SEMANTIC SEGMENTATION OF SATELLITE IMAGERY USING POSITIVE AND UNLABELED LEARNING

by

MOHAMMAD ESHGHI

A DISSERTATION

Presented to the Department of Geography
and the Division of Graduate Studies of the University of Oregon
in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy

June 2022

DISSERTATION APPROVAL PAGE

Student: Mohammad Eshghi

Title: Semantic Segmentation of Satellite Imagery Using Positive and Unlabeled Learning

This dissertation has been accepted and approved in partial fulfillment of the requirements for the Doctor of Philosophy degree in the Department of Geography by:

Prof. Amy Lobben	Chair
Prof. Hui (Henry) Luan	Core Member
Prof. Lucas Silva	Core Member
Prof. Daniel Lowd	Institutional Representative

and

Krista Chronister	Vice Provost for Graduate Studies

Original approval signatures are on file with the University of Oregon Division of Graduate Studies.

Degree awarded June 2022

© 2022 Mohammad Eshghi
This work is licensed under a Creative Commons
Attribution 4.0 License.

DISSERTATION ABSTRACT

Mohammad Eshghi

Doctor of Philosophy

Department of Geography

June 2022

Title: Semantic Segmentation of Satellite Imagery Using Positive and Unlabeled Learning

The recent advances of deep learning in computer vision field have revolutionized digital image processing. The adoption of vision-based deep learning models in remote sensing has been promising. However, despite their success in remote sensing image processing, deep learning models suffer from labeled data scarcity, which is defined as the lack of large scale labeled datasets. This drawback is important to pay attention to since manually labeling data is labor-intensive and time-consuming. In addition, in many applications, the only information of interest is the presence of the application-specific landcover or object within an image, and thus it is not reasonable to spend extra time and cost to fully label the rest of an image. Therefore, remote sensing image processing benefits greatly from positive and unlabeled learning, which is a more general setting of semi-supervised learning and addresses the availability of only a few labeled examples of the presence of the application-specific event in a dataset.

This dissertation investigates the possibility of leveraging transfer learning and ensemble learning frameworks in a positive and unlabeled learning setting for semantic segmentation of satellite imagery. First, I create positive and unlabeled satellite imagery datasets from an available binary positive and negative dataset to be used for model development. Next, I develop a deep homogenous transfer positive and unlabeled learning which utilizes two distinct positive and negative as well as positive and unlabeled satellite imagery datasets acquired by a same satellite sensor (i.e. similar domain images). Building upon this, I extend the homogenous aspect of the developed model to the heterogeneous case. In doing so, the developed model will be able to not only learn from similar domain satellite images and non-satellite images but also to leverage satellite images from dissimilar domains. In the next stage, I develop a deep ensemble positive and unlabeled learning model in order to incorporate the advantages of multiple different models for a same task. Then, I investigate the possibility of a mixture of the proposed models in transfer learning and ensemble learning frameworks for PU learning. Finally, I conclude this dissertation by discussing the possible next steps for future works.

CURRICULUM VITAE

NAME OF AUTHOR: Mohammad Eshghi

GRADUATE AND UNDERGRADUATE SCHOOLS ATTENDED:

University of Oregon, Eugene, OR, USA
K.N. Toosi University of Technology, Tehran, Tehran, Iran

DEGREES AWARDED:

Doctor of Philosophy, Geography, 2022, University of Oregon
Master of Science, Computer and Information Science, 2021, University of Oregon
Master of Science, Geographic Information Systems Engineering, 2016, K.N. Toosi University of Technology
Bachelor of Science, Geodesy and Geomatics Engineering, 2013, K.N. Toosi University of Technology

AREAS OF SPECIAL INTEREST:

Machine Learning
GIScience
Computer Vision
Remote Sensing

PROFESSIONAL EXPERIENCE:

Graduate Employee, University of Oregon, Eugene, OR, USA, 2016-2022

Data Scientist (Intern), Apple, Cupertino, CA, USA, 2021

Machine Learning Research Scientist (Intern), Radiant Earth Foundation, Washington D.C., D.C., USA, 2021

Data Engineer (Intern), Apple, Cupertino, CA, USA, 2020

Instructor, Machine Learning (GeoAI), University of Oregon, Eugene, OR, USA, 2020

Full-Stack Web Developer and Data Analyst, University of Oregon, Eugene, OR, USA, 2017-2020

Instructor, GIScience-I, University of Oregon, Eugene, OR, USA, 2017

GRANTS, AWARDS AND HONORS:

Graduate Research Fellow (National Science Foundation Funded), National Socio-Environmental Synthesis Center (SESYNC), 2021

Mather Graduate Fellowship, Department of Geography, University of Oregon, 2021

Graduate Research Pursuit (National Science Foundation Funded), National Socio-Environmental Synthesis Center (SESYNC), 2019-2021

Travel grant for Summer Institute on Cyberinfrastructure for Socioenvironmental Synthesis (National Science Foundation Funded), National Socio-Environmental Synthesis Center (SESYNC), 2019

Travel grant for Summer School on Reproducible problem solving with CyberGIS and Geospatial Data Science (National Science Foundation Funded), American Association of Geographers-University Consortium for Geographic Information Science (AAG-UCGIS), 2019

Rippey Graduate Student Research Award, Department of Geography, University of Oregon, 2019, 2021

Graduate Student Research Travel Award, Department of Geography, University of Oregon, 2018

PUBLICATIONS:

Eshghi, M., & Schmidtke, H. R. (2018). An approach for safer navigation under severe hurricane damage. *Journal of Reliable Intelligent Environments*, (Vol. 4), pp. 161–185.

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to my advisor, Prof. Amy Lobben, whose consistent support and encouragement I will never forget. Furthermore, I am thankful to and grateful for my parents and my sisters whose constant limitless and unconditional love and support keep me motivated and confident. My accomplishments and success are because of their sacrifices for me. Last but not least, I thank my dear partner, Emily, for her love, patience, and support.

To my mother and father, to whom I owe everything.

To my sisters and my partner, for being driving force in my life.

TABLE OF CONTENTS

Chapter Page

I. INTRODUCTION 1
  1.1. Dissertation Scope & Contributions 2
    1.1.1. Deep Transfer Positive and Unlabeled Learning 3
    1.1.2. Deep Ensemble Positive and Unlabeled Learning 4
  1.2. Dissertation Outline 5
II. LITERATURE REVIEW 6
  2.1. Positive and Unlabeled Learning 6
  2.2. Transfer Learning 16
  2.3. Ensemble Learning 22
III. DATASETS 30
  3.1. Standard Open Image Datasets 30
  3.2. Massachusetts Buildings Dataset 30
  3.3. Inria Aerial Image Dataset 34
    3.3.1. Positive and Unlabeled Dataset 36
      3.3.1.1. Homogeneous Case 39
IV. DEEP TRANSFER POSITIVE AND UNLABELED LEARNING 41
  4.1. Background 42
    4.1.1. Semi-supervised Learning 42
    4.1.2. Weakly-supervised Learning 45
    4.1.3. Semi-supervised Domain Adaptation 46
    4.1.4. Unsupervised Domain Adaptation 48
    4.1.5. The motivation behind this work 50

Chapter Page

- 4.2. Baselines 50
  - 4.2.1. Homogeneous case 52
  - 4.2.2. Heterogeneous case 53
- 4.3. Methodologies 53
  - 4.3.1. Homogeneous case 54
    - 4.3.1.1. Model architecture: UNET 55
    - 4.3.1.2. Model backbone: VGG16 56
    - 4.3.1.3. Learning module: supervised 57
    - 4.3.1.4. Learning module: self-training 58
    - 4.3.1.5. Learning module: positive-unlabeled 59
    - 4.3.1.6. Performance metric: Accuracy 61
    - 4.3.1.7. Performance metric: IoU 61
    - 4.3.1.8. Performance metric: F1-score 62
  - 4.3.2. Heterogeneous case 62
    - 4.3.2.1. Model architecture: DeepLabV2 65
    - 4.3.2.2. Model backbone: ResNet-101 66
    - 4.3.2.3. Learning module: consistency-training 68
- 4.4. Results 69
  - 4.4.1. Homogeneous case 69
    - 4.4.1.1. Conclusions 82
  - 4.4.2. Heterogeneous case 85
    - 4.4.2.1. Ablation Study 90
    - 4.4.2.2. Conclusions 91
- 4.5. Conclusion 91

Chapter Page

V. DEEP ENSEMBLE POSITIVE AND UNLABELED LEARNING 92

5.1. Background 93

5.2. Methodologies 97

5.2.1. Dropout 98

5.2.2. EMA 98

5.2.3. SWA 99

5.2.4. Feature ensemble 99

5.2.5. Model ensemble 99

5.2.6. TreeNet 100

5.2.7. Contextual ensemble 100

5.2.8. Multi-scale ensemble 100

5.3. Results 103

5.4. Conclusion 111

VI. CONCLUSIONS & FUTURE WORK 113

6.1. Future Work 115

REFERENCES CITED 116

LIST OF FIGURES

Figure Page

1. From left to right: fully-labeled binary data and classifier, one-class labeled data and classifier, partially-labeled binary data and unlabeled data and semi-supervised classifier, and positive and unlabeled data and classifier. Blue points are positive data, red points are negative data; and bright colors are labeled data and pale colors are unlabeled data. 7

2. Transfer Learning is a framework for reusability of the extracted knowledge in a large-data source domain to train a model in a limited-data target domain 16

3. Sample images from ImageNet-ILSVRC dataset (Russakovsky et al., 2015). 31

4. (a) Left and right columns show images and their corresponding labels, respectively. From top to bottom: a sample from Massachusetts buildings dataset (Mnih, 2013) and samples from its derived patches, one from the middle, and the rest from top-left, top-right, bottom-left, and bottom-right corners, respectively, showing the mirroring effect when not enough pixels are available for patch creation. (b) Selected final patches: images and their corresponding labels in the left and right columns. 33

5. (a) Left and right columns show images and their corresponding labels, respectively. From top to bottom: a sample from Inria Aerial Image Dataset (Maggiori, Tarabalka, Charpiat, & Alliez, 2017) and samples from its derived patches, one from the middle, and the rest from top-left, top-right, bottom-left, and bottom-right corners, respectively, showing the mirroring effect when not enough pixels are available for patch creation. (b) Selected final patches: images and their corresponding labels in the left and right columns. 35

Figure Page

6. PU data is created using a moving window over labels corresponding to each image within the train set. At each location of the moving window pixels that belong to the positive class are changed to unlabeled with complete randomness (i.e. no selection bias). 37

7. Left and right columns show images and their corresponding PU labels created by the moving window strategy. 38

8. Different PU train datasets created from Inria Aerial Image dataset with different missing probabilities for the positive class, all of which satisfying the SCAR assumption—graphs in black, blue, green, yellow, red, and orange show the distribution of the positive class and of the positive data in PU5, PU6, PU7, PU8, and PU9 datasets, respectively. 39

9. Homogeneous transfer positive and unlabeled learning architecture. . . . 55

10. UNET architecture; the left-side is the contracting module (i.e. encoder), the right-side is the expansive module (i.e. decoder), the brown square is the segmentation output, and, finally, the grey squares are intermediary feature maps to be concatenated to their corresponding upsampled layer. 57

11. The electromagnetic spectrum captured by satellite remote sensing (shown as SRS in the image). Image adopted from Pettorelli, Schulte to Bühne, C. Shapiro, and Glover-Kapfer (2018). . . . 63

12. The misalignments of different remote sensors across different bands. Image adopted from Rocchio and Barsi (n.d.). 63

13. Heterogeneous transfer positive and unlabeled learning architecture. . . 65

14. DeepLabV2 architecture; top-to-bottom are three repetitions of left-to-right path in each row over 1.0, 0.75, and 0.5 downscaled input images; left-to-right path are multi-scale feature generators with different atrous rates from 6 to 24; the final feature map is upsampled to generate the input-size-like output. 67

15. A sample building block showing the idea of residual learning. Figure adopted from K. He, Zhang, Ren, and Sun (2016). 68

Figure Page

16. Residual learning block: 2- versus 3-layer. Figure adopted from K. He et al. (2016). 68

17. The training loss (in red) and validation loss (in green) for Target-PU model without ImageNet's pre-trained weights on (a) PU5, (b) PU6, (c) PU7, (d) PU8, (e) PU9 datasets. 74

18. The training loss (in red) and validation loss (in green) for Target-PU model with ImageNet's pre-trained weights on (a) PU5, (b) PU6, (c) PU7, (d) PU8, (e) PU9 datasets. 75

19. The training loss (in red) and validation loss (in green) for Target-FT model without ImageNet's pre-trained weights on (a) PU5, (b) PU6, (c) PU7, (d) PU8, (e) PU9 datasets. 76

20. The training loss (in red) and validation loss (in green) for Target-FT model with ImageNet's pre-trained weights on (a) PU5, (b) PU6, (c) PU7, (d) PU8, (e) PU9 datasets. 77

21. The training loss (in red) and validation loss (in green) for Seamless-U model without ImageNet's pre-trained weights on (a) PU5, (b) PU6, (c) PU7, (d) PU8, (e) PU9 datasets. 78

22. The training loss (in red) and validation loss (in green) for Seamless-U model with ImageNet's pre-trained weights on (a) PU5, (b) PU6, (c) PU7, (d) PU8, (e) PU9 datasets. 79

23. The training loss (in red) and validation loss (in green) for Seamless-PU model without ImageNet's pre-trained weights on (a) PU5, (b) PU6, (c) PU7, (d) PU8, (e) PU9 datasets. 80

24. The training loss (in red) and validation loss (in green) for Seamless-PU model with ImageNet's pre-trained weights on (a) PU5, (b) PU6, (c) PU7, (d) PU8, (e) PU9 datasets. 81

25. Left-to-right: Sample image patches from homogeneous PU Inrial Dataset, the corresponding ground truth, and predictions from four models trained on the PU5 case: (i) Target-PU without pre-trained weights, (ii) Target-FT without pre-trained weights, (iii) Seamless-U with pre-trained weights, and (iv) Seamless-PU with pre-trained weights. 83

Figure Page

26. Left-to-right: Sample image patches from homogeneous PU Inrial Dataset, the corresponding ground truth, and predictions from four models trained on the PU9 case: (i) Target-PU without pre-trained weights, (ii) Target-FT without pre-trained weights, (iii) Seamless-U with pre-trained weights, and (iv) Seamless-PU with pre-trained weights. 84

27. Left-to-right: Sample image patches from heterogeneous PU Inrial Dataset, the corresponding ground truth, and predictions from (i) PixMatch(*Augmentations*) , (ii) Seamless-PU with UNET trained on PU5, (iii) Seamless-PU with DeepLabV2 trained on PU5, (iv) Seamless-PU with UNET trained on PU9, (v) Seamless-PU with DeepLabV2 trained on PU9. 89

28. Different deep ensemble architectures. 101

29. self-ensemble multi-scale UNET. 103

30. Cosine similarity of model weights for TreeNet (Decoder) model over different checkpoints at different epochs during the training phase for PU9 dataset. 106

31. Cosine similarity of model weights for TreeNet (Encoder) model over different checkpoints at different epochs during the training phase for PU9 dataset. 108

32. Cosine similarity of model weights for ModelNet model over different checkpoints at different epochs during the training phase for PU9 dataset. 108

33. Left-to-right: Sample image patches from homogeneous PU Inrial Dataset, the corresponding ground truth, and predictions from the baseline model and the proposed model trained on PU9 dataset. 110

LIST OF TABLES

Table Page

1. Massachusetts Buildings Dataset Characteristics 32

2. Inria Aerial Image Dataset Characteristics 36

3. The performance of the proposed model and the baselines in the target domain 72

4. Performance drop of PUDA model (Seamless-PU) without the consistency loss on heterogeneous PU data from Inria Image Dataset. The model uses a warm start using the ImageNet pre-trained weights. 86

5. Performance of unsupervised domain adaptation model (PixMatch) and the proposed PUDA model (Seamless-PU). All utilize a warm start using the ImageNet pre-trained weights. 88

6. The effect of the degree of the magnitude that the consistency loss is incorporated, which is shown for the case of PU9 dataset. 90

7. The performance of different ensemble models compared to the performance of a single model. 107

8. The mixture of Ensemble and Transfer PU models. 112

CHAPTER I

INTRODUCTION

Remotely-sensed (RS) imagery is among the most valuable geographically referenced spatial big data that provides a unique opportunity for understanding earth's surface (Miller & Goodchild, 2015; S. Wang et al., 2015). Such an opportunity is of great importance in different scientific domains including but not limited to geography, ecology, epidemiology, social sciences, and emergency management (S. Wang et al., 2015). In these scientific domains, it is crucial to have access to and to be able to process the most recent imageries in order to extract the necessary information. This requires frequent image acquisition and automatic image processing. In terms of image acquisition, the production rate of RS imagery has exploded in recent years due to increase in the number of satellites and advances in drone imagery. On the other hand, with regards to automatic image processing,the recent advances in deep learning in computer vision have revolutionized digital image processing. For example, as a result of the success of convolutional neural networks in different vision tasks such as object detection (Girshick, Donahue, Darrell, & Malik, 2015; Redmon, Divvala, Girshick, & Farhadi, 2016) and semantic segmentation (Long, Shelhamer, & Darrell, 2015; Noh, Hong, & Han, 2015; Ronneberger, Fischer, & Brox, 2015a), the interest in deep learning models has greatly increased within the remote sensing research community (X. X. Zhu et al., 2017).

Semantic segmentation task (also known as pixel-based classification in remote sensing research community) refers to assigning a label (of a landcover type or an object) to every pixel within an image (Kemker, Salvaggio, & Kanan, 2018). Semantic segmentation of RS imagery usually requires labeled datasets with samples that are representative of all landcover types within the area in order to train a classifier (W. Li, Guo, & Elkan, 2010), whereas only one specific class of landcover types (or a specific class of objects) might be of interest (Foody, Mathur, Sanchez-Hernandez, & Boyd, 2006). For example, landcover classes such as agricultural lands may not be of interest when the objective is to identify the manmade structures such as roads or buildings, and vice versa (Foody et al., 2006; W. Li et al., 2010). In such cases, the intrinsically labor-intensive and time-consuming approach of creating a set of representative training samples can be avoided. However, this results in unsuitability of the widely-used supervised learning framework, and thus alternative approaches, such as positive and unlabeled learning framework, should be taken.

Positive and unlabeled (PU) learning is a more general setting of semi-supervised learning, in which a binary classifier is learned based on a set of (limited) labeled data for the positive class only (i.e. the class of interest) and a very large amount of unlabeled data containing both positive and negative classes (Bekker & Davis, 2020). Despite its significance, little attention has been paid to PU learning within the remote sensing literature, whereas machine learning and deep learning for remote sensing have a rich literature—for more details on machine learning and deep learning in remote sensing, see Maulik and Chakraborty (2017) and Y. Li, Zhang, Xue, Jiang, and Shen (2018).

1.1 Dissertation Scope & Contributions

In this dissertation, I investigate the potentials of deep learning-based PU learning for semantic segmentation of satellite imagery. The contributions of this dissertation can be categorized into two major learning frameworks: (i) Transfer Positive and Unlabeled Learning, and (ii) Ensemble Positive and Unlabeled

Learning. The following pages present the two research questions and an overview of the main contributions of this dissertation.

> ***RQ1.*** *How can Transfer Learning be incorporated in the context of positive and unlabeled learning for semantic segmentation of satellite imagery?*
>
> ***RQ2.*** *How can Ensemble Learning be incorporated in the context of positive and unlabeled learning for semantic segmentation of satellite imagery?*

1.1.1 Deep Transfer Positive and Unlabeled Learning.

Transfer learning (TL) is the idea of transferring the informative knowledge from one domain (i.e. a source domain) to another (i.e. a target domain) (Torrey & Shavlik, 2010). TL is suitable when learning from the source domain with large amount of labeled data can be informative to find a model using the limited labeled data in the target domain, which can do better compared to models trained only by the limited labeled data in the target domain (Karbalayghareh, Qian, & Dougherty, 2018)–for example, the limited labeled data in the target domain could be the case for PU data (J. Chen & Liu, 2014). Although Transfer PU learning has been studied outside of remote sensing literature to some extent (J. Chen & Liu, 2014; Mignone & Pio, 2018), no study has been found (especially with a deep learning-based approach) for semantic segmentation of RS imagery.

TL can be categorized into homogeneous and heterogeneous TL. In homogeneous TL, the source domain and target domain have the same feature space $X_S = X_T$. Heterogeneous TL, however, is considered when the source domain and target domain have different feature spaces $X_S \neq X_T$ (Bashath et al., 2022a). I consider homogeneous TL for situations when a model is developed on a set of satellite images from a geographic location, and then it is used on images from the same geographic location. The problem here is that the spectral signature of objects is not exactly the same–i.e. different marginal distributions, $P_S(X) \neq P_T(X)$, but same feature spaces, $X_S = X_T$. I propose a solution which focuses on both shared feature spaces and shared parameters in a deep learning-based framework. Next, I consider heterogenous TL for situations when a model is developed on a set of general image datasets (like ImageNet) or satellite images from a geographic location for a learning task, and it is used on satellite images or satellite images from another geographic location. In this case, since the two domains are related (i.e. both are images) transferring information is still possible. I extend my proposed model for homogeneous case to heterogenous case, in which the goal is to close the gap between feature spaces of the source and target domains.

