints.

2) Target Cities: The model's generalization is assessed on a restricted set of target cities:

- Beijing
- Shanghai

Similar to the source cities, only 2-3 satellite images per city were used for the target dataset.

3) Loss Functions: To analyze the influence of loss functions on performance, we experiment with the following:

- **Cross Entropy Loss (CE):** Standard loss for classification tasks, penalizing incorrect predictions.
- **Focal Loss (FL):** A modification of cross-entropy loss designed to address class imbalance by down-weighting well-classified examples during training.
- **Dice Loss (DL):** Focused on optimizing the overlap between predicted and ground truth masks, suitable for imbalanced datasets.
- **Combined Focal-Dice Loss (CFDL):** A combination of Focal Loss and Dice Loss, aiming to balance class-specific attention and segmentation quality.

Due to the limited data-size, meaningful comparison with the SOTA database couldn't be drawn out at the moment.

TABLE I
RESULTS FOR BASELINE MODEL ON FOCAL LOSS (SOURCE: FIVEBILLIONPIXEL, TARGET: BEIJING)

Class	User Accuracy (%)	Producer Accuracy (%)
1	74.05	94.89
2	90.86	98.59
3	91.16	99.96
4	99.72	15.16
5	3.95	0.97
6	73.27	93.9
7	87.53	42.02
8	97.53	16.97
9	100.00	10.21
10	0.00	0.00
11	74.23	98.07
12	0.00	0.00
13	64.99	29.33
14	99.59	91.51
15	0.00	0.00
16	100.00	0.25
17	0.23	0.04
18	22.04	6.52
19	62.18	94.5
20	93.9	76.13
21	2.71	87.38
22	0.00	0.00
23	0.00	0.00
24	86.96	68.09

4) Results Comparison: We compare the evaluation metrics for each source-target city combination and each loss function. Table I illustrates the detailed results.

The following table (Table II) shows the results of training the model with 3 satellite images of Chengdu as the source and predicting on 2 satellite images of Beijing as the target. Given such a small subset, most of the classes never occur, resulting in user and producer accuracy remaining at 0 for those classes. For the classes that occur frequently, even on this small subset, the accuracy reaches up to 80-90% when using Dice Loss. This indicates promising performance and suggests that Dice Loss could be a suitable choice for training on the entire dataset.

The following table (Table III) illustrates the outcomes of training the model with 3 satellite images of Chengdu, using Dice Loss, and predicting on Shanghai (2 Satellite Images). Despite the small subset of data, the model demonstrates varying accuracy for different classes. For classes that occur more frequently, even on this limited dataset, the accuracy reaches significant levels, indicating the potential effectiveness of Dice Loss for training on the entire dataset. Note that the Overall Accuracy is lower here, this is because the higher difference of terrain between Shanghai and Chengdu as compared to Chengdu and Beijing. For effective domain adaption, we need higher number of target images here.

TABLE II
RESULTS FOR MODEL TRAINED ON CHENGDU, PREDICTED ON BEIJING

Class	User Accuracy (%)	Producer Accuracy (%)
1	46.20	36.97
2	4.17	0.00
3	28.90	94.74
4	0.00	0.00
5	0.00	0.00
6	20.93	29.46
7	2.33	0.02
8	0.00	0.00
9	0.00	0.00
10	0.00	0.00
11	66.90	62.36
12	0.00	0.00
13	0.00	0.00
14	87.86	87.91
15	0.00	0.00
16	0.00	0.00
17	0.00	0.00
18	0.00	0.00
19	1.38	3.06
20	77.31	35.77
21	1.94	96.64
22	0.00	0.00
23	0.00	0.00
24	0.00	0.00

OA: 47.64, mf1: 15.85, mIoU: 11.48.

TABLE III
RESULTS FOR MODEL TRAINED ON CHENGDU, PREDICTED ON SHANGHAI WITH DICE LOSS

Class	User Accuracy (%)	Producer Accuracy (%)
1	62.37	50.76
2	37.24	0.03
3	46.61	88.43
4	0.00	0.00
5	0.00	0.00
6	26.66	40.10
7	2.95	0.15
8	0.00	0.00
9	0.00	0.00
10	0.00	0.00
11	61.21	43.60
12	0.00	0.00
13	0.00	0.00
14	5.47	91.53
15	0.00	0.00
16	0.00	0.00
17	35.14	1.08
18	0.00	0.00
19	6.02	2.10
20	6.56	74.09
21	1.68	69.93
22	0.00	0.00
23	0.00	0.00
24	0.00	0.00

OA: 30.19, mf1: 10.05, mIoU: 6.62.

VI. CONCLUSION

In conclusion, our work has introduced novel contributions to the field of semantic segmentation for land cover mapping using satellite imagery. We have explored and implemented new loss functions, such as the Dice Loss and Combined Focal-Dice Loss, to enhance the model's ability to handle imbalanced datasets and improve segmentation performance.

The experimental results on small datasets, particularly training on a limited subset of the five billion pixel dataset and predicting on specific target cities, demonstrate promising outcomes. Notably, the use of the Dice Loss function has shown substantial accuracy for classes even in such constrained scenarios. The Combined Focal-Dice Loss function also exhibits potential in balancing class-specific attention and segmentation quality.

Our contribution to data augmentation, including random vertical flipping and color jittering, could be a crucial factor contributing to the model's good performance despite the very small dataset. These techniques play a vital role in diversifying the training data, enhancing the model's ability to generalize to unseen scenarios.

It is important to note that our current results are based on training the model on very small datasets. To fully assess the capabilities of the introduced loss functions and the U-Net architecture, further experiments are required on the entire five billion pixel dataset and comprehensive target sets. This will allow us to gauge the impact of different loss functions on a larger scale and draw more robust conclusions regarding their effectiveness.

In future work, we plan to scale up our experiments to encompass the complete datasets, considering multiple source and target cities. This will enable a comprehensive evaluation of the proposed model's generalization capabilities and provide insights into the optimal choice of loss functions for diverse land cover mapping scenarios.

References

[1] X.-Y. Tong, G.-S. Xia, and X. X. Zhu, "Enabling country-scale land cover mapping with meter-resolution satellite imagery," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 196, pp. 178–196, 2023.