1.1.2 Deep Ensemble Positive and Unlabeled Learning.

Ensemble learning (EL) is essentially a methodological framework based on a voting mechanism, in which the predictive power of multiple different learning algorithms is fused in order to achieve a better predictive performance through the collective voting–as opposed to any of the participating learning algorithms individually (Dong, Yu, Cao, Shi, & Ma, 2020a; Sagi & Rokach, 2018a). In remote sensing, research has shown the robustness of ensemble classifiers rather than a single method strategy (e.g. X. Huang & Zhang, 2012a). However, little research has focused on the ensemble PU learning in remote sensing leaving a gap in this area of research. R. Liu et al. (2018) is among the few research targeting the ensemble PU learning for RS imagery. Therefore, this dissertation investigates deep ensemble learning frameworks in a PU learning setting. Building upon this, I propose a deep ensemble PU learning that benefits from both ensemble and PU worlds for semantic segmentation of RS imageries.

1.2 Dissertation Outline

The remainder of this dissertation is organized as follows. I provide an overview of studies related to (i) PU learning, (ii) TL, and (iii) EL, in general, and with a focus on remote sensing literature, in particular, in Chapter II. Next, in Chapter III, I introduce the datasets used in this dissertation as well as the PU dataset that I create on top of one of these datasets. Chapter IV presents my approach for homogenous and heterogeneous transfer PU learning. Then, I present the proposed ensemble PU learning in Chapter V. Finally, I conclude and summarize my contributions and further discuss the opportunities for future work in Chapter VI.

CHAPTER II

LITERATURE REVIEW

In this chapter, I review the recent research in positive and unlabeled (PU) learning and other learning frameworks that are used as the base for developing my methodologies. The literature review is categorized into three sections: (i) PU Learning, (ii) Transfer Learning, and (iii) Ensemble Learning.

2.1 Positive and Unlabeled Learning

PU learning is a more general setting of both semi-supervised learning and one-class classification, in which a binary classifier is learned based on a limited set of labeled data from only one class and a very large amount of unlabeled data containing examples from both positive and negative classes. Therefore, PU learning is considered for situations where the data for negative class is either absent or distributed too diversely (X. Chen, Gong, & Yang, 2021; M. Du Plessis, Niu, & Sugiyama, 2015). As shown in Fig. 1, PU learning differs from semi-supervised learning because it uses only labeled examples from positive class. It also differs from one-class classification since it utilizes unlabeled data examples (Bekker & Davis, 2020). It should be mentioned that the negative and unlabeled (NU) learning task is applied when the labeled data belongs to the negative class, and thus NU learning is essentially the same as PU learning (Hammoudeh & Lowd, 2020; Niu, du Plessis, Sakai, Ma, & Sugiyama, 2016). The significance of PU learning is due to its high number of applications (X. Chen et al., 2021). However, PU learning is challenging due to the difficulty in model selection, model building, and model assessment (W. Li, 2013).

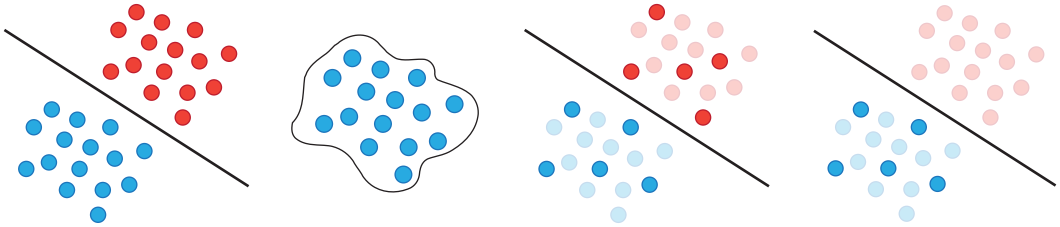


Figure 1. From left to right: fully-labeled binary data and classifier, one-class labeled data and classifier, partially-labeled binary data and unlabeled data and semi-supervised classifier, and positive and unlabeled data and classifier. Blue points are positive data, red points are negative data; and bright colors are labeled data and pale colors are unlabeled data.

In binary classification, a classifier is trained using a set of training examples of form $\{x, y\}$ where x is a set of attributes/predictors/variables and y is the class label; usually $y = 1$ refers to class positive and $y = 0$ refers to class negative. In PU learning, however, the training data set contains examples of form $\{x, y, s\}$ where x and y are the same as the binary case. The new element here is s which is used to denote whether the example is labeled ($s = 1$) or not ($s = 0$). In situations with PU data, the binary loss functions are not valid anymore resulting in the emergence of different PU learning algorithms.

Categories of PU learning algorithms include: two-step techniques, biased learning models, and learning with incorporation of the positive class prior (Bekker & Davis, 2020; Jaskie & Spanias, 2019). The two-step techniques assume that positive data are very similar and are different from negative data. Therefore, by identifying reliable negative samples from unlabeled data at the first step, a supervised or semi-supervised learning model can be trained in the second step by utilizing positive, reliable negative, and the rest of the unlabeled samples (Bekker & Davis, 2020). For example, Dhurandhar and Gurumoorthy (2020) propose an approach in which, using an unsupervised framework and independent from classifiers, they identify and weigh positive and negative examples from unlabeled examples. B. Liu, Lee, Yu, and Li (2002) employ a *spy* technique in which, first, they put a subset of randomly selected positive examples into the unlabeled set. Then, they determine a probability threshold to identify negative examples from the unlabeled data. P. Yang, Liu, and Yang (2017) propose an adaptive sampling approach for identifying reliable negative data from the unlabeled data. They first consider all unlabeled data as negative data. Then, using a predictive model trained on positive and negative (PN) data, the negative class probability for unlabeled data is calculated. The probabilities for all unlabeled data belonging to positive or negative classes are calculated iteratively by repeating the last two steps, which, at the end, results in one or more robust negative dataset(s) based on likelihood threshold. The two-step strategy has a drawback: identifying negative samples is not always reliable and cannot always be accurate, thus such situations result in poor performances.

In biased techniques, the unlabeled data are considered noisy samples of the negative class. For example, Biased-SVM considers a modified cost function to address the noisy data (Jaskie & Spanias, 2019; B. Liu, Dai, Li, Lee, & Yu, 2003). W. S. Lee and Liu (2003) perform weighted logistic regression in which positive samples have larger weights in comparison to the negative samples (Bekker & Davis, 2020). H. Shi, Pan, Yang, and Gong (2018) solve PU learning by focusing on risk minimization. They propose a decomposition technique for the loss function in order to model the noisy negative labels. They show that estimations of the negative class centroid can reduce the adverse effect caused by noisy negative labels. To improve this further, Gong et al. (2019) propose a kernelized version of the algorithm proposed by H. Shi et al. (2018) in order to address the non-linear classifiers/decision boundaries. All in all, the performance of biased techniques depends on the number of positive samples within the unlabeled set. In situations where a large number of unlabeled data belong to the positive class, negative unlabeled data adds too much uncertainty to the learning process, which results in poor performances.

PU models that incorporate a class prior are probabilistic approaches that rely on an estimation of the ratio of the positive instances in a population. Such estimation requires an assumption on labeling mechanism, one of the most popular being Selected Completely At Random (SCAR) (Bekker & Davis, 2020). In SCAR, it is assumed that the labeled data are uniformly selected from the positive data resulting in possibility of using traditional classifiers (Bekker & Davis, 2020; Elkan & Noto, 2008). For example, Elkan and Noto (2008) first train a non-traditional classifier on PU data to estimate the weights (i.e. the label frequency or class prior) of examples on a validation set. Then, they convert the non-traditional classifier to a traditional classifier using the found weights. Recent research has been trying to relax the SCAR assumption and consider less restrictive assumptions on labeled data. For example, Bekker, Robberechts, and Davis (2019) relax the SCAR assumption to the less restrictive Selected At Random (SAR) assumption that considers non-uniform labeling mechanisms. The SCAR assumption also relies on availability of the class prior on which research (i.e. M. C. Du Plessis, Niu, & Sugiyama, 2016; Jain, Delano, Sharma, & Radivojac, 2020; Łazęcka, Mielniczuk, & Teisseyre, 2021; Perini, Vercruyssen, & Davis, 2020) has demonstrated how to estimate the class prior from data. On the other hand, in some cases (C. Zhang, Hou, & Zhang, 2020), the goal is to learn a model without pre-estimating the class prior (in an isolated step).

A further advancement in class prior-based category is to find suitable risk estimators for PU learning. The approaches in this category aim to apply distinct loss functions that satisfy specific conditions to PU risk estimators. For example, M. C. Du Plessis, Niu, and Sugiyama (2014) propose a cost-sensitive learning between positive data and unlabeled data which results in using non-convex loss functions, such as the ramp loss, in order to avoid the systematic estimation bias. The cost-sensitive classifier does depend on the class prior probability estimation unless the unlabeled data contains mostly positive examples. As an improvement on the work by M. C. Du Plessis et al. (2014), M. Du Plessis et al. (2015) propose a convex PU loss called double hinge loss that considers an ordinary convex loss function for unlabeled samples and a composite loss function for positive samples. Their loss function performs as accurate as the non-convex ramp loss while (i) it can still cancel the systematic estimation bias, (ii) it causes estimators to converge to the optimal solutions at the optimal parametric rate, and (iii) it has a much lower computational burden. As an improvement on this approach, Kiryo, Niu, Du Plessis, and Sugiyama (2017) propose a non-negative risk estimator for PU learning, which modifies the lower-bound of the double hinge loss. In comparison to the unbiased risk estimators, the non-negative risk estimator will not produce negative empirical risks in case of learning flexible models such as deep neural networks. The non-negative risk estimator is more robust against overfitting, and, thus, it can be used on deep neural networks even when positive data are limited. Building on non-negative risk estimator, X. Chen, Liu, Tu, Cao, and Yang (2018) propose a manifold non-negative risk estimator by adding a manifold regularizer to the non-negative risk estimator.

Following on deep learning-based models, H. Chen, Liu, Wang, Zhao, and Wu (2020) propose a general variational principle for PU learning which introduces a loss function which can be efficiently calculated without involving class prior estimation. Na et al. (2020) propose a generative PU learning based on variational auto encoders called VAE-PU that does not rely on SCAR assumption. The proposed deep generative model is used to virtually generate unobserved data of PU data in order to satisfy the risk function requirement resulted from SCAR-independence condition. Guo et al. (2020) propose PUGAN which is based on deep Generative Adversarial Networks (GANs), in which the assumption is that the data produced by the generator should be considered as unlabeled rather than negative, and, thus, the problem in GAN models become a PU learning rather than a standard PN learning. Hou, Chaib-Draa, Li, and Zhao (2017) propose GenPU in which GAN framework is used to identify both positive and negative data distributions. Hu et al. (2021) propose a GAN-based model called PAN in which they develop a new objective function based on Kullback–Leibler divergence in order to address the problem with GAN which is that it focuses disproportionately on the positive class (which in the case of GANs is the real data). Zheng, Yuan, Wu, Li, and Lu (2019) propose a one-class GAN for fraud detection with only benign users as training data. The generator searches for the probability distribution of the malicious users, and the discriminator attempts to distinguish between the generated malicious examples from the generator and the benign examples from training data. Chiaroni, Rahal, Hueber, and Dufaux (2018) apply GANs to learn the distribution of the negative class by generating fake images, the distribution of which is closer to the distribution of negative class within the unlabeled dataset. Then, a supervised convolutional neural network

(CNN) classifier is trained on the positive and fake generated negative samples. Finally, in comparison to most PU research focusing on classification, Y. Yang, Liang, and Carin (2020) propose the object detection problem as a PU problem due to the challenging nature of collecting complete labels for object detection because of the large number of instances, which makes this task more difficult than simply a classification task. They utilize the Faster-RCNN architecture (Ren, He, Girshick, & Sun, 2015) with NNPU loss function adopted from Kiryo et al. (2017).

Finally, Niu et al. (2016) compare PU learning to PN learning and demonstrate that, under certain conditions, PU learning can perform as well as PN learning. They establish risk bounds of risk minimizers in cases of PN and PU data in order to investigate settings in which PU could outperform PN learning. They estimate error bounds of the risk minimizers and discover that the PU bound could be tighter than the PN bound under some assumptions such as *Rademacher complexity* of the decision function being bound by a constant as well as the size of the dataset. The aforementioned situation of the error band is proved in both finite-sample and asymptotic cases in which, the size of unlabeled data should increase at a rate faster than the increase in the sizes of positive and negative data. Their results are independent of the specific forms of the function class and/or data distribution. However, they still rely on availability of the class-prior probability.

- **Positive and Unlabeled Learning in Remote Sensing.** Remote sensing classification is a complex process and requires consideration of many factors, including selection of training samples and suitable classification approaches (Lu & Weng, 2007). With regards to the type of classification approach, there are three general categories: (i) pixel-, (ii) sub-pixel-, and (iii) object-based methods. Pixel-based semantic segmentation algorithms assign a label to every pixel in an image (Kemker et al., 2018), for which supervised learning has been traditionally used (Dey, Zhang, & Zhong, 2010; W. Li et al., 2010). A crucial step in the supervised approach is to collect representative training samples for all landcover types in the image to ensure the success of training a classifier (W. Li, 2013; Tuia, Persello, & Bruzzone, 2016). If the goal is to identify only a specific class of landcover types, such as urban regions vs non-urban regions (W. Li et al., 2010), then it is not reasonable to collect negative data in a representative way since the negative class is too diverse (M. Du Plessis et al., 2015). Therefore, many researchers have started investigating PU learning classifiers for semantic segmentation of RS imagery, including both active and passive imageries.

PU learning in remote sensing research originates from ecological niche modeling for when the only available data is information about the presence of the specie-of-interest. One-class classifiers such as one-class Support Vector Machine (SVM) (Schölkopf, Platt, Shawe-Taylor, Smola, & Williamson, 2001) and Maximum Entropy (MaxEnt) models (Phillips, Anderson, & Schapire, 2006) are the most used approaches for one-class species distribution modeling (W. Li, Guo, & Elkan, 2011). These methods are adopted for species distribution extraction, such as different types of vegetation, using semantic segmentation of RS imageries (Rapinel & Hubert-Moy, 2021). For example, X. Liu, Liu, Gong, Lin, and Lv (2017) study the effectiveness of different one-class classification methods to detect invasive plants. In addition, one-class classifiers are used for other tasks such as single-land-cover detection (W. Li & Guo, 2010), building detection (Krupiński, Lewiński, & Malinowski, 2019; W. Yang, Yin, Song, Liu, & Xu, 2013), mapping forced labor (McDonald et al., 2021), and flood masks (Brill, Schlaffer, Martinis, Schröter, & Kreibich, 2021). Therefore, because of their extensive adoption, these methods are extensively analyzed for one-class classification of RS imagery, demonstrating that even the best performers among these methods still functions poorly in many cases (Mack & Waske, 2017). Although other flavors of one-class classification, such as sparse representation (Y. Chen, Nasrabadi, & Tran, 2012; Ran, Zhang, Li, & Du, 2016; Song, Li, & Jia, 2020; Song, Li, Li, & Plaza, 2016), are studied with competitive performances to other methods in their category, there are a few important challenges regarding training an accurate one-class classification model, including (i) when the size of and/or the number of the training image(s) is large, and (ii) when the number of positive data examples is small (Mack et al., 2016).

In order to take advantage of the unlabeled data, based on the idea introduced by Elkan and Noto (2008), W. Li et al. (2010) propose a PU learning which includes training a classifier on PU data based on the SCAR assumption, followed by dividing the classifier by the constant probability that a positive example is labeled to achieve the final desired classifier. An advantage of their proposed model is the possibility of its implementation with neural networks. Their model appears to be superior to Biased SVM, Gaussian domain descriptor, one-class SVM (W. Li et al., 2010) as well as to binary classifiers such as SVMs and maximum likelihood classifiers (Deng, Li, Liu, Guo, & Newsam, 2018). Therefore, their proposed model is used for many applications including extraction of built-up areas (Djerriri, Benyelles, Attaf, & Cheriguene, 2019; Xia, Yokoya, & Adriano, 2021), RS images' time series analysis (Desloires, Ienco, Botrel, & Ranc, 2022), and crop mapping (L. Zhao et al., 2019). B. Liu, Zhu, Wang, Liu, and Yu (2011) have also demonstrated the success of the model by Elkan and Noto (2008) for passive SAR imageries. C. Zhu, Liu, Yu, Liu, and Yu (2012) point out that the PU model by Elkan and Noto (2008) is very sensitive to the precision of estimating the positive class frequency, which could be low when few positive data are present. Therefore, they try to mitigate this issue by implementing a two-step approach: (i) unreliable positive and reliable negative examples are identified using the Spy detection method by B. Liu et al. (2002)—these examples are assigned different weights calculated by the method in Elkan and Noto (2008), and, then, (ii) these examples are used along with the (reliable) positive examples, and their respective weights to train a binary classifier. Finally, Gui, Xu, Wang, Yang, and Pu (2020) propose a two-step PU learning approach for extracting built-up areas from PolSAR imageries.

The possibility of using PU learning in a deep learning framework is important in remote sensing research. Xia and Yokoya (2021) investigate a self-paced PU learning for identifying and assessing damage to buildings. Their approach includes a two-step PU learning and a three-part loss function, one of which is a hybrid loss function of cross-entropy supervised loss and the unbiased risk estimator introduced by M. Du Plessis et al. (2015). Jian, Chen, and Cheng (2021) investigate the method proposed in Zheng et al. (2019) for change detection in time series RS imageries. Finally, Lei et al. (2021) propose a deep learning-based PU learning which takes advantage of CNNs and PU loss function by Kiryo et al. (2017). They make up the PU data from PN data for a training set, but keep the PN data for test set. The positive data is created very carefully to ensure zero noise. In addition, they create a subset of unlabeled data/pixels within an image, which could be due to the architecture style they choose. In this case, the architecture is a point-wise classification rather than a full segmentation architecture, which may slow the prediction time. Finally, Santara et al. (2019) explore PU learning for hyperspectral imageries using NNPU-based risk minimization by Kiryo et al. (2017), and a one-versus-all classification based on a two-step PU learning approach.

2.2 Transfer Learning

Transfer learning (TL) is the idea of transferring informative knowledge from one domain (i.e. source domain) to another domain (i.e. target domain) (Fig. 2). TL is suitable when learning from source domain with large amount of labeled data can inform the target domain model with limited labeled data—which performs better relative to training from scratch in the target domain with limited label (Torrey & Shavlik, 2010). PU data domain is an example of limited data domain (Mignone & Pio, 2018). Transferring informative knowledge among domains is all the more important when a domain may not have enough information from labeled data, but, when domains combined do have enough information to construct a representative model—examples for this in PU learning are Mignone and Pio (2018) and B. Liu et al. (2022). Therefore, the key with TL is the ability to transfer only the beneficial part of the knowledge learned in the source domain to the target domain (Day & Khoshgoftaar, 2017).

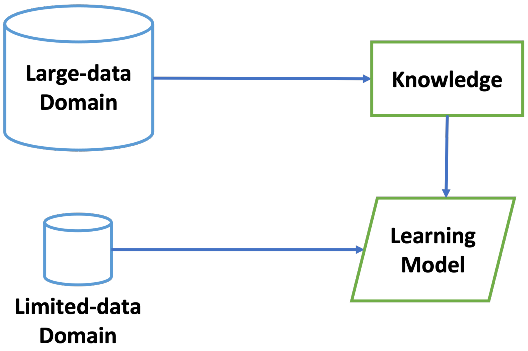


Figure 2. Transfer Learning is a framework for reusability of the extracted knowledge in a large-data source domain to train a model in a limited-data target domain

Concepts in TL include domain and its corresponding task(s). A domain is defined by the feature space, X, and marginal probability distribution over the feature space, $P(X)$. Given a domain, a task is defined by the label space, Y, and a predictive function, $f(\cdot) = P(Y|X)$, that provides a prediction for the label of a given data example (e.g. pixel) (Tan et al., 2018). Since TL is a family of different strategies and is not referring to a specific methodology (Bashath et al., 2022b), TL categorization has evolved to include two approaches, homogeneous and heterogeneous (Weiss, Khoshgoftaar, & Wang, 2016). TL is considered homogeneous when the data in the source and target domains have a same feature space ($X_S = X_T$) and labels ($Y_S = Y_T$). As a note, if $X_S = X_T$ but the feature distributions are different ($P_S(X) \neq P_T(X)$), it is called domain adaptation (DA). On the other hand, TL is defined as heterogeneous when the data in the source and target domains have nonequivalent (and non-overlapping) feature spaces ($X_S \neq X_T$) with possible nonequivalent labels ($Y_S \neq Y_T$) (Day & Khoshgoftaar, 2017).

Homogeneous TL can be categorized into instance-, feature- (symmetric or asymmetric), parameter-, relational-informational-, and hybrid-based techniques (Weiss et al., 2016), some of which are of interest in my dissertation. In instance-based techniques, the data samples in the source domain are reweighed to be directly used in the target domain—e.g. Asgarian et al. (2018). Feature-based techniques target the gap between the marginal and conditional distributions of the source and target domains, and try to close the gap using asymmetric or symmetric feature transformation. Parameter-based techniques try to share model parameters between source and target domain. For example, Oquab, Bottou, Laptev, and Sivic (2014) investigate the idea of transferring image representations learned in a CNN-based model trained on a source domain to be re-used in a model on a target domain. Finally, hybrid-based techniques transfer knowledge using both instances and shared parameters (Weiss et al., 2016). On the other hand, in heterogeneous TL, since the problem is the difference in feature spaces, only feature-based techniques are considered for heterogeneous TL (Day & Khoshgoftaar, 2017).

DA has gained a lot of attention as well. For example, Ganin and Lempitsky (2015) propose an approach that trains deep architectures on labeled data in the source domain and unlabeled data in the target domain. Their architecture uses a deep feature extractor that is connected to a deep label predictor for the source domain data and a deep domain predictor that performs the unsupervised DA on both source and target domain data. Therefore, the latter ensures the extraction of domain-invariant features. Universal training (i.e. shared encoders/feature extractors) can focus on either one domain or multiple domains—for example Nam, Lee, Park, Yoon, and Yoo (2021), Ouali, Hudelot, and Tami (2020), Kim and Kim (2020), Z. Wang et al. (2020), Kalluri, Varma, Chandraker, and Jawahar (2019), and Hoffman, Wang, Yu, and Darrell (2016). In addition, there are GAN-based approaches to DA such as Musto and Zinelli (2020) that are used for translating source domain images to target domain images using GAN—it should be noted that GAN-based translation approaches are widely used in remote-sensing literature.

The potential for leveraging the advantages of TL and DA in PU learning is huge. However, research in PU learning mostly focuses on the knowledge of single domain for learning a classifier; not enough work has focused on utilizing TL for PU learning (B. Liu et al., 2022). Among the few, for example, B. Liu et al. (2022) propose a mix of transfer and ensemble learning framework for PU learning. They take a parameter-based approach, in which SVM parameters and regularization terms are shared between source and target domains. Meng, Xie, and Sun (2021) take an instance-based approach, in which they use a PU learning model to identify examples that are close to target domain to be used later with the data in the target domain with a gradient boost decision tree model. As a limitation of their work, they show that negative transfer (i.e. when TL does not improve and even worsen the quality of learning) can occur in case of multi-source tasks. Mignone, Pio, D'Elia, and Ceci (2020) study a setting in which both source and target domains have PU data. They propose an approach in which, first, the unlabeled data are converted to weighted labeled data in each of source and target domains separately, and they then use a parametric-based approach to create a new classification model based on different models trained separately in each of the source and target domains. Bhat and Culotta (2017) use a feature-based approach for document classification, in which they reweigh the features' importances based on information extracted from a base document classifier on PU data. Hammoudeh and Lowd (2020) investigate the PU learning under covariate shift where the distribution of positive data shifts in the test data in comparison to the training data (as opposed to a fully-labeled source domain and PU target domain setting). Loghmani, Vincze, and Tommasi (2020) convert the open-set DA problem to a PU problem. In other words, instead of a source domain with label space Y_S and a target domain with label space Y_T such that $Y_S \subset Y_T$, they consider the source data as positive and the target data as unlabeled, which violates the SCAR assumption. Therefore, they use a mix of an autoencoder loss (as the domain agnostic discriminative loss of choice) and the NNPU loss by Kiryo et al. (2017) in order to avoid the unreliable predictions that occur in deep neural networks when NNPU with logarithmic loss is used with the unlabeled samples belonging to a different domain. Finally, Sonntag, Behrens, and Schmidt-Thieme (2022) investigate PU-DA where source domain is completely labeled and the target domain has PU data. They propose a two-step approach, in which reliable positive and negative pseudo-labels are selected in the target domain using a PU loss on target domain and a classic loss on the source domain. They then train a supervised classifier on the target domain.

- **Transfer Learning in Remote Sensing.** TL in remote sensing literature has gained a lot of attention. M. Zou and Zhong (2018) investigate a parameter-based TL in which model parameter values (i.e. weights) from a pre-trained model on an open image dataset such as ImageNet (Russakovsky et al., 2015) are used to for fine-tuning a model in the target domain (i.e. the domain of interest). They do this by either adjustmenting of all weights or reusing them, depending on whether the size of the training data in the target domain is large enough or not, respectively. A. X. Wang, Tran, Desai, Lobell, and Ermon (2018) also show the success of using a pre-trained model on images of one geographic location to be used and fine-tuned on images of another geographic location. X. He, Chen, and Ghamisi (2019) use a fully connected layer (as the start point of their model) in order to convert hyperspectral images with more than three channels into three channel data, which will then use pre-trained model weights on ImageNet for the rest of the model. Wurm, Stark, Zhu, Weigand, and Taubenböck (2019) take advantage of transferring pre-trained fully convolutional networks for mapping slums in various satellite images. Other examples of such TL approach are Najjar, Kaneko, and Miyanaga (2017) and Z. Zhou et al. (2021). Their success in using pre-trained model parameters is due to the similarities in low-level features of the two domains (M. Zou & Zhong, 2018). However, different studies such as Xie, Jean, Burke, Lobell, and Ermon (2016) have been able to get the same benefits using data rich proxy labels for cases in which the two learning tasks are not very related (Ghosh, Jia, & Kumar, 2021). All in all, although this type of parameter-based approach results in model improvement, as opposed to training a model from scratch, better results could be obtain from a feature-based approach or a hybrid of both.

The spectral shifts among the feature distributions of RS images of different geographic locations can cause the model trained on one geographic location to fail when used in a rather different geographic location. Therefore, in order to have a model that is robust to shift among the datasets, remote sensing turns to DA methods (Tuia et al., 2016). DA can be done in an unsupervised or semi-supervised approach: with unsupervised DA, the two unlabeled domains must match, while semi-supervised DA assumes that the source domain contains fully labeled data and the target domain contains a set of unlabeled data (Tuia et al., 2016).

Tasar, Happy, Tarabalka, and Alliez (2020b) address the domain shift between train and test datasets. In their approach, they use a generative adversarial style-transfer network to create fake images that have the style of test images. These fake images are used to fine-tune a model trained with the training data. Tasar, Tarabalka, Giros, Alliez, and Clerc (2020) address the multi-source DA. They use a generative adversarial style-transfer network to remove the marginal gap between each source domain and the target domain, which results in all the data having similar marginal distributions. Then, a model is trained on the transformed source data and used for prediction on the data in the target domain. There are many other style-transfer approaches such as Ji, Wang, and Luo (2020), Tasar, Giros, Tarabalka, Alliez, and Clerc (2020), L. Shi, Wang, Pan, and

Shi (2020), D. Zhao, Li, Yuan, and Shi (2021), Tasar, Happy, Tarabalka, and Alliez (2020a), Y. Zhang et al. (2019), and M.-Y. Liu, Breuel, and Kautz (2017). Finally, in their transformation-based DA, Chakraborty and Roy (2020) adopt a hierarchy of different approaches, including stacked auto-encoder neural network and fuzzy theory.

Lucas, Pelletier, Schmidt, Webb, and Petitjean (2021) address the semi-supervised DA in which some labeled samples are available in the target domain. Their proposed model is first trained on a source domain, and then trained on the target domain using a source-regularized loss function that shrinks the model weights estimations with respect to those learned on the source domain. To alleviate the fails in general unsupervised DA methods on RS imagery due to its difference with other datasets, such as urban scene/driving datasets, Iqbal and Ali (2020) propose a weakly-supervised DA for built-up region segmentation, in which they assume the availability of weak-labels that are image level labels in the target domain with pixel-level unlabeled images. They guide the adaptation in the segmentation model by the use of image classification for the weak-labels during the training process.

2.3 Ensemble Learning

Ensemble learning (EL) is a methodological framework, with the goal of establishing a better predictive performance by training and combining multiple learners each one of which, individually, may not produce as strong of predictive performance (Dong, Yu, Cao, Shi, & Ma, 2020b)—if the individual learning models are of same type, this approach is known as homogeneous EL, or heterogeneous EL otherwise (Ganaie, Hu, et al., 2021). The fundamental idea of EL is based on the voting mechanism implemented using different strategies including decreasing variance (bagging), decreasing bias (boosting), or ideally both (stacking) (Z.-H. Zhou, 2021). A primary tenet behind EL is to avoid inductive bias by allowing a better search in the hypothesis space or even extending it such that the generalization performance is improved by avoiding a single poor learner and/or a model hypothesis space that does not properly approximate the ground truth hypothesis (Z.-H. Zhou, 2021). Although EL is very successful in machine learning research, it faces new challenges with incorporation of deep neural networks due to their huge increase in training time and space (Y. Yang, Lv, & Chen, 2021). Efforts to address these challenges resulted in the emergence of deep EL that combines the advantages of both deep learning and EL (Ganaie et al., 2021).

Deep EL algorithms can be categorized into three main approaches: (i) extracting features from deep learning models and using them as input to other traditional classifiers, (ii) having an ensemble of the output of the individual learners that are deep learning models, and (iii) training end-to-end deep learning architectures that incorporate EL principles (Y. Yang et al., 2021). Some of these deep EL categories utilize the traditional EL approaches such as bagging, boosting, and stacking, and are mostly accompanied with a TL approach—for example, Doshi and Yilmaz (2020) propose an ensemble model for object detection task.

For the first category of deep EL algorithms, Korzh, Joaristi, and Serra (2018) propose a four-step deep transfer ensemble learning for classification task. They first fine-tune each of multiple CNNs with different architectures, and then they fine-tune an ensemble of these models, which include the initial weights identified from the previous step. Finally, they use the ensemble of the CNNs as a feature extractor for another classifier such as SVMs. Although, their approach could be considered in the second category of deep EL algorithms too, their approach has more overlap with the first category.

For the second category of deep EL algorithms, Kandaswamy, Silva, Alexandre, and Santos (2015) propose an approach to reduce the impact of layer selection in TL by an ensemble of multiple different-style transfer for generic features that are previously learned. Kumar, Kim, Lyndon, Fulham, and Feng (2016) propose the ensemble of CNNs, in which they first fine-tune the pre-trained CNNs on a large dataset of natural images, and then they apply the fine-tuned CNNs as either a feature extractor for a SVM classifier or a classifier for medical image classification. As mentioned, their approach investigates the first category of deep EL algorithms as well (i.e. CNNs as feature extractors for SVM classifiers).

Finally, for the third category of deep EL algorithms, Devassy and Antony (2021) propose a mix of transfer and ensemble learning, in which they use pre-trained models to create intermediate feature maps to be concatenated and fed to a set of dense layers with a binary classification output—the pre-trained weights and layers are freezen in the EL stage. Yu et al. (2021) propose a CNN-based EL for image dehazing, in which two subnetworks are used, one for capturing global image representation using TL, and the other subnetwork for capturing domain-specific representation. These two subnetworks create two feature-maps that are then concatenated and inserted into the rest of the main network for purpose of outputting dehazed images. Bousselham et al. (2021) propose an end-to-end deep EL for semantic segmentation that avoids the heavy cost of multi-stage training of ensembles of deep networks. In order to do this, they take advantage of multiple independent decoders, each of which gets trained on one of the multi-scale features set produced by a feature pyramid network approach.

Some other approaches in deep EL that do not fit into the aforementioned categorization are also worth mentioning. Nozza, Fersini, and Messina (2016) demonstrate the positive effect of EL through different ensemble methods such as simple-voting on reducing the cross-domain generalization error that occurs in domain adaptation. Nigam, Huang, and Ramanan (2018) investigate the possibility of an ensemble of models learned from dataset that differ in scene structure, viewpoints, and objects statistics in order to transfer knowledge from multi diverse domains to the target domain.

In terms of EL for PU learning, P. Yang, Humphrey, James, Yang, and Jothi (2016) propose an ensemble of PU learning models in order to alleviate the class imbalance within the dataset. They first, for each classifier, create a balanced training set by randomly subsampling from the unlabeled set. Then, they train SVM classifiers using a biased PU approach. They use a correction factor approach (Elkan & Noto, 2008) to reduce the prediction bias on the unlabeled data by considering all of them as negative. This bias is further reduced by an ensemble of each of these SVM models that are used to produce the final prediction. Nguyen, Li, and Ng (2012) apply an ensemble of classifiers for PU problem in time series classification. Claesen, De Smet, Suykens, and De Moor (2015) propose a bagging-base ensemble of a variation of SVM models for PU learning. In their approach: (i) the variability between different models is increased by resampling from both positive and unlabeled data, which results in more robustness against false positives, and (ii) the relative misclassification penalty between positive and unlabeled examples is controlled by introducing an extra degree of freedom in the model. Basile, Di Mauro, Esposito, Ferilli, and Vergari (2019) propose a general probabilistic generative approach to PU learning, in which the goal is to find reliable negative examples that are in the lowest density regions of the positive class distribution. They next create a mixture of generative models using the adoption of bagging ensemble from the discriminative framework. P. Yang, Li, Chua, Kwoh, and Ng (2014) propose a two-step PU learning approach in which, after identifying the reliable negative examples, both positive and reliable negative examples are used to assign weights to unlabeled data. The assigned weights represent the likelihood of the unlabeled data belonging to either positive or negative class. Finally, an ensemble of classifiers are used for the final prediction. Jowkar and Mansoori (2016) propose a two-step PU learning approach in which the second step consists of a graph-based ensemble of three weighted classifiers.

- **Ensemble Learning in Remote Sensing.** In remote sensing, researchers have demonstrated the robustness of ensemble classifiers over a single method strategy—for example, Benediktsson, Chanussot, and Fauvel (2007), X. Huang and Zhang (2012b), and Rahman, Smith, and Timms (2013). X. He and Chen (2020) take advantage of TL by reusing pre-trained models in order to alleviate limited train data in the target domain. Then, they use a classifier-level ensemble of multiple different fine-tuned models. Since their approach considers target domain of hyperspectral images, and the pre-trained models come from domains with three channels, they randomly select three channels of hyperspectral images for the fine-tuning process. In a different vein, some researchers have focused away from a random selection of channels from hyperspectral images, and have proposed different approaches that take advantage of all available channels within hyperspectral images. For example, X. He et al. (2019) propose a fully connected layer added to the beginning of the three-channel compatible networks' architectures to map the multi-channel hyperspectral images onto three channels input for such networks. Korzh et al. (2018) propose a four-step approach for an ensemble of pre-trained CNNs. First, they fine-tune each CNN separately; then, they construct an ensemble of these fine-tuned networks. Their approach results in improved EL performance in two ways: (i) using the ensemble of models, themselves, as either the final classifier or the feature extractor, then fed into another model as the final classifier, and (ii) fine-tuning the ensemble of models using joint loss function. Fan, Xu, and Zhang (2021) propose a classifier-level bagging-based ensemble for classifying house damage. Jamali et al. (2021) investigate the use of two different deep EL approaches for complex wetland classification: (i) majority voting classifier-level ensemble, and (ii) feature-level ensemble of multiple deep networks to be fed into a final classifier. They show that the latter results in improved predictive performance. Iyer, Sriram, and Lal (2021) utilize an ensemble of deep learning models, in which they calculate the mean of the probability scores of each model output per class, and, then make a prediction based on the maximum of the averaged probability values.

Plazas, Ramos-Pollán, and Martínez (2021) propose a classifier-level ensemble-based semi-supervised deep learning approach which iteratively labels those unlabeled data with high confidence predictions. Gu et al. (2022) propose an ensemble semi-supervised learning in which, first, supervised deep CNNs classifiers are trained on labelled data to be used as feature extractor. Then, in the second stage, unlabeled data are exposed to a self-learning process, which exploits pseudo-labelling continuously in order to improve the model. For self-learning, the fine-tuned models in the first step are used to produce different feature maps to increase the diversification among the learners in the ensemble module. The ensemble framework uses cross-checking mechanism among the classifiers such that their disagreement is minimized. This way, the ensemble components are trained together as opposed to traditional separate training strategy.

Although research has shown promising results of ensemble classifiers in remote sensing, little research has focused on the ensemble PU learning in remote sensing leaving a gap in this area of research. In one of the few examples, R. Liu et al. (2018) investigate EL for PU learning. Three ensemble methods (majority vote, weighted average, and weighted vote combination rules) are compared with different stand-alone one-class classification and PU learning methods. They show the superiority of weighted average and weighted vote ensembles over other models for classifying the urban areas from RS imagery. Wu, Qiu, Jia, and Liu (2020) investigate the bagging-based approach to PU learning using decision tree method for generating a landslide susceptibility map. Finally, X. Liu, Liu, Datta, Frey, and Koch (2020) propose a weighted voting ensemble of three one-class classification models for mapping invasive plants.

The literature reviewed in this chapter, covered PU, Transfer, and Ensemble learning paradigms. The goal of these paradigms of learning models is to alleviate the labeled data scarcity in many real-world applications, including vision-based applications. Although each of these methods are studied extensively, in general, and are introduced and used in remote sensing, in particular, the potentials of the combined advantages of these methods have not been researched much. Such research is still in its infancy with a few research demonstrating promising results in different leaning tasks. Therefore, the goal of this dissertation is to investigate such potentials for semantic segmentation task. More specifically, I investigate two of the possible avenues that are Transfer PU learning and Ensemble PU learning for semantic segmentation of RS imagery where labeled data scarcity is a burden on developing learning models.

CHAPTER III

DATASETS

Several non-remotely-sensed and remotely-sensed imageries and their derived positive and unlabeled variation datasets were used in this dissertation research. All of the datasets are online and freely available. In the sections below, I briefly describe each of these datasets: (i) the standard open image dataset ImageNet, (ii) Massachusetts Buildings Dataset, and (iii) Inria Aerial Image Buildings Dataset, in that order.

3.1 Standard Open Image Datasets

ImageNet, ILSVRC image dataset (Russakovsky et al., 2015), is used as the standard open image dataset for this dissertation. The ImageNet consists of 1281167 train images, 50000 validation images and 100000 test images for 1000 object classes—see Fig. 3 for sample images. Although this dataset is not directly used for training the models in this dissertation, it is indirectly used by incorporating the pre-generated parameters/weights of models pre-trained on them (as the starting point or warm start) for some of my models and baselines.

3.2 Massachusetts Buildings Dataset

The Massachusetts buildings dataset (Mnih, 2013) is a high resolution dataset with a spatial resolution of 1.0 meter. This dataset consists of a single source covering almost 340 km^2 in total of mostly urban and suburban areas of various sizes of buildings. The reference map consists of labels for each image that shows pixels that either do or don't belong to buildings—most of the labels are derived automatically with the test and validation sets being manually corrected for better performance assessments.

The dataset consists of 151 images of size 1500 × 1500 including 137 images for the train set, 4 images for the validation set and 10 images for the test set. Images are converted into patches of size 256 × 256—if not enough pixels are available at the edges of an images, enough pixels are created with the mirroring technique (*from OpenCV library*) of the patch's edges in order to create a 256 × 256 patch. This process results in 5436 total image patches including 4932 images for the train set, 144 images for validation set and 360 images for test set—see Fig. 4a for a sample image with some of its corresponding derived patches. Finally, in order to avoid the class imbalance problem (since it is out of the scope of this dissertation), a patch selection is performed based on the criterion that, within each patch, at least 25% of the pixels must belong to the class building. It should be noted that an upper bound for the positive class is not needed since class imbalance in favor of the class of interest is not problematic. Therefore, the final dataset consists of a total of 3559 image patches including 3201 images for the train set,

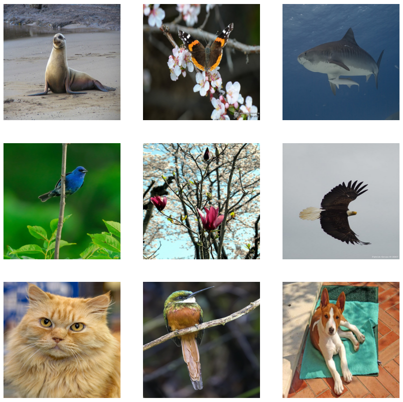


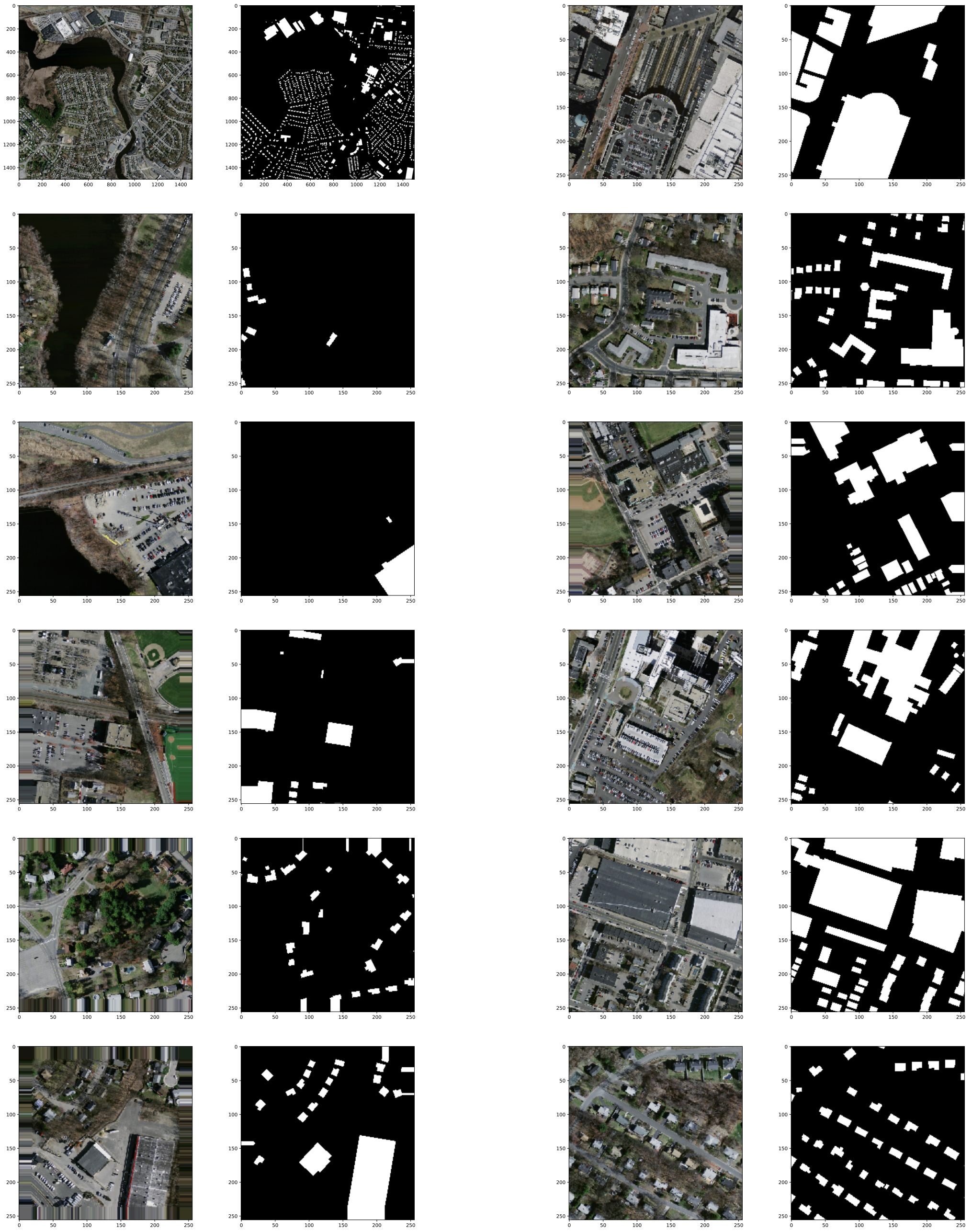
Figure 3. Sample images from ImageNet-ILSVRC dataset (Russakovsky et al., 2015).

83 images for validation set and 275 images for test set—see Fig. 4b for a sample of final patches. Table 1 represents a summary of the dataset characteristics before and after pre-processing. Finally, the dataset is standardized using equation 3.1, where $c \in C$ refers to a specific channel at the time and μ^c and σ^c are the mean and the standard deviation for that specific channel calculated over the entire train set.

$$\hat{x}^c = \frac{x^c - \mu^c}{\sigma^c} \tag{3.1}$$

Table 1. Massachusetts Buildings Dataset Characteristics

Dataset	Total Images	Training Images	Validation Images	Testing Images	Image Size	Resolution (m/p)
Original	151	137	4	10	1500 × 1500	1.0
Processed	3559	3201	83	275	256 × 256	1.0



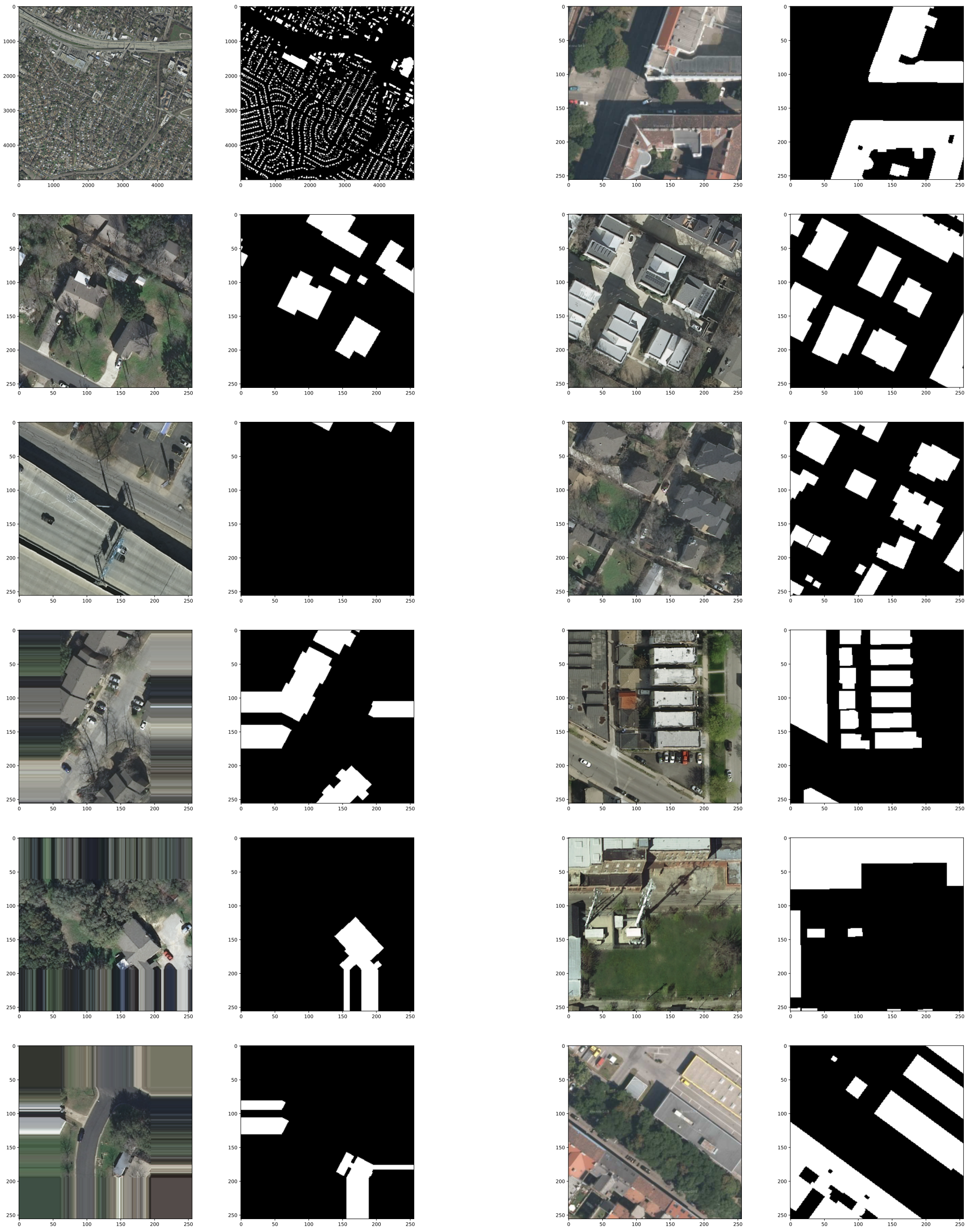
(a) (b)

Figure 4. (a) Left and right columns show images and their corresponding labels, respectively. From top to bottom: a sample from Massachusetts buildings dataset (Mnih, 2013) and samples from its derived patches, one from the middle, and the rest from top-left, top-right, bottom-left, and bottom-right corners, respectively, showing the mirroring effect when not enough pixels are available for patch creation. (b) Selected final patches: images and their corresponding labels in the left and right columns.

3.3 Inria Aerial Image Dataset

The Inria Aerial Image Labeling Dataset (Maggiori et al., 2017) consists of public domain aerial orthorectified RGB-color imagery with a spatial resolution of 0.3 meter. The images cover dissimilar urban settlements for different geographic locations such as cities and towns with different building densities. The dataset covers a total of $810\,km^2$, $405\,km^2$ in each of train and test sets. Due to unavailability of reference maps for the test set, the available original train set is used to create my train, validation, and test sets with proportions of 70%, 15%, and 15% of the original train set for this dissertation. The reference map consists of public domain official building footprints with labels for each pixel either belonging or not belonging to buldings.

The dataset consists of 180 images of size 5000×5000. As with the Massachusetts buildings dataset, images are converted into patches of size 256×256 with the same policy that if not enough pixels are available at the edges of an images, enough pixels are created with mirroring technique in order to have a 256×256 patch. This process results in 72000 total image patches of size 256×256. Then, again, in order to avoid the class imbalance problem, a patch selection is performed based on the criterion that, within each patch, at least 25% of the pixels must belong to the class building. Therefore, the final dataset consists of a total of 18702 image patches. Thus, with the 70%, 15%, and 15% portion strategy, the train set, validation set, and test set have respectively 13092, 2805, and 2805 images. See Table 2 and Fig. 5 for, respectively, a summary of the dataset characteristics (before and after pre-processing) and a sample image with some of its corresponding derived patches (before and after pre-processing). Finally, the dataset is standardized in the same manner as for Massachusetts buildings dataset.



(a) (b)

Figure 5. (a) Left and right columns show images and their corresponding labels, respectively. From top to bottom: a sample from Inria Aerial Image Dataset (Maggiori et al., 2017) and samples from its derived patches, one from the middle, and the rest from top-left, top-right, bottom-left, and bottom-right corners, respectively, showing the mirroring effect when not enough pixels are available for patch creation. (b) Selected final patches: images and their corresponding labels in the left and right columns.

Table 2. Inria Aerial Image Dataset Characteristics

Dataset	Total Images	Training Images	Validation Images	Testing Images	Image Size	Resolution (m/p)
Original	180	–	–	–	5000 × 5000	0.3
Processed	18702	13092	2805	2805	256 × 256	0.3

3.3.1 Positive and Unlabeled Dataset. A bottleneck for developing models on domains with limited or no labeled data (such as positive and unlabeled (PU) learning problem) is the validation of the model itself since there are no fully-labeled samples available for such validation (Tuia et al., 2016). Although research such as W. Li and Guo (2013) has been trying to develop performance metrics on naturally PU data for PU model evaluation and assessment, commonly acceptable and used performance metrics are still those developed for positive and negative (PN) data. Therefore, for this dissertation, I create PU train data on top of the PN train data from Inria Aerial Image dataset. In doing so, PU data are used for model development, and PN data are used for model evaluation and assessment.

I use the pre-processed data that I created above from Inria Aerial Image dataset to create the PU data. I choose these data owing to the fact that the data do not suffer from a class-imbalance problem, which could lead to generating suboptimal classifiers due to the heavy influence of the majority class in comparison to the influence of minority class on model training (Chawla, Japkowicz, & Kotcz, 2004). Although there is research such as X. Chen et al. (2021) that investigates the class-imbalance problem in PU learning, it is out of the scope of this dissertation, and, thus, the PU dataset is created in a way that avoids such problem (as shown in Fig. 6). The process for creating the PU data involves applying a moving window of the size of 32 × 32 and stride 32 that scans the label maps corresponding to each image within the train set. At each location of the moving window, the pixels that belong to the positive class are changed randomly to unlabeled, with some probability (which, in this case, is 0.5). Therefore, since there is complete randomness and no selection bias, it is safe to say that the created data follow the Selected Completely At Random (SCAR) assumption. It could be argued that the chosen size for the moving window may introduce a bias, and, thus, the SCAR assumption may not be met; however, the counter argument is that the moving window's size is random and chosen independently from the positive class characteristics, arguably still fulfilling the SCAR assumption. Fig. 7 shows some samples of PU train data resulting from the PU data making process.

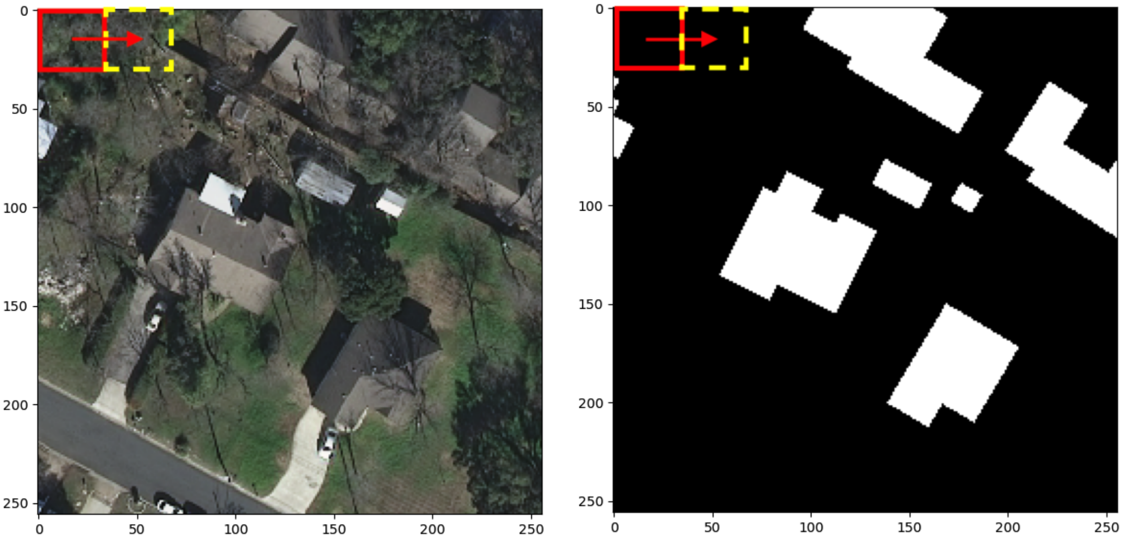


Figure 6. PU data is created using a moving window over labels corresponding to each image within the train set. At each location of the moving window pixels that belong to the positive class are changed to unlabeled with complete randomness (i.e. no selection bias).

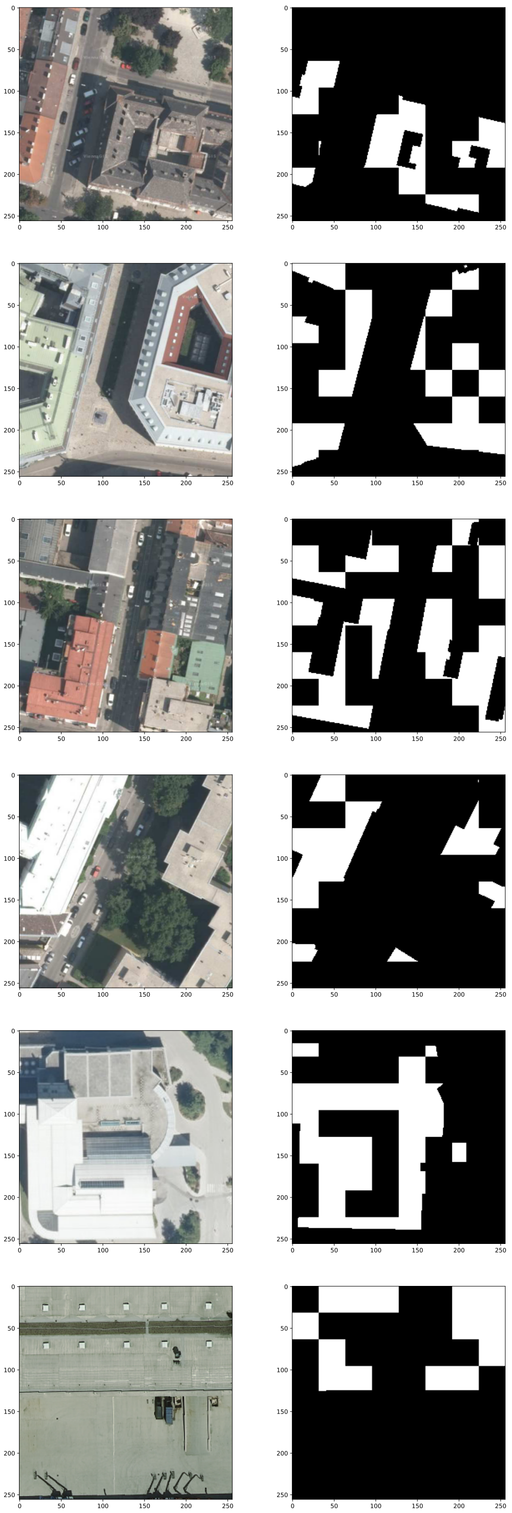


Figure 7. Left and right columns show images and their corresponding PU labels created by the moving window strategy.

Next, I repeat the aforementioned procedure in order to create different PU train datasets with different levels of missing labels from the positive class. Therefore, the probability with which the pixels of positive class that are under the moving window are converted to unlabeled pixels is changed to 0.6, 0.7, 0.8, and 0.9 for a total of four additional cases. Thus, there are five totally different PU train dataset from Inria Aerial Image dataset that are named PU5, PU6, PU7, PU8, and PU9, all of which satisfy the SCAR assumption. See Fig. 8. In other words, PU9 means that the probability that a pixel that belongs to the positive class is labeled is $1 - 0.9 = 0.1 = p(s = 1|y = 1)$.

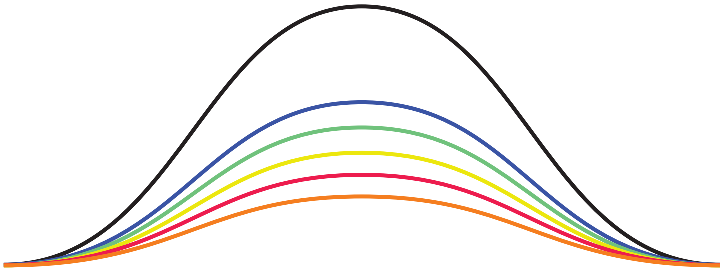


Figure 8. Different PU train datasets created from Inria Aerial Image dataset with different missing probabilities for the positive class, all of which satisfying the SCAR assumption—graphs in black, blue, green, yellow, red, and orange show the distribution of the positive class and of the positive data in PU5, PU6, PU7, PU8, and PU9 datasets, respectively.

3.3.1.1 Homogeneous Case. As mentioned in the previous chapter, the spectral shifts among the feature distributions of remotely-sensed images of different geographic locations can cause the model trained on one geographic location to fail when used in a different geographic location (Tuia et al., 2016). This failure can be considered as the heterogeneous case of transfer learning/domain adaptation. However, if the training and testing (or future inference) happens at the same geographic location, then the homogeneous transfer learning/domain adaptation is the case. Since this dissertation investigates both homogeneous and heterogeneous transfer learning/domain adaptation and the PU dataset created so far is for heterogeneous case, here, I explain the dataset that is created for homogeneous case. The Inria Aerial Image dataset includes different geographic locations (i.e. cities). So, I selected image patches and their corresponding label maps for one of the cities—in this case Chicago with 6855 patches. Next, I randomly selected a subset of the patches containing half of the patches with PN labels with the 85% and 15% portion strategy for train and validation sets. Then, the other half is used for creating PU data with the 70%, 15%, and 15% portion strategy for train, validation, and test sets. This results in 2914 and 514 images for train and validation sets in the PN data, and 2399, 514, and 514 images for train, validation, and test sets in the PU data, respectively. Like the previous case, this selection process resulted in PU5, PU6, PU7, PU8, and PU9 datasets, each of which represent a different level of missing labels from the positive class.

All of the datasets described above will be used in the following two chapters in which I develop learning models that address the research questions of this dissertation.

CHAPTER IV

DEEP TRANSFER POSITIVE AND UNLABELED LEARNING

The large quantities of frequently-acquired multi-temporal and multi-source remotely-sensed (RS) imageries provide great opportunities for real-time earth monitoring (Tuia et al., 2016). However, such monitoring requires automatic image processing for which supervised learning models have shown success, though with the downside of relying on availability of fully-labeled data for model training. In addition, there is no guarantee that models trained on images from one geographic location will perform well on images from another geographic location due to distribution shifts within the RS imageries between the two geographic locations that arise from differences in acquisition techniques, atmospheric conditions, objects of interest, and so forth (Tuia et al., 2016). There are two lines of research that address these issues: (i) transfer learning (TL) and (ii) positive and unlabeled (PU) learning. TL and one of its specific variations, domain adaptation (DA), leverage the benefits of multi-temporal and multi-source aspects of RS imageries. On the other hand, the goal in PU learning is to alleviate the labeled data scarcity, especially when creating a fully-labeled data may not be fundamentally justified due to the interest in only one specific landcover class or object (W. Li et al., 2010). Both TL/DA (D. Zhao et al., 2021) and PU learning (Deng et al., 2018) have shown promising results in semantic segmentation of RS imageries and are of great importance for (close-to) real-time predictions on newly acquired RS imageries. However, researchers have not yet attempted to combine these two approaches, especially in the context of RS imageries. Therefore, in this chapter, I investigate the possibility of hybrid methodologies that benefit from both TL/DA and PU learning. Specifically, I start with the simpler scenario that is homogeneous

TL/DA and PU learning. Then, I discuss the advantages and disadvantages of such scenario. Next, I relax the homogeneous assumption for the proposed model such that it can handle the heterogeneous case. Finally, I discuss the limitations and opportunities for further improvements.

4.1 Background

Deep learning models have been successful in many computer vision tasks, including one of the most important of them, semantic segmentation. Such success is mainly due to the supervised learning approach that depends on collecting large amount of dense pixel-wise labels (Mittal, Tatarchenko, & Brox, 2019). However, the high cost of collecting labeled data has resulted in researchers shifting focus to other promising areas, such as semi-supervised learning, weakly-supervised learning, and unsupervised and semi-supervised domain adaptation.

4.1.1 Semi-supervised Learning. Semi-supervised learning (SSL) focuses on target domain (i.e. the domain of interest) and tries to alleviate the labeled data scarcity by learning a generalized model applied to a large amount of unlabeled data and a few limited labeled data. French, Aila, Laine, Mackiewicz, and Finlayson (2019) investigate the consistency regularization technique for semi-supervised semantic segmentation. The idea behind consistency regularization is that the trained model should generate consistent predictions for inputs of the same unlabeled image under different perturbations. They show that incorporating augmentation techniques such as CutOut (DeVries & Taylor, 2017) and CutMix (Yun et al., 2019) (with CutMix showing the superior results) improve the performance of semi-supervised semantic segmentation. The learning approach is further improved by French and Mackiewicz (2021) who incorporate color augmentation. Finally, Olsson, Tranheden, Pinto, and Svensson (2021) propose the ClassMix augmentation technique in which two unlabeled images are used to create a new augmented image based on superimposing part of one image on to the other image with the help of object boundaries mask extracted from network's prediction segmentations. In addition, the corresponding prediction segmentations for each of the two input unlabeled images are also augmented to create the corresponding label map for the augmented output image.

Souly, Spampinato, and Shah (2017) use generative adversarial networks (GANs) for SSL such that the generator creates additional training data while the discriminator is a segmentation network with an additional class to count for the *fake* class, which is produced by the generator. The three-part loss function takes into account a standard generative adversarial network GAN loss for the image produced by the generator, a loss for unlabeled data, and a standard cross-entropy loss for the labeled data. Hung, Tsai, Liou, Lin, and Yang (2018) propose an adversarial network-based semi-supervised semantic segmentation. Their method consists of a segmentation network and an adversarial discriminator in which they alternate the typical GAN discriminators to a fully convolutional network (FCN)-based discriminator for distinguishing the ground truth segmentation distribution from the predicted probability maps by the segmentation network. The discriminator is the key for the semi-supervised setting since it provides predicted labels as pseudo labels for unlabeled data to be used for the training of the segmentation network. Similarly, Mittal et al. (2019) propose a model in which the generator is the segmentation network and the discriminator distinguishes the ground truth segmentation maps from the generated ones. However, in addition to the GAN branch, they add a multi-label classifier as the second branch based on a modified Mean Teacher (Tarvainen & Valpola, 2017), which is responsible for filtering the segmentation network's false positive predictions. Alonso, Sabater, Ferstl, Montesano, and Murillo (2021) propose a teacher-student model with a memory bank that contains relevant and high-quality feature vectors from labeled data. The teacher network creates pseudo-labels for unlabeled data to be used along with labeled data information from labeled data for training the student network. The key element in their approach is the memory bank and its corresponding contrastive learning loss that enforces the output features from the student network to be similar to the ones from the teacher network's memory bank. Ke, Qiu, Li, Yan, and Lau (2020) propose a general framework for pixel-wise SSL tasks that can be applied to a wide range of pixel-wise tasks without structural adaptation. Their model utilizes two (segmentation) models with different initializations to form perturbations between them, and a flaw detector network that checks the differences among models' outputs and ground-truth where in the corresponding losses, for unlabeled data, the terms for such differences are assumed to be zero.

Though, while SSL has been successful, it still requires some fully-labeled images to capture the complete semantics within an image of a complete scene or object. RS imageries are large; fully labeling one encumbers the same labeling burdens. And small patches of such images may not contain complete semantics of objects/classes. In addition, the negative class, in case of RS imageries, is too diverse such that being interested in only one class for prediction makes labeling far more undesirable. Finally, and most importantly, SSL does not address the more challenging problem of distribution shifts among RS images, a problem that makes deploying the trained models for inference much more difficult.

4.1.2 Weakly-supervised Learning.

Weakly-supervised approaches assume availability of zero labeled data and take into account auxiliary information instead, one of the most studied ones being the image-level class labels (Ahn & Kwak, 2018; Z. Huang, Wang, Wang, Liu, & Wang, 2018; J. Lee, Kim, Lee, Lee, & Yoon, 2019). For example, Ahn and Kwak (2018) assume the availability of image-level class labels. First, they generate pixel-level segmentation labels for training images given their image-level class labels. Then, they train a semantic segmentation network using the generated segmentation labels. Other non-image-level-based weakly-supervised methods are also proposed. For example, Papandreou, Chen, Murphy, and Yuille (2015) adopt an approach in which either bounding boxes of semantic object in an image or image-level labels are available. D. Lin, Dai, Jia, He, and Sun (2016) take advantage of scribble-based image labels that are easier and faster to generate. They propose a two-part model in which a graphical model creates fully labeled images by exploring the unmarked pixels using the propagated information from scribbles, and a FCN that is trained using the generated labels from the graphical model and provides semantic prediction for the graphical model.

While researchers have achieved some success, there are two problems with weakly-supervised approaches. First, the need for auxiliary information is too vague for the diverse information represented in RS images (or even their image patches). In addition, such auxiliary information is still too difficult to generate since RS imageries cover a large geographic location. Creating auxiliary information requires going through so many different small image patches which, although it does not represent as much work as creating dense pixel-level information, still takes considerable time and energy. Second, although weakly-supervised approaches incorporate weakly labeled examples in addition to pixel-level labels, they do “not exploit the unlabeled data to extract additional training signal” (Ouali et al., 2020, p. 2).

4.1.3 Semi-supervised Domain Adaptation. Semi-supervised domain adaptation (SSDA) tries to reduce the data distribution shift between source and target domains using fully labeled source data and partially limited labeled target data. SSDA is mainly focused on image classification with different approaches, including consistency training of different forms such as entropy minimization (Berthelot et al., 2019), pseudo-labeling (Sohn et al., 2020), divergence-based consistency (Gong, Wang, & Liu, 2021), and adaptive consistency (Abuduweili, Li, Shi, Xu, & Dou, 2021). However, semantic segmentation task is more challenging and it requires methods originally developed for it rather than adjusting image classification methods for it (S. Chen, Jia, He, Shi, & Liu, 2021).

There are approaches that use Image-to-Image translation techniques in order to reduce the feature discrepancy between source and target images. For example, Musto and Zinelli (2020) propose a mixture of pixel-level and feature-level domain alignments in a generative adversarial (e.g. GAN) framework for translating source images to target style. These kind of methods either are done stand alone before the segmentation process or are co-trained with the segmentation network. For the former, it is an extra step and burden for fast training, and for the latter, it adds extra computation burdens. However, there is research that takes advantage of adversarial learning in a different way. For example, Z. Wang et al. (2020) propose an adversarial approach with shared feature generator strategy. In addition to the cross-entropy loss for the labeled data, they introduce two adversarial losses for cross domain feature alignment: (i) a global adaptation using a GAN-based discriminator trying to distinguish the input feature maps of the source and target domains, and (ii) a semantic-level adaptation on both source and target data. S. Chen et al. (2021) denotes that such approach is unstable in training due to co-occurrence of adversarial training and weak supervision. They also claim that the full potential of labeled data from the two domains are not utilized; therefore they propose region-level and sample-level data mixing methods applied to source and target labeled data to reduce the data distribution gap. Two complementary teacher models are considered for training on the mixed data, where each of them is fed with data from one of the mixing methods. Then, an ensemble of these two pre-trained domain-mixed teachers is used as the teacher network for the student network in the knowledge distillation pipeline that includes cross-entropy loss for target labeled data and Kullback–Leibler (KL)-divergence loss for unlabeled data and its corresponding pseudo labels from the mixed teacher network. Finally, they use a self-training component to train the teachers for further improvement using the generated pseudo labels by the student network.

Finally, the idea of shared networks, specifically sharing feature generators (i.e. encoders) has been receiving attention due to their potential of decreasing the deployment costs in favor of real-time inference. In addition, Kalluri et al. (2019) propose a universal segmentation framework that shares an encoder among different domains, where each has its own decoder; thus, their model can handle label space discrepancy. In addition, to addressing feature alignment among domains, they propose a pixel-level entropy regularization as a separate module for propagating information among labeled and unlabeled images within all domains. Ouali et al. (2020) also propose an approach based on shared encoders between labeled and unlabeled images within a domain and among domains. They claim that, for semantic segmentation, it is easier to capture the low density regions in the hidden representations rather than the input images, and, thus, they perform a cross-consistency training technique based on the resiliency of the model (i.e. the shared encoder and main decoder) to the perturbations on the generated features of unlabeled images by the shared encoder. Auxiliary decoders are used to help propagate such resiliency of the shared encoder and main decoder using an unsupervised loss calculated on the output of the main and auxiliary decoders. The shared encoder and main decoder are further trained using the labeled images. They also show that this SSL technique can be expanded to SSDA by an alternate-training approach among domains using domain-specific main and auxiliary decoders while the shared encoder remains unique across domains.

Although, in comparison to SSL, SSDA addresses the problem of distribution shifts that happens often among RS imageries, it still requires some fully labeled images that may not be easily available in case of RS imageries.

4.1.4 Unsupervised Domain Adaptation.

Unsupervised domain adaptation (UDA) tries to reduce the data distribution shift between source and target domains using fully labeled source data and unlabeled target data (as opposed to partial labels considered in SSDA). The approaches in UDA can be broadly categorized into (i) adversarial learning by either transferring the style of labeled source data to target domain (Hoffman et al., 2018) or adding a discriminator network to a segmentation network for improving the segmentation network (Luo, Zheng, Guan, Yu, & Yang, 2019), (ii) self-training (i.e. pseudo labeling) (Y. Zou, Yu, Kumar, & Wang, 2018), (iii) consistency training (Melas-Kyriazi & Manrai, 2021), and (iv) a mix of some or all of them (Y.-C. Chen, Lin, Yang, & Huang, 2019; Y. Li, Yuan, & Vasconcelos, 2019), which is usually the case in recent models. For example, Vu, Jain, Bucher, Cord, and Pérez (2019) address the gap between data distributions using two entropy minimization approaches on pixel-wise predictions: (i) a direct MinEnt entropy minimization and (ii) an adversarial AdvEnt entropy minimization. Pan, Shin, Rameau, Lee, and Kweon (2020) adopt the adversarial AdvEnt entropy minimization approach to address the inter-domain adaptation, while they propose an intra-domain adaptation using image-level mean predicted entropy maps for target domain images to classify them into easy and hard subdomains. Then, they use the pseudo labels from the easy subdomain to adapt the segmentation network from the easy to hard subdomain. However, to capture the domain gaps among predicted labels within pixels, Yan et al. (2021) propose a pixel-level intra-domain adaptation as an alternative to image-level intra-domain adaptation. Saito, Watanabe, Ushiku, and Harada (2018) propose an adversarial approach in which the best discriminative features are learned through an alternative-training of a shared generator and two classifiers. Truong et al. (2021) propose a generalized form of the Adversarial Entropy Minimization, Bijective Maximum Likelihood, that does not take into account pixel independence and incorporates a bijective mapping network (learned using the label set of the source domain) for calculating the loss on unlabeled target data.

Research in UDA has achieved notable success in semantic image segmentation. However, “the domain gap cannot be fully alleviated due to the lack of strong supervision in the target domain” (S. Chen et al., 2021, p. 2). Also, the successful adversarial approaches are computationally expensive, are not stable during training time, and are challenging with respect to hyperparameters (Melas-Kyriazi & Manrai, 2021), which makes training the models challenging.

4.1.5 The motivation behind this work.

Although collecting dense pixel-wise fully-labeled images is not preferable in RS research, it is justifiable to collect some limited labeled pixels for the landcover class of interest. Such partially labeled pixels within images can be generated with quick eyeballing and skimming of RS imageries in comparison to the difficult process of creating dense pixel-wise labels either through field investigation or visual interpretation. Therefore, with such labeled data situation, the aforementioned techniques are not quite applicable. However, they provide a reasonable foundation that can be assembled along with ideas from PU research to provide a seamless PU method that takes advantage of RS imageries across domains. Considering this, I start with the naive homogeneous assumption for model development and show that such assumption is practically hard to meet. Therefore, I move on to the heterogeneous case and its practicality in RS imageries. The rest of this chapter is organized into multiple sections, each of which covers both homogeneous and heterogeneous cases. In § 4.2, I discuss the baselines, against which I evaluate my model. Next, I present the proposed approach for PU domain adaptation in § 4.3. In § 4.4, I examine the results, and, finally, I summarize and discuss future work in § 4.5.

4.2 Baselines

There is a dearth of research on positive and unlabeled domain adaptation (PUDA), making it hard to consider good baselines. A few different studies Lei et al. (2021); Loghmani et al. (2020); Sonntag et al. (2022) implicitly or explicitly address the problem of PUDA. However, none of these research papers are suitable to be considered for a baseline here, and I explain why in the following page.

Lei et al. (2021) do not explicitly target PUDA. However, they consider multi-modality (including different times and locations) in the RS imageries they use. With the way of defining train and test datasets, one can consider this approach as PUDA. Still, there are some major problems with their approach that make it hard to use it as a baseline. Their model consists of a feature extractor and fully connected classification heads. Therefore, it introduces high computation and time cost because of the point-wise predictions. The positive (and negative) labeled data are selected very carefully using field investigation and interpretation, which negates the promises of PU learning for removing heavy costs of labeling the data. In addition, such cherry-picked labeled data makes it hard to leverage more images for training, which can be seen based on the very limited number of pixels that are used in their experiments. Next, they select a limited number of unlabeled data, and my interpretation is that they do so to avoid the class imbalance problem. However, in such way, they do not take advantage of all of the available unlabeled data in the images. Finally, their novelty seems to come from their use of NNPU loss (Kiryo et al., 2017). Therefore, I define one of my baselines to be a more general case of Lei et al. (2021) that addresses such issues while using the NNPU loss as they did.

Loghmani et al. (2020) formulate the PUDA such that the source data are considered as positive and the target data as unlabeled. Therefore, they completely ignore the available labels in target domain, which makes it UDA-like. However, their approach is open set DA, which is different from (closed set) UDA. Therefore, since the open set DA assumption that they consider is not applicable in the cases considered in this dissertation, instead of their approach, I consider a state-of-the-art UDA as a baseline.

Sonntag et al. (2022) address PUDA by integrating UDA and PU learning for image classification. Although this research is the closest to what is done in this chapter, there are still problems, which prevent it from being considered as a baseline. First, they assume a same feature space for source and target domains. While this does not hold the method for the homogeneous case, it is definitely problematic for the heterogeneous case. What is more important is that their approach is geared towards image classification. The major element in their model is the candidate set that is constituted during learning phase in step-1 that identifies reliable positive and negative examples for a supervised learning in step-2. The creation of the candidate set may be feasible for image classification with tens' of thousands images at most (like the size of the datasets that they used). However, the creation of such a candidate set is not practically possible for pixels in semantic segmentation. Even if it were possible, the step-2 of the approach will be similar to the method used by Lei et al. (2021), which introduces the difficulties that were mentioned before. For example, the positive and negative pixels will be distributed sparsely in a salt-and-pepper manner preventing the supervised model in step-2 to capture global and local relationships for semantic segmentation. Therefore, following the same approach as Sonntag et al. (2022), I consider PU learning and UDA, with different settings that will be discussed later, as baselines for my proposed model.

The above research is not suitable to be considered as the baselines due to the mentioned reasons. Next, I discuss the baselines that I use for my proposed model.

4.2.1 Homogeneous case.

I consider three different baselines for this setting: (i) PU learning on target-only data with and without transferring pre-trained weights from ImageNet, (ii) training a model in a supervised manner using fully-labeled source data, and, then, fine-tuning it in a PU learning manner on PU target data. The supervised section is considered with and without transferring pre-trained weights from ImageNet, and (iii) the model that I propose when there is no PU loss meaning that target data is considered as unlabeled. The reason for the last baseline is to show that power of the proposed model in comparison to the other two baselines, and that how PU data in target domain can help improve the model's performance. Showing the potentials of the proposed model and incapability of the PU baseline for the heterogeneous case, I use a UDA model as the baseline in the heterogeneous case as it is explained below.

4.2.2 Heterogeneous case. I consider one of the state-of-the-art models in UDA, PixMatch (Melas-Kyriazi & Manrai, 2021). The reason behind choosing PixMatch is its great performance as well as a comprehensive performance comparison that its authors did with all the competitive and state-of-the-art models at the time of the PixMatch paper. There are several perturbation settings, such as Fourier and CutMix, in PixMatch model, which are considered as different baselines. In addition, I adopt the network architecture and backbone from the PixMatch model for the proposed model in order to have a further and more fair comparison of the two models. Finally, it should be mentioned that only a UDA model as the baseline is sufficient since the proposed model is already compared against the NNPU-based baselines in the homogenous setting.

4.3 Methodologies

Source and target domains provide two different sets of images. The source domain $\mathcal{D}_S = \{(x_s, y_s)\}_{s\in S}$ contains n_S images of form x_s with their corresponding fully-labeled ground truth semantic maps of form y_s. The target domain $\mathcal{D}_T = \{(x_t, y_t)\}_{t\in T}$ contains n_T images of form x_t with their corresponding

PU semantic maps of form $y_t = y_{t_p} \cup y_{t_u}$ where y_{t_u} contains pixels that come from both positive and negative classes. The ratio of labeled-positive to unlabeled data is controlled in 5 different datasets PU5-9 as described in Chapter III. In all settings, the models are trained on train and validation sets and evaluated on the test set. In the following pages, I present the proposed PUDA approach starting with the homogeneous case, and then moving on to the heterogeneous case.

4.3.1 Homogeneous case.

The assumption here is that if the two source and target domains contain images from the same geographic location and acquired by the same sensors, then their feature space should be the same, but with different feature distributions due to differences in acquisition times and atmospheric effects. At the core of the proposed approach is the shared feature generator (i.e. encoder) idea that is used by researchers such as Kalluri et al. (2019), Ouali et al. (2020), and Z. Wang et al. (2020). As shown in Fig. 9 and equation 4.1, the model consists of a supervised learning module, a self-training learning module, and a PU learning module. For the proposed model, UNET is chosen as the semantic segmentation architecture with VGG16 as its encoder backbone. In the following section, I start with reviewing the UNET architecture and the VGG16 encoder backbone and the proposed model architecture; then, I discuss each of the aforementioned learning modules. Finally, I mention the performance metrics used for models' assessment and evaluation: accuracy, IoU, and F1-score.

$$\begin{aligned}\mathcal{L}_{homogeneous_S} &= \mathcal{L}_{supervised} \\ \mathcal{L}_{homogeneous_T} &= \lambda_{pu}\mathcal{L}_{T_{pu}} + \lambda_{pseudo}\mathcal{L}_{T_{pseudo}}\end{aligned} \tag{4.1}$$

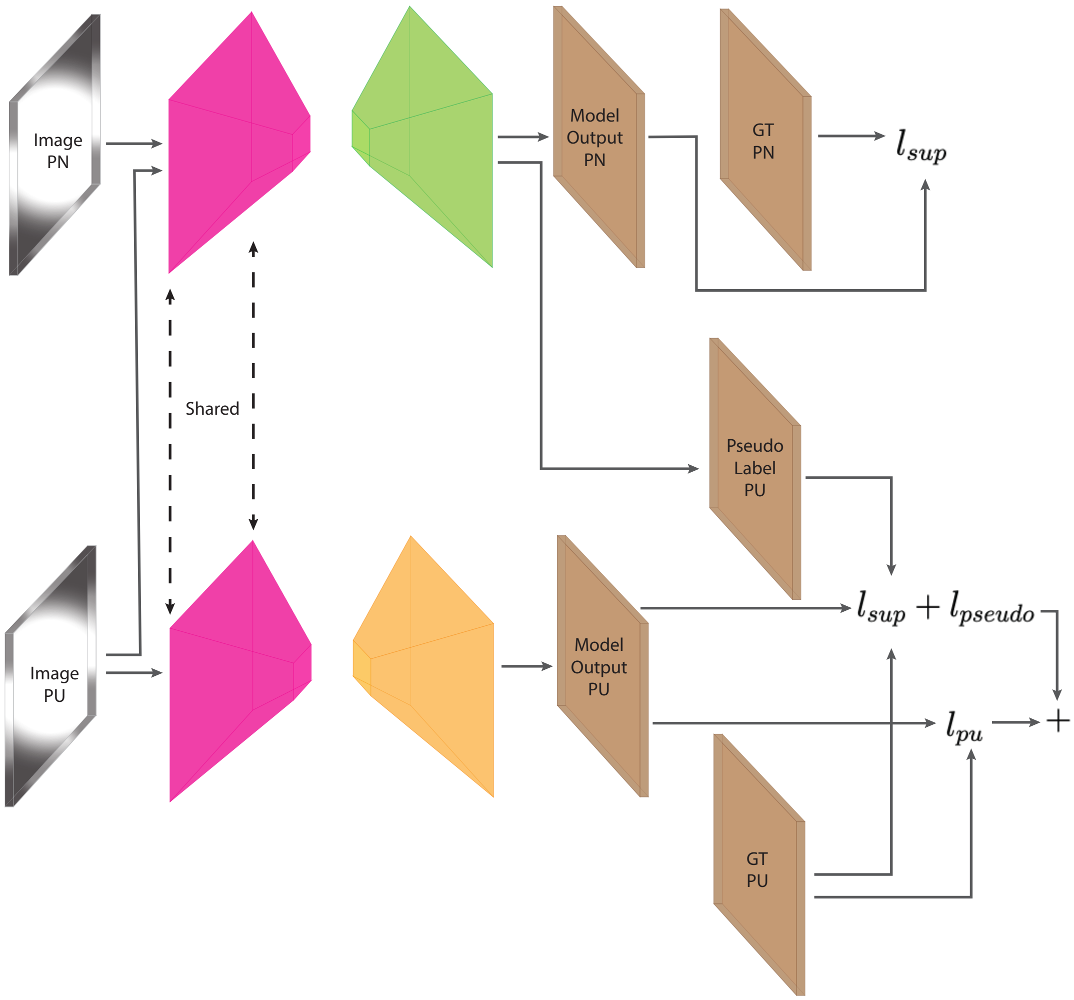


Figure 9. Homogeneous transfer positive and unlabeled learning architecture.

4.3.1.1 Model architecture: UNET. Unet, by Ronneberger, Fischer, and Brox (2015b), is a segmentation network that was originally developed for Biomedical images. The Unet architecture consists of two modules, a contracting module and an expansive module. The contracting module (i.e. encoder or feature extractor) consists of a stack of convolutional, batch-normalization, activation, and max-pooling layers. The contracting module captures the context of the input image and outputs a feature map. The expansive module is somewhat symmetric to the contracting module, which replaces max-pooling layers with transposed-convolution layers as the upsampling operators in order to reconstruct an output with the same size as the input image. Between each two upsampling operations a high-resolution intermediary feature map from the contracting path is attached to the current upsampled layer for localization and propagating context information, and then is fed to a convolution layer before getting passed to the next upsampling operator. Since the network is based purely on convolutional layers and does not have any fully connected layer, it can accept input images of any size (e.g. 256×256 in case of this dissertation). The batch-normalization layer is not part of the original network; rather, it is added later in order to allow the network to capture the data distribution for better convergence. Fig. 10 shows the architecture of the Unet model: (i) the contracting module consists of the repeated application of a 3×3 convolution layer with padding and stride 1, a batch-normalization layer, and a rectified linear unit (ReLU) activation layer two or three times, followed by a 2×2 max pooling operation with stride 2 for downsampling, and (ii) the expansive module consists of repeating steps of a 4×4 transposed-convolution for feature upsampling, a ReLU activation layer, a concatenation of the corresponding feature map from the contracting path, a 3×3 convolution layer to condense features, and another ReLU activation layer. Finally, a 1×1 convolution layer is applied to the last upsampled layer in order to map the feature layer to the desired number of classes.

4.3.1.2 Model backbone: VGG16. VGG16, by Simonyan and Zisserman (2014), is a Convolutional Neural Network (CNN) Architecture developed for image classification. Since semantic segmentation networks are evolved from image classification networks, they use one of the many available image classification networks as the backbone for their feature generators. There are either two or three 3 × 3 convolutional layers followed by a max-pooling layer. Such a stack of layers is applied repeatedly, which results in a reduction in the number of hyper-parameters. Finally, two Fully-Connected (FC) layers are applied, followed by a softmax layer that generates class probabilities for the output. When considered for semantic segmentation networks, the FC and softmax layers are factored-out and the remaining of the network is used for as the feature generator. As mentioned in the Unet section, I use a modified version of the original VGG16 architecture, in which there are batch-normalization layers added to the network.

4.3.1.3 Learning module: supervised. The supervised loss function is the standard cross-entropy loss on the source domain. As shown in equation 4.2, for the source input image x_s: $p_s^{h,w}$ is the output probability distribution at

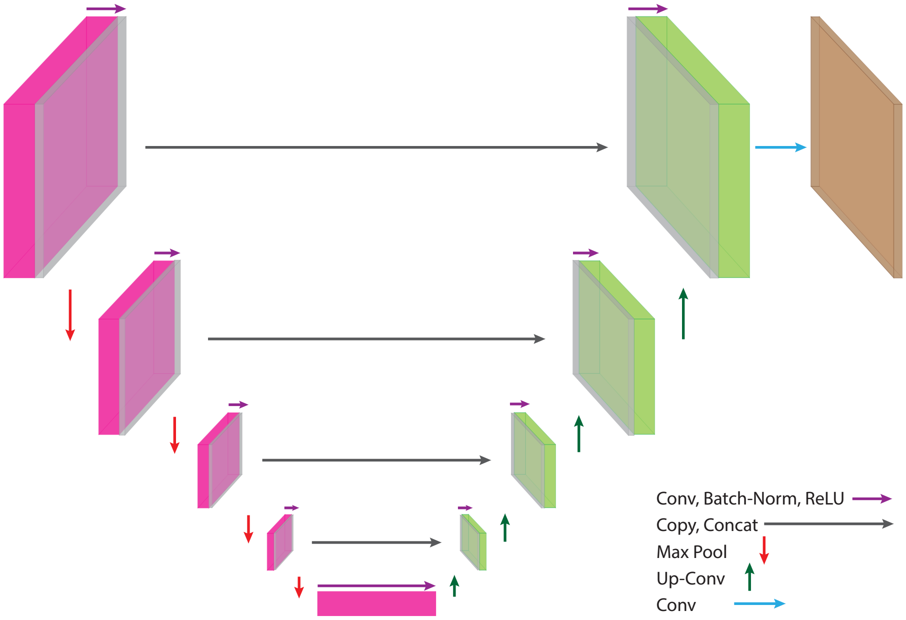


Figure 10. UNET architecture; the left-side is the contracting module (i.e. encoder), the right-side is the expansive module (i.e. decoder), the brown square is the segmentation output, and, finally, the grey squares are intermediary feature maps to be concatenated to their corresponding upsampled layer.

pixel (h, w); $y_s^{h,w}$ is the corresponding ground truth label for that pixel; y_s is the corresponding ground truth semantic map; n_S is the total number of pixels in the source domain S; H is binary cross-entropy loss shown in equation 4.3; and $\mathcal{L}_S$ is the total loss over all images in the source domain.

$$\mathcal{L}_S = \frac{1}{n_S} \sum_{s \in S} \sum_{h,w \in s} H(y_s^{h,w}, p_s^{h,w}(y_s|x_s)) \tag{4.2}$$

$$H(y, \hat{y}) = -ylog(\hat{y}) - (1 - y)log(1 - \hat{y}) \tag{4.3}$$

4.3.1.4 Learning module: self-training. The self-training loss function is also the standard cross-entropy loss on the target domain. However, it consists of two parts: one for unlabeled pixels that get their pseudo-labels from the source model and one for the labeled positive pixels. Considering that labeled positive pixels are used in the PU loss too, having an extra loss term in the self-training loss for the labeled positive pixels means putting more emphasis on available information for model training. In particular, this is appropriate since the labeling mechanism is under SCAR assumption, and thus such double emphasis on the labeled positive data does not add any bias to the learning procedure.

For calculating the self-training loss, first, I need to calculate the pseudo-labels for unlabeled pixels within all images in the target domain. Therefore, for each of the target images, x_t, I pass the image through the source model to obtain the pseudo-labels $\hat{y}_{pseudo} = argmax(p_s(y_s|x_t))$ for the unlabeled pixels. For calculating the total loss over all images in the target domain, $\mathcal{L}_{T_{pseudo}}$, as shown in equation 4.4, the followings terms are needed for the target input image x_t: (i)

$\hat{y}^{h,w}_{pseudo}$ the pseudo-label generated by the source model for unlabeled pixel (h, w), (ii) $y^{h,w}_{t_p}$ the ground truth label for positive-labeled target pixel (h, w), (iii) $p^{h,w}_t$ the target output probability distribution at pixel (h, w), (iv) y_t the corresponding target ground truth semantic map, and (v) n_T the total number of pixels in the target domain T. The extent to which the self-training loss contributes to the final loss is controlled by the hyper-paramenter λ_{pseudo} as it is shown in equation 4.1.

$$\mathcal{L}_{T_{pseudo}} = \frac{1}{n_T} \sum_{t \in T} \left\{ \sum_{h,w \in t_u} H(\hat{y}^{h,w}_{pseudo}, p^{h,w}_t(y_t|x_t)) + \sum_{h,w \in t_p} H(y^{h,w}_{t_p}, p^{h,w}_t(y_t|x_t)) \right\} \tag{4.4}$$

4.3.1.5 Learning module: positive-unlabeled. In comparison to positive-negative (PN) problem where fully-labeled data in label space $Y \in \{-1, +1\}$ are available, PU problem addresses availability of some positive labeled data and unlabeled data from both positive and negative classes. In the PN problem, positive and negative data are sampled from their corresponding distributions $p_p(x) = p(x|Y = +1)$ and $p_n(x) = p(x|Y = -1)$. When the goal is to learn the decision function $\mathcal{F}$ for the problem, then the empirical risk of $\mathcal{F}$ is estimated from PN data as follows:

$$\hat{R}_{pn}(\mathcal{F}) = \pi_p \hat{R}^+_p(\mathcal{F}) + \pi_n \hat{R}^-_n(\mathcal{F}) \tag{4.5}$$

Where $\pi_p = p(Y = +1)$ and $\pi_n = p(Y = -1) = 1 - \pi_p$ are the positive and negative class-prior probabilities, respectively. And, $\hat{R}^+_p(\mathcal{F}) = \frac{1}{n_p} \sum_{i=1}^{n_p} \ell(\mathcal{F}(x^p_i), +1)$

and $\hat{R}_n^-(\mathcal{F}) = \frac{1}{n_p}\sum_{i=1}^{n_p} \ell(\mathcal{F}(x_i^n), -1)$. In PU learning, the negative data is absent, and, instead, unlabeled data are available with distribution $p(x)$. In addition, having $\pi_n p_n(x) = p(x) - \pi_p p_p(x)$ helps to replace $\pi_n R_n^-(\mathcal{F})$ with $R_u^-(\mathcal{F}) - \pi_p R_p^-(\mathcal{F})$. Therefore, equation 4.5 can be re-written as follows:

$$\hat{R}_{pu}(\mathcal{F}) = \pi_p \hat{R}_p^+(\mathcal{F}) + \hat{R}_u^-(\mathcal{F}) - \pi_p \hat{R}_p^-(\mathcal{F}) \tag{4.6}$$

where $\hat{R}_p^-(\mathcal{F}) = \frac{1}{n_p}\sum_{i=1}^{n_p} \ell(\mathcal{F}(x_i^p), -1)$ and $\hat{R}_u^-(\mathcal{F}) = \frac{1}{n_u}\sum_{i=1}^{n_u} \ell(\mathcal{F}(x_i^u), -1)$. The equation 4.6 can result in negative empirical risks when flexible models such as neural networks overfit the data. Given that semantic segmentation models use neural networks, this type of overfitting can pose a significant problem. Therefore, a modification to equation 4.6 can solve the negative empirical risk problem. This is shown in the following:

$$\hat{R}_{pu}(\mathcal{F}) = \pi_p \hat{R}_p^+(\mathcal{F}) + \max\Big\{0, \hat{R}_u^-(\mathcal{F}) - \pi_p \hat{R}_p^-(\mathcal{F})\Big\} \tag{4.7}$$

The equation 4.7 (proposed by Kiryo et al., 2017) is referred to as the non-negative PU (NNPU) risk estimator. This PU loss function is used in the target domain and is re-written in equation 4.8, where $p_{t_p}^{h,w}$ and $p_{t_u}^{h,w}$ are the output probability distributions of the labeled and unlabeled pixels at locations (h, w), respectively, over the label space y_t; and n_{T_p} and n_{T_u} are the total number of labeled and unlabeled pixels in the target domain T, respectively. The extent to which the PU loss contributes to the final loss is controlled by the hyper-paramenter λ_{pu} as it is shown in equation 4.1.

$$\mathcal{L}_{T_{pu}} = \frac{\pi_p}{n_{T_p}} \sum_{t_p \in T} \sum_{h,w \in t_p} H(+1, p_{t_p}^{h,w}(y_t|x_{t_p})) \\ + \max\left\{0, \frac{1}{n_{T_u}} \sum_{t_u \in T} \sum_{h,w \in t_u} H(-1, p_{t_u}^{h,w}(y_t|x_{t_u})) - \frac{\pi_p}{n_{T_p}} \sum_{t_p \in T} \sum_{h,w \in t_p} H(-1, p_{t_p}^{h,w}(y_t|x_{t_p}))\right\} \tag{4.8}$$

4.3.1.6 Performance metric: Accuracy. Pixel accuracy measures the percentage of correctly classified pixels within an image, which then can be averaged over all images within the dataset. Since I am addressing binary classification, the per-class and global pixel accuracies are the same representing the pixel accuracy for the class of interest. The pixel accuracy is calculated using equation 4.9, where TP, TN, FP, and FN are true positive, true negative, false positive, and false negative rates that are calculated within the confusion matrix.

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \tag{4.9}$$

4.3.1.7 Performance metric: IoU. Pixel accuracy can be misleading if there is class-imbalance within the dataset. Therefore, pixel accuracy is always accompanied with other performance metrics such as IoU (or Jaccard index) and F1-score (or Dice coefficient), especially for semantic segmentation. Intersection over Union (IoU), shown in equation 4.10, quantifies the percentage of overlapping pixels between the model's prediction output and the ground truth semantic map. This metric can be reported per class or averaged over all classes as mean-IoU, but in my case that concerns binary classes, I report IoU for the class of interest.

$$IoU = \frac{TP}{TP + FP + FN} \tag{4.10}$$

4.3.1.8 Performance metric: F1-score.

F1-score (or Dice coefficient) combines two popular metrics, the precision and the recall, and is designed to handle data with class-imbalance. As shown in equation 4.11, F1-score equals to 2 times of the overlapping pixels divided by the total number of pixels in both models' prediction output and the ground truth semantic map.

$$F1 = \frac{2TP}{2TP + FP + FN} \tag{4.11}$$

4.3.2 Heterogeneous case.

The assumption under which I am working is that the homogeneous case is naive since it's rare to have such situations and, a more realistic case happens when the two source and target domains contain images from different geographic locations, acquired by the different sensors (i.e. satellites), and have different acquisition times, and, thus different atmospheric effects (Tuia et al., 2016). As shown in Fig. 11, remote sensors are designed to capture information from Earth's surface within specific range intervals of the electromagnetic spectrum called bands.

The RS imagery's bands may not overlap for different sensors causing different feature spaces for images acquired by those sensors. As shown in Fig. 12, there are misalignments of different remote sensors across different bands. In addition, the magnitude of the captured reflectance (one of the atmospheric effects) may not be the same for different satellites that are also affecting the feature space.

Therefore, heterogeneous case is applicable since the feature spaces are different mainly due to differences in sensors.

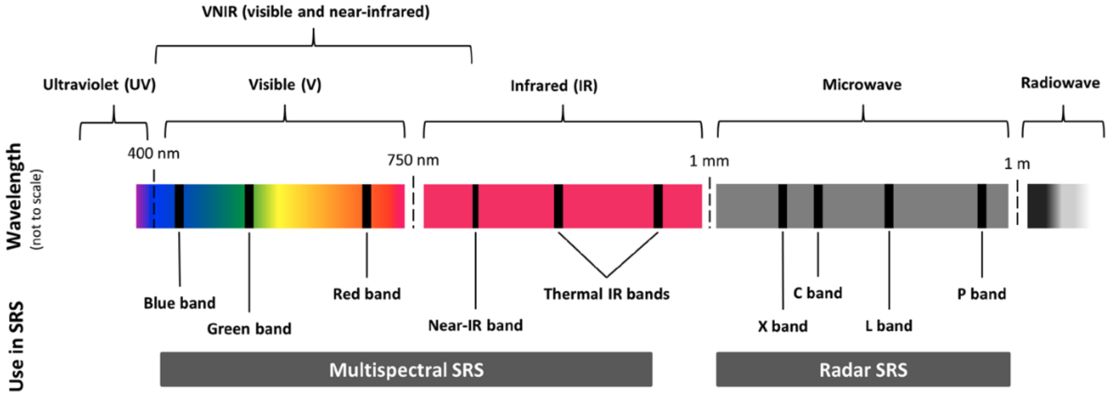
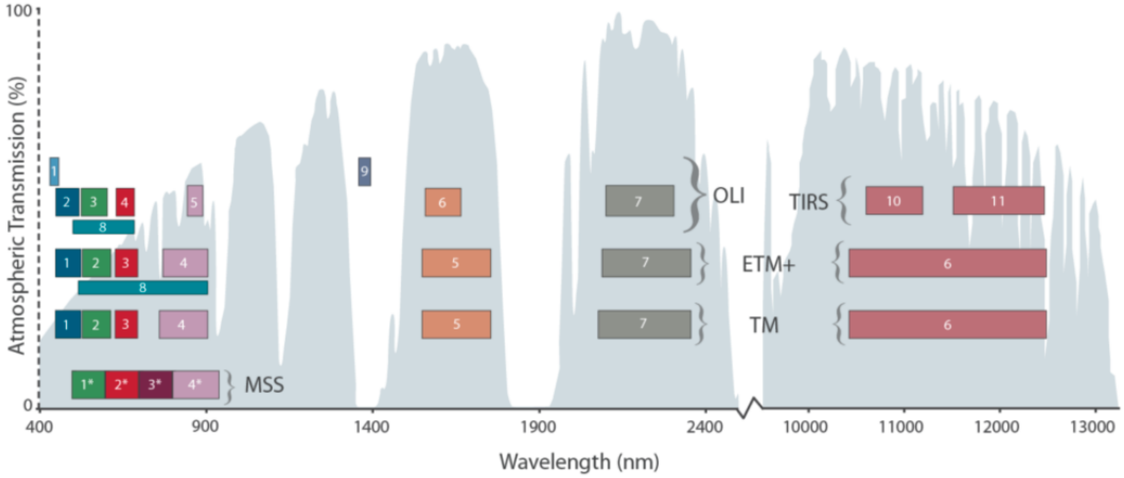


Figure 11. The electromagnetic spectrum captured by satellite remote sensing (shown as SRS in the image). Image adopted from Pettorelli et al. (2018).



* MSS bands 1–4 were known as bands 4–7, respectively, on Landsats 1–3

Figure 12. The misalignments of different remote sensors across different bands. Image adopted from Rocchio and Barsi (n.d.).

I propose to take advantage of consistency-training in order to tackle the differences among feature spaces. The reason behind this is that, although feature spaces are different, they overlap such that the centers of the bands are close or similar with differences in upper bound and/or lower bound for interval range of the bands. Consistency-training approaches have shown their effectiveness in training models that are robust to changes such as perturbation-based changes in input data (e.g. Y. Yang & Soatto, 2020). Therefore, I add a consistency-loss to the proposed model in the homogeneous section, which is shown in Fig. 13 and equation 4.12. The meaning of such action can be understood from a teacher-student network perspective (G. Hinton, Vinyals, Dean, et al., 2015) such that the source model can be considered as the teacher to the student network in the target domain. The differences considered in the electromagnetic spectrum range for different bands (RGB in the case of this dissertation) can be considered as some sort of perturbations that are added to images from one sensor to resemble images from another sensor. Thus, the proposed consistency-loss, explained below, makes sure that the model is robust to such changes in the input data, which results in a better feature generator for the model.

Besides the UNET architecture and its VGG16 encoder backbone, I incorporate DeepLabV2 architecture with ResNet101 as its encoder backbone as well since they are used in the PixMatch model. In the following section, I review the DeepLabV2 architecture and the ResNet101 encoder backbone and the proposed consistency-training module for the proposed model. The performance metrics used for models' assessment and evaluation are the same as before: accuracy, IoU, and F1-score.

$$\begin{aligned} \mathcal{L}_{heterogeneous_S} &= \mathcal{L}_{homogeneous_S} \\ \mathcal{L}_{heterogeneous_T} &= \mathcal{L}_{homogeneous_T} + \lambda_{consistency}\mathcal{L}_{consistency} \end{aligned} \tag{4.12}$$

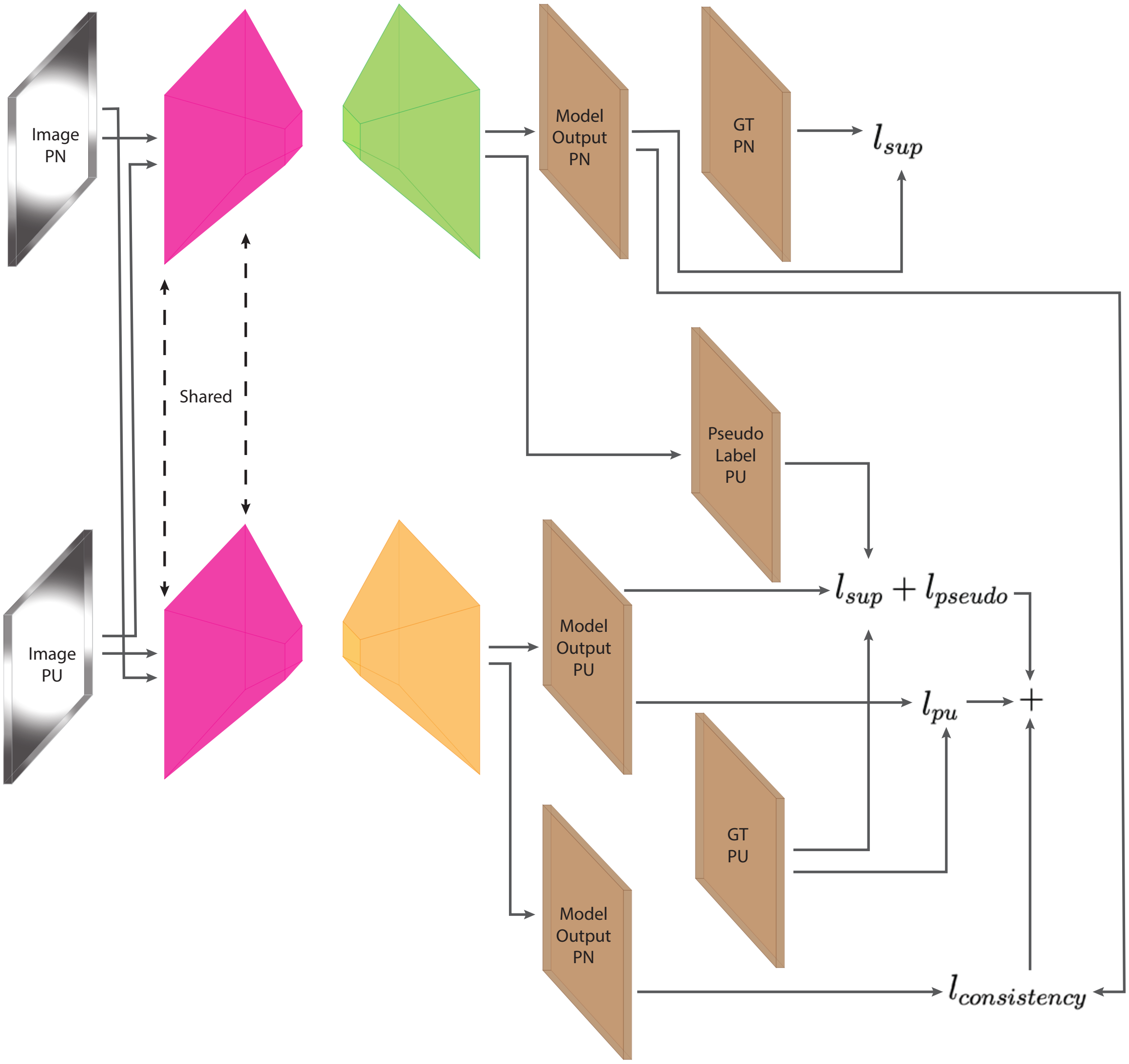


Figure 13. Heterogeneous transfer positive and unlabeled learning architecture.

4.3.2.1 Model architecture: DeepLabV2.

DeepLabV2, by L.-C. Chen, Papandreou, Kokkinos, Murphy, and Yuille (2017), is also a segmentation network with innovative Atrous Convolution at its core. DeepLabV2 leverages high-level global features and fine-grained details by a multi-scale feature aggregation technique with a final element-wise summation in order to create the final feature map. Next, the final feature map is passed to a bi-linear interpolation layer to reconstruct an output with the same size as input image. Fig. 14 shows the architecture schema of the DeepLabV2 model. The Atrous Convolution is an improvement on convolution kernels (for semantic segmentation in comparison to image classification) and provides the opportunity for capturing the global relationships among each pixel and its neighboring pixels at different levels—that is controlled by the atrous (i.e. dilation) rate, while keeping the computation costs low. DeepLabV2 utilizes the Atrous Convolution using the Spatial Pyramid Pooling (SPP) technique in order to create multi-scale feature-maps created by different atrous rates from 6 to 24. Finally, the whole aforementioned path for creating the final feature-map is repeated three times over 1.0, 0.75, and 0.5 downscaled input images for another way of multi-scale feature fusion. It should be mentioned that DeepLabV2 also uses a Fully-connected Conditional Random Field (CRF) module. The CRF module is a probabilistic method that helps with better label predictions of pixels around the boundaries of objects in images based on pixels correlations.

4.3.2.2 Model backbone: ResNet-101. ResNet-101, by K. He et al. (2016), is also a CNN Architecture developed for image classification and is also based on layered stacks of convolutional and max-pooling layers followed, at the end, by an average-pooling, a FC, and a softmax layer. Again, when it is used for semantic segmentation, the last three layers are filtered out. The innovation of general ResNet architecture—the 101 part refers to the total number of layers—is the residual block that solves the vanishing gradient problem in deep networks, causing accuracy degradation as more layers get added to an existing network. Residual blocks (Fig. 15) take advantage of the residual function such that instead of learning a mapping function, $\mathcal{H}(x)$, (i.e. a stack of convolutional layers) between the input and the output, the attempt is to learn a residual function, $\mathcal{F}(x)$ where $\mathcal{F}(x) = \mathcal{H}(x) - x$ (or i.e. $\mathcal{H}(x) = \mathcal{F}(x) + x$). In this way, learning the residual function is much easier and avoids the vanishing gradient problem. As opposed to ResNets, with smaller number of layers, the 101-layer version uses 3-layer (instead of 2-layer) deep residual blocks. In each residual block, if the output dimensions is

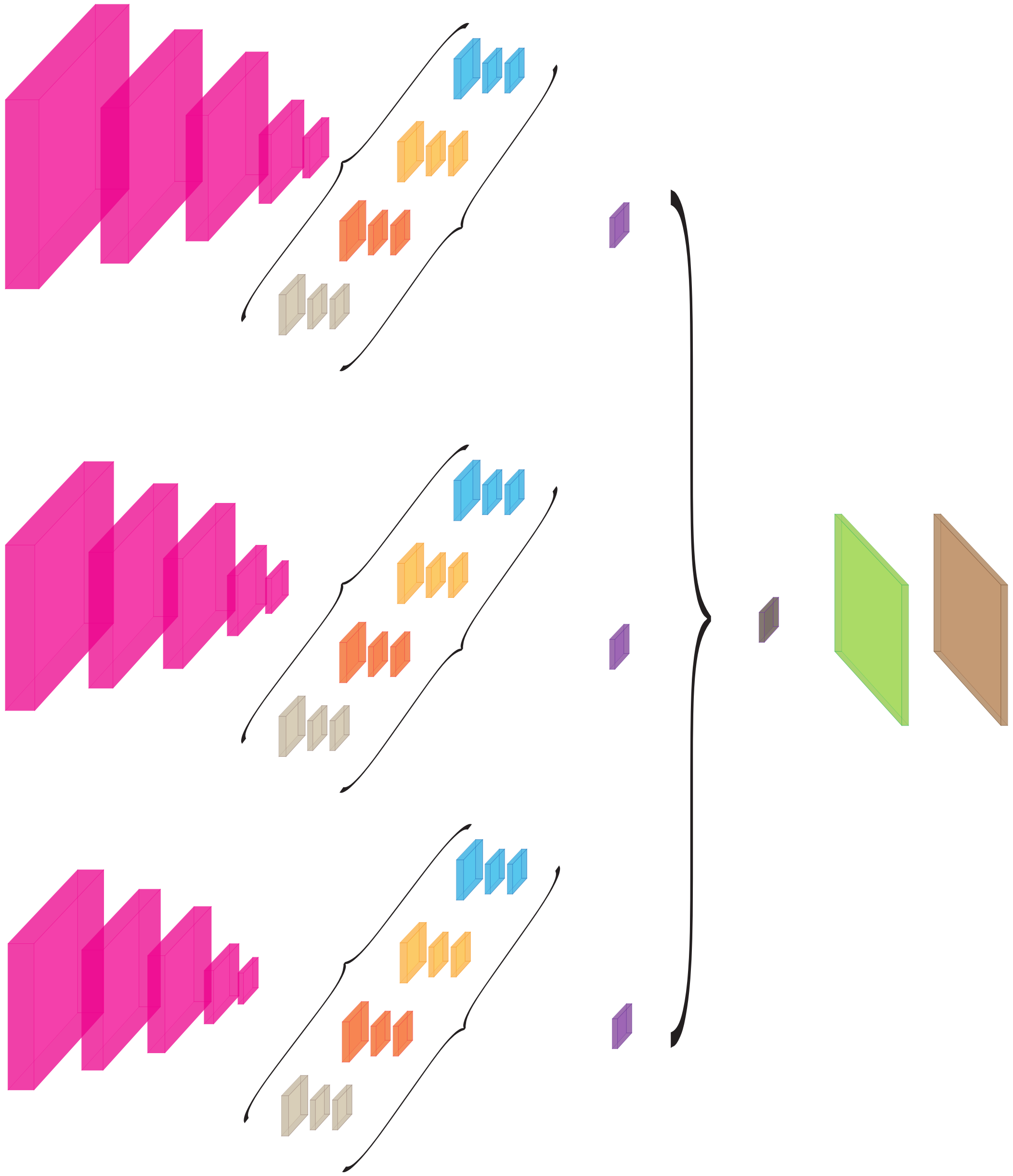


Figure 14. DeepLabV2 architecture; top-to-bottom are three repetitions of left-to-right path in each row over 1.0, 0.75, and 0.5 downscaled input images; left-to-right path are multi-scale feature generators with different atrous rates from 6 to 24; the final feature map is upsampled to generate the input-size-like output.

the same as the input, the input is added directly to the output; otherwise, first, a linear projection is applied to ensure that dimensions match.

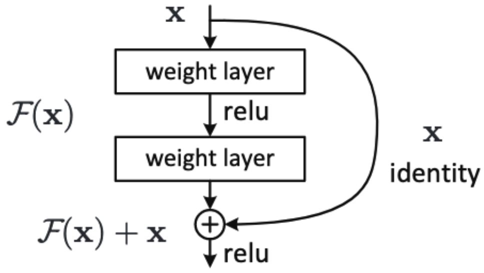


Figure 15. A sample building block showing the idea of residual learning. Figure adopted from K. He et al. (2016).

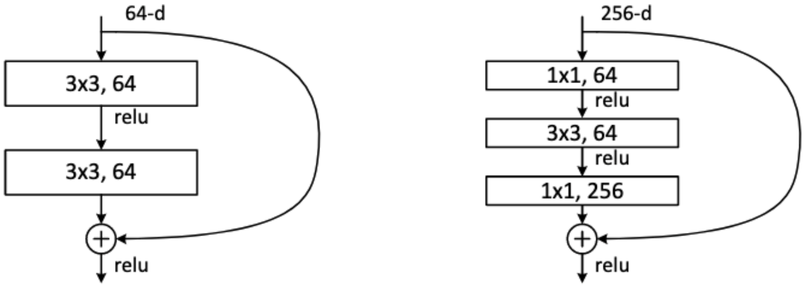


Figure 16. Residual learning block: 2- versus 3-layer. Figure adopted from K. He et al. (2016).

4.3.2.3 ***Learning module: consistency-training.*** The consistency-training loss is applied on the target model and measures the ℓ_1-norm (shown as $|| \cdot ||_1$ in equation 4.13) or the absolute value of difference between probability predictions from the source model output and the target model output given the input source image, x_s. In equation 4.13, $p_s^{h,w}$ and $p_t^{h,w}$ are the output probability distributions of the source and target models at pixel (h, w), respectively, over the label spaces $y_s = y_t$; and n_S is the total number of pixels in the source domain S. The extent to which the consistency loss contributes to the final loss is controlled by the hyper-paramenter $\lambda_{consistency}$ as it is shown in equation 4.12.

$$\mathcal{L}_{T_{consistency}} = \frac{1}{n_S} \sum_{s \in S} \sum_{h,w \in s} ||p_s^{h,w}(y_s|x_s) - p_t^{h,w}(y_t|x_s)||_1 \tag{4.13}$$

4.4 Results

In this section, I present the results from the proposed method and the corresponding baselines. I again start with the homogeneous case, and then move on to the heterogeneous case.

4.4.1 Homogeneous case. The performance of models in this section is evaluated using PN and PU datasets as the source and target domains from the homogeneous dataset created from Inria Image Dataset, which is discussed in Chapter III. As for the baseline, two different models are considered. First, a PU model with the NNPU loss (Kiryo et al., 2017) that is trained only on PU images within the target domain in two different scenarios: (i) with and (ii) without pre-trained weights from ImageNet dataset—this model is called *Target-PU*. The next model, *Target-FT*, is an extension of the first model in a way that the model first is trained on images from the source domain with and without pre-trained weights from ImageNet dataset. Then, the weights of this model are used as the warm start for the PU learning phase on PU images within the target domain.

The effect of the proposed model is analyzed with and without the PU learning module resulting in two models respectively called *Seamless-U* and *Seamless-PU*. Each of these two models, again, are considered in two cases: (i) with and (ii) without pre-trained weights from ImageNet dataset. The reason behind having Seamless-U alongside Seamless-PU is to investigate and understand to what extent the assumptions made on homogeneous cases are effective in practice.

The use of three channels of RGB for the models allows me to be able to use the pre-trained weights from ImageNet dataset. The optimizer is chosen to be Stochastic Gradient Descent with learning rate, momentum, and weight decay equal to 0.01, 0.9, and 1e-4, respectively. I adjust the learning rate during the training phase by reducing the learning rate according to *LambdaLR* scheduler. It is assumed that the class-prior probability is known and equal to $\pi_p = 0.5$. Although the exact value of the class-prior is calculated from the dataset and is equal to ~ 0.42, I consider a buffer since having the exact and precise value of the class-prior is too ideal and rarely (i.e. almost never) happens in a real scenario. This assumption is justifiable in a sense that it does not affect the comparison of the models since Kiryo et al. (2017) show robustness of NNPU to the misspecification of the class-prior probability within the range of $[0.8\pi_p, 1.2\pi_p] = [0.36, 0.54]$. Finally, the rest of the settings of NNPU loss, such as values for α and β, are set to be the same as the original paper by Kiryo et al. (2017).

Table 3 shows the results of running all models on PU5, PU6, PU7, PU8, and PU9 datasets for different level of positive-class labeled probabilities—since Sealmless-U model uses the target domain images as unlabeled, its performance is the same across different PU datasets, and this is shown by arrows in Table 3.

The performance of both baselines degrades when they are trained with pre-trained weights from ImageNet dataset. However, such pre-trained weights help improve the proposed model in both U and PU cases. One interpretation of this could be that the proposed approach finds a feature generator that is more domain invariant, and, thus, it can incorporate other related domains such as ImageNet dataset better without introducing negative transfer effect. However, this is not the case for the baselines resulting in having negative transfer when using image domains that are not quite similar to RS imageries. The best performance among all models for all PU datasets belongs to the proposed approach with domain invariant feature generator that also takes advantage of the PU learning module as well as ImageNet's pre-trained weights.

Next, I will address the models' behavior in different PU datasets. Looking at Table 3, it can be seen that, among the PU datasets, the PU dataset on which models perform the best is not consistent. In other words, adding more labeled positive data does not result in a consistent behavior in terms of reducing models error and improving their performances. This is against the general belief that the more the labeled data, the better the trained model. However, what is important is that the proposed model outperforms all other models on all, and, especially the PU9 dataset, which makes it valuable considering the few amount of positive labeled data required for such performance.

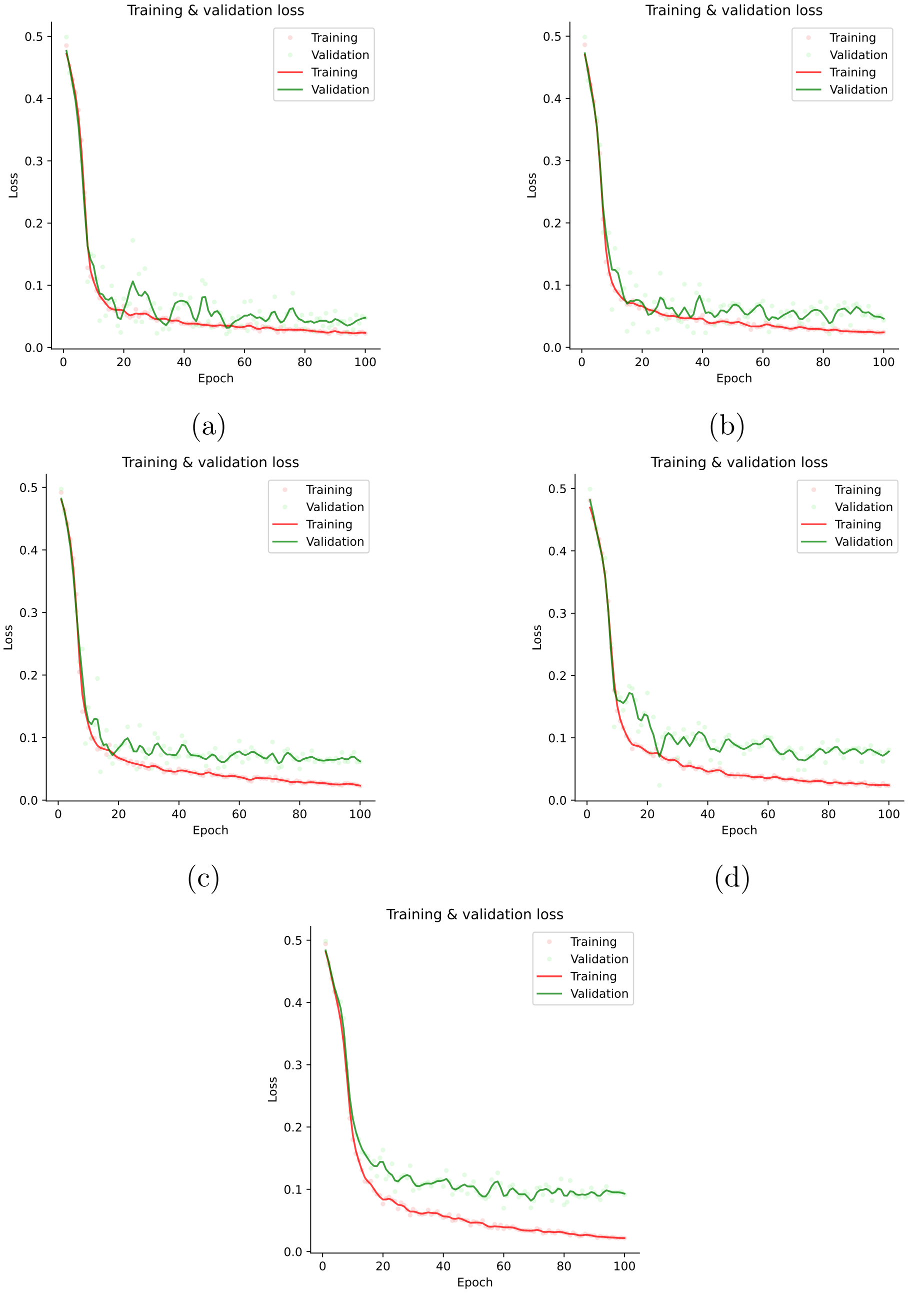
In the case of Target-PU and Target-FT models, as shown in Fig. 17, and 19, the less the labeled data are available, the more the gap between the train and validation losses expands. This means that the models are less generalizable since they receive fewer signals from the smaller amount of labeled data. In general, there is one other reason for such divergence, which is that the train and val sets are not representative of each other in some PU datasets. However, this is not the case here since the images are a blend of the same geographic location, same satellite sensor, same time, and same atmospheric conditions. The divergence between train and validation losses is more noticeable when using ImageNet pre-trained weights as opposed to training from initial random weights (Fig. 17 vs. Fig. 18; and Fig. 19 vs. Fig. 20), which can be seen as the negative transfer effect that is discussed above. However, such divergence does not appear for

Table 3. The performance of the proposed model and the baselines in the target domain

	PU5			PU6			PU7			PU8			PU9			Best Case
Method	Acc	IoU	F1	Acc	IoU	F1	Acc	IoU	F1	Acc	IoU	F1	Acc	IoU	F1	
Target-PU	77.69	63.45	77.57	81.06	67.46	80.49	81.52	66.95	80.13	80.75	66.37	79.69	81.38	66.59	79.87	PU6
Target-PU(pre-train)	73.94	53.42	69.55	74.50	53.71	69.76	74.68	51.67	67.87	74.26	51.96	68.27	74.12	51.54	67.52	PU6
Target-FT	77.70	64.00	77.89	80.31	67.16	80.30	81.98	67.94	80.81	81.96	68.11	80.94	82.50	68.44	81.21	PU9
Target-FT(pre-train)	77.34	63.51	77.57	79.61	66.28	79.66	81.26	67.16	80.26	80.73	66.45	79.79	80.09	65.65	79.22	PU7
Seamless-U	87.62	73.69	84.79	←	←	←	←	←	←	←	←	←	←	←	←	N/A
Seamless-U(pre-train)	88.81	76.01	86.31	←	←	←	←	←	←	←	←	←	←	←	←	N/A
Seamless-PU	87.67	74.48	85.31	87.70	74.56	85.37	87.80	74.05	85.05	87.44	73.89	84.92	87.67	74.44	85.30	PU6
Seamless-PU(pre-train)	88.85	76.63	86.71	89.21	77.21	87.09	89.19	76.54	86.66	88.92	76.41	86.58	88.58	76.01	86.21	PU6

Seamless-U and Seamless-PU models since these models try to learn the maximum information available in a joint information space by the two domains. These results are illustrated in Fig. 21, 22, 23, and 24. These figures, however, show some irregularities in the downward trend of the loss function, which could be due to switching the encoder back and forth between the two domains during the learning phase.

Fig. 25 qualitatively supports the quantitative results in Table 3. Fig. 25 shows a sample of image patches, their corresponding ground truth, and predictions from the best performing model between with and without pre-trained weights for each model trained on the PU5 case: (i) Target-PU without pre-trained weights, (ii) Target-FT without pre-trained weights, (iii) Seamless-U with pre-trained weights, and (iv) Seamless-PU with pre-trained weights. Both baselines—even when provided with the maximum amount of labeled positive data (i.e. PU5)—result in very large amount of false positive, while both Seamless-U and Seamless-PU do not suffer from that. Also, it seems that the false negatives are of a higher rate in Seamless-U than Seamless-PU, which, along with the quantitative results, makes the Seamless-PU model, the best model among all. Finally, Fig. 26 shows the models' outputs for the other extreme side when there is the minimum amount of labeled positive data available (i.e. PU9). Again, it can be seen that the proposed Seamless-PU model outperforms the rest of the models.



(a)

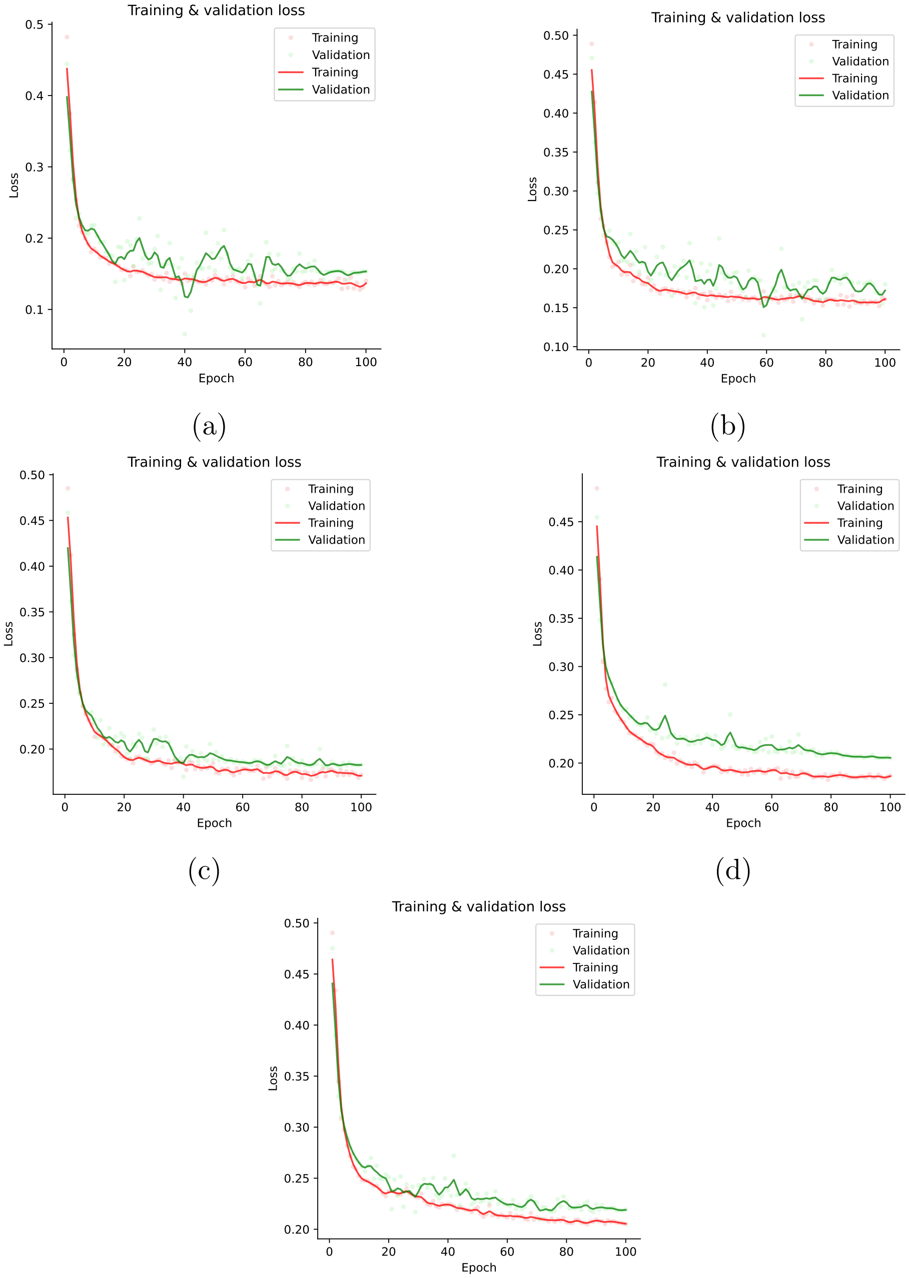
(b)

(c)

(d)

(e)

Figure 17. The training loss (in red) and validation loss (in green) for Target-PU model without ImageNet's pre-trained weights on (a) PU5, (b) PU6, (c) PU7, (d) PU8, (e) PU9 datasets.

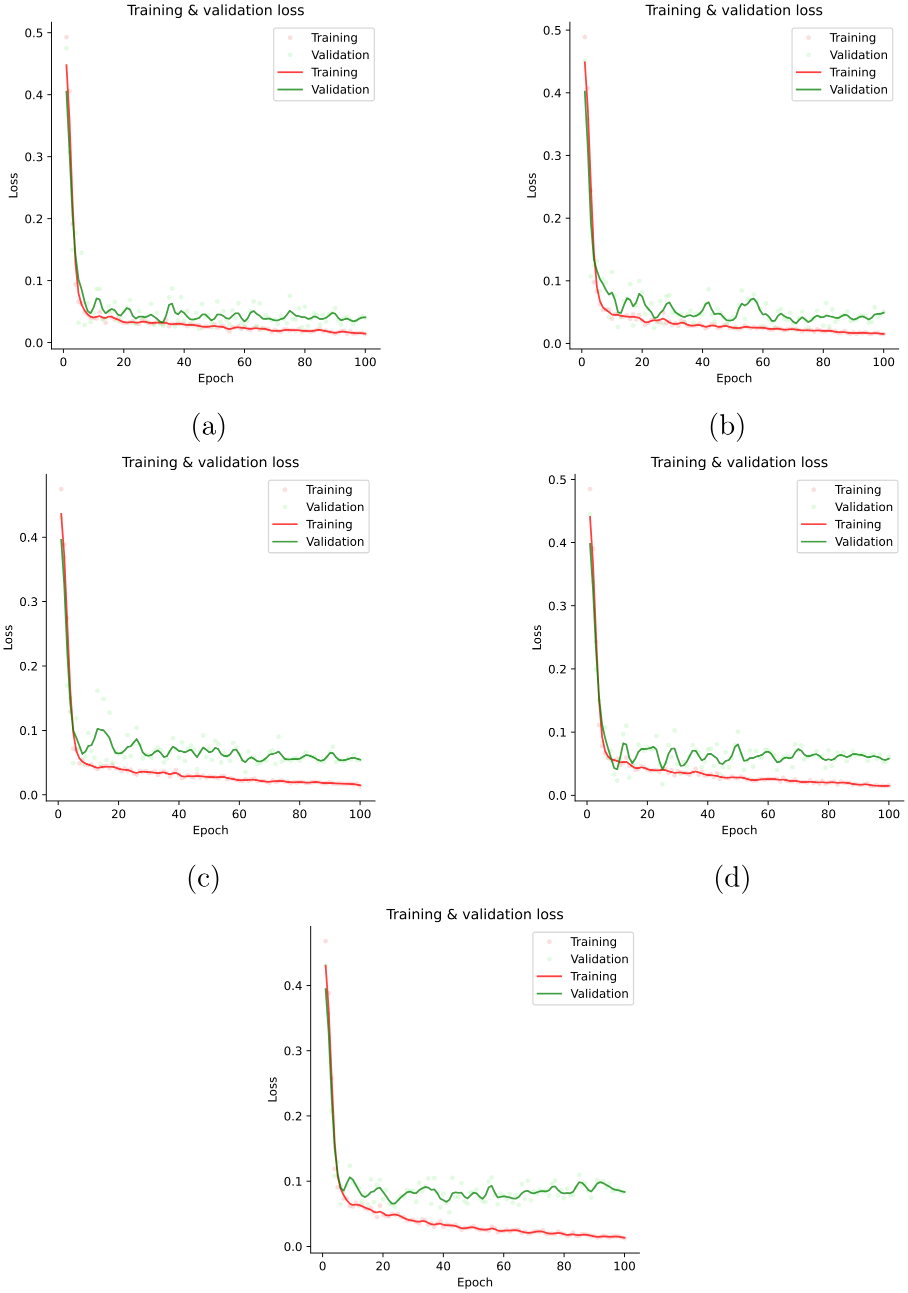


(a) (b)

(c) (d)

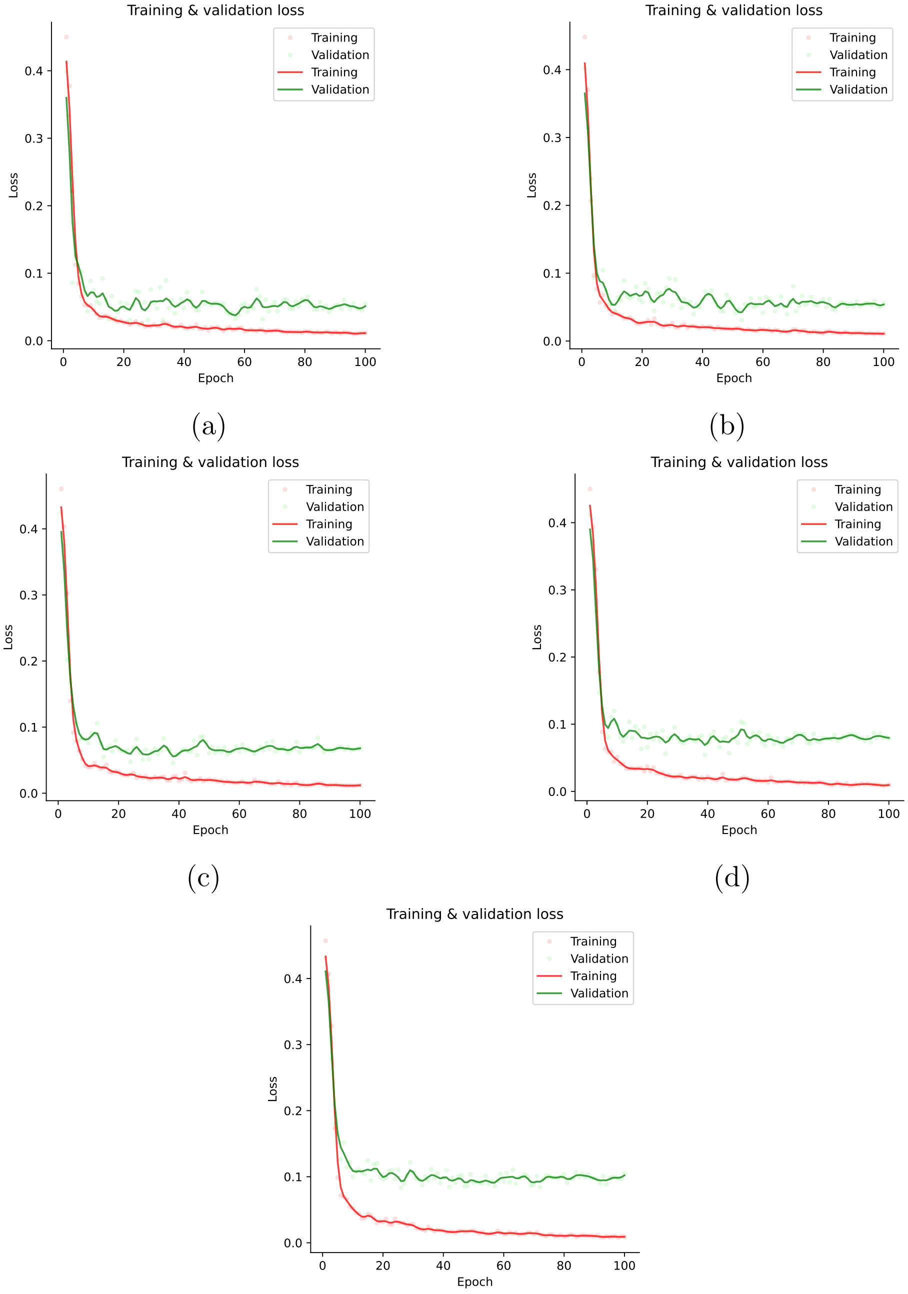
(e)

Figure 18. The training loss (in red) and validation loss (in green) for Target-PU model with ImageNet's pre-trained weights on (a) PU5, (b) PU6, (c) PU7, (d) PU8, (e) PU9 datasets.



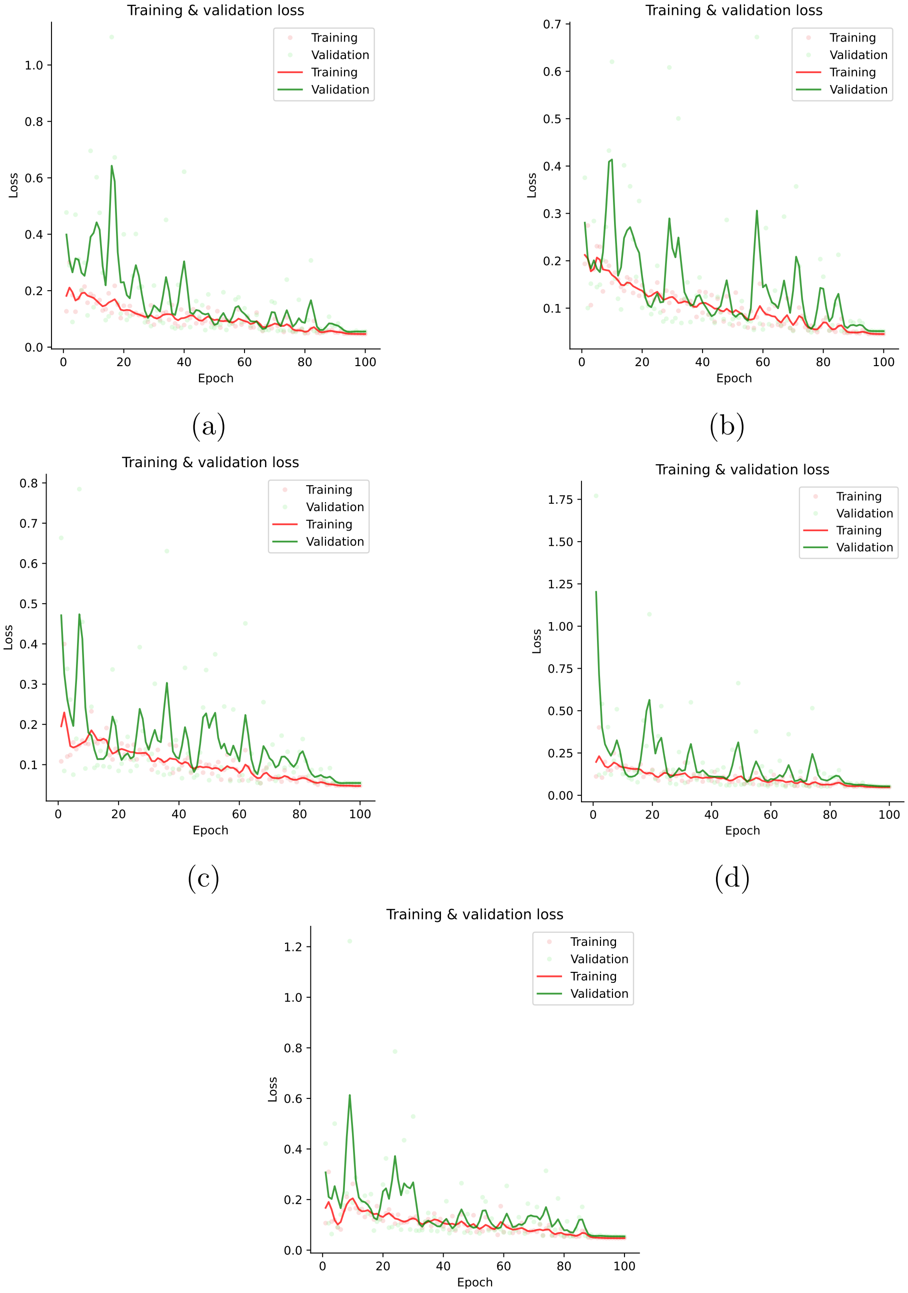
(a) (b) (c) (d) (e)

Figure 19. The training loss (in red) and validation loss (in green) for Target-FT model without ImageNet's pre-trained weights on (a) PU5, (b) PU6, (c) PU7, (d) PU8, (e) PU9 datasets.



(a) (b) (c) (d) (e)

Figure 20. The training loss (in red) and validation loss (in green) for Target-FT model with ImageNet's pre-trained weights on (a) PU5, (b) PU6, (c) PU7, (d) PU8, (e) PU9 datasets.



(a)

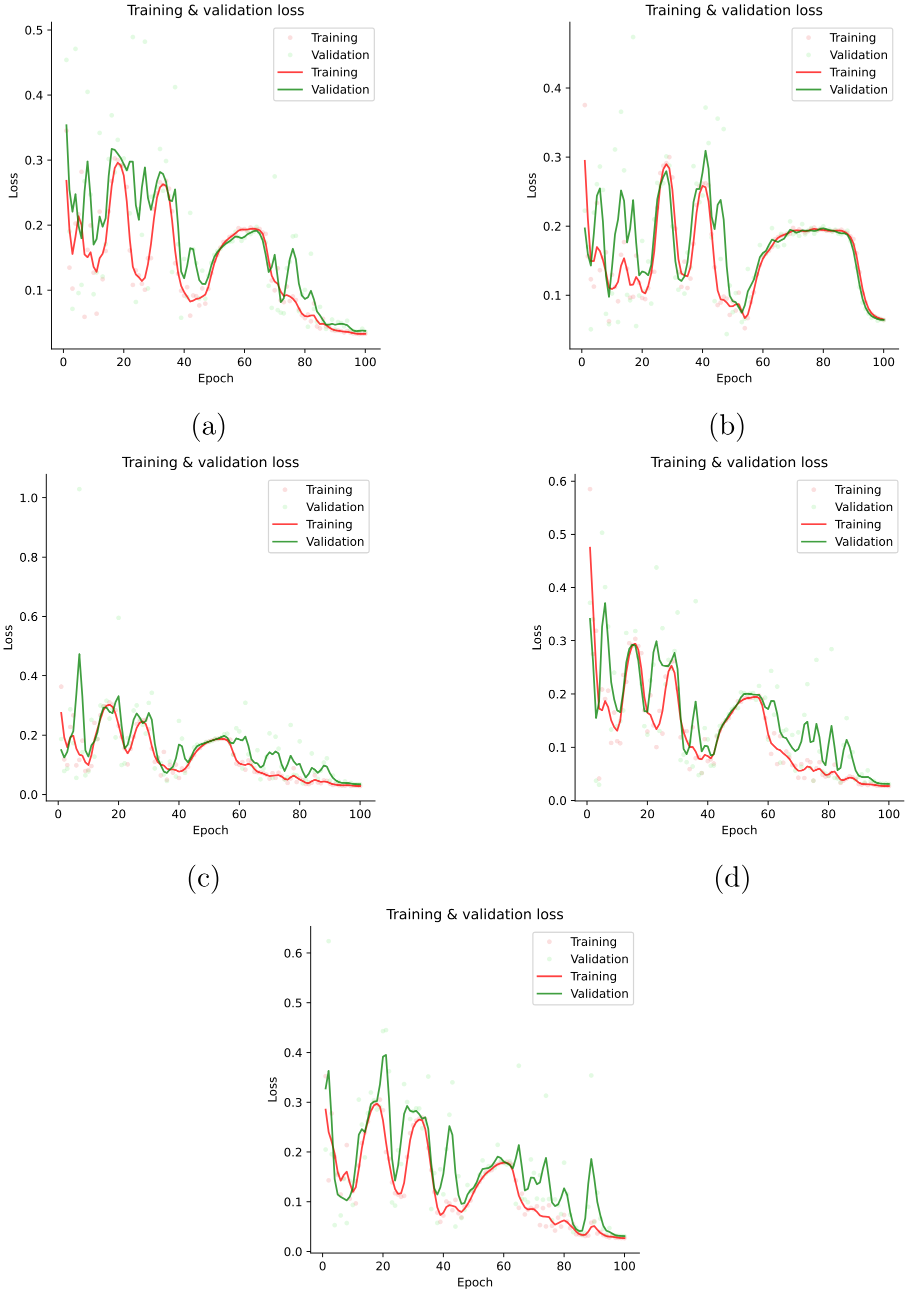
(b)

(c)

(d)

(e)

Figure 21. The training loss (in red) and validation loss (in green) for Seamless-U model without ImageNet's pre-trained weights on (a) PU5, (b) PU6, (c) PU7, (d) PU8, (e) PU9 datasets.



(a)

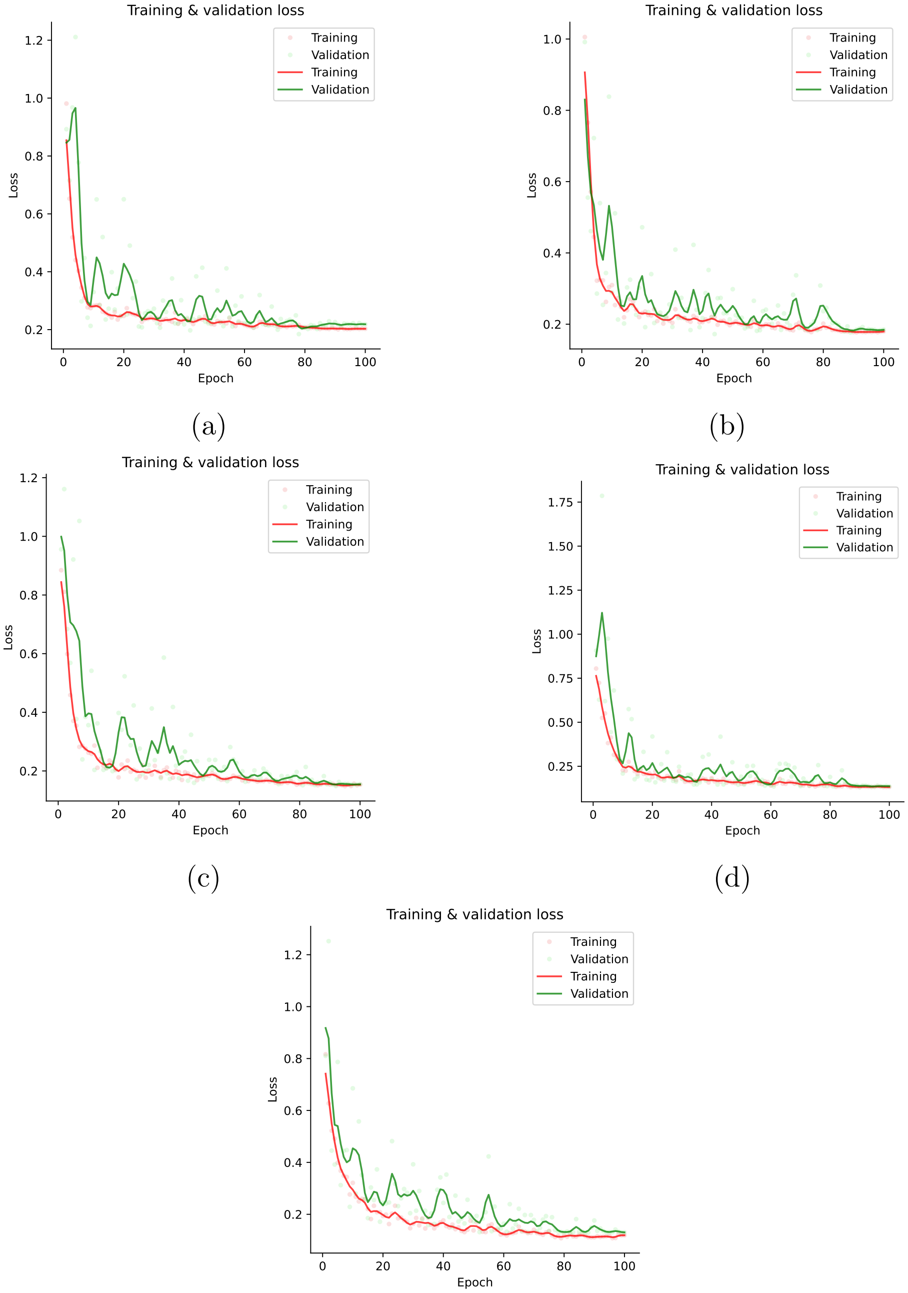
(b)

(c)

(d)

(e)

Figure 22. The training loss (in red) and validation loss (in green) for Seamless-U model with ImageNet's pre-trained weights on (a) PU5, (b) PU6, (c) PU7, (d) PU8, (e) PU9 datasets.



(a)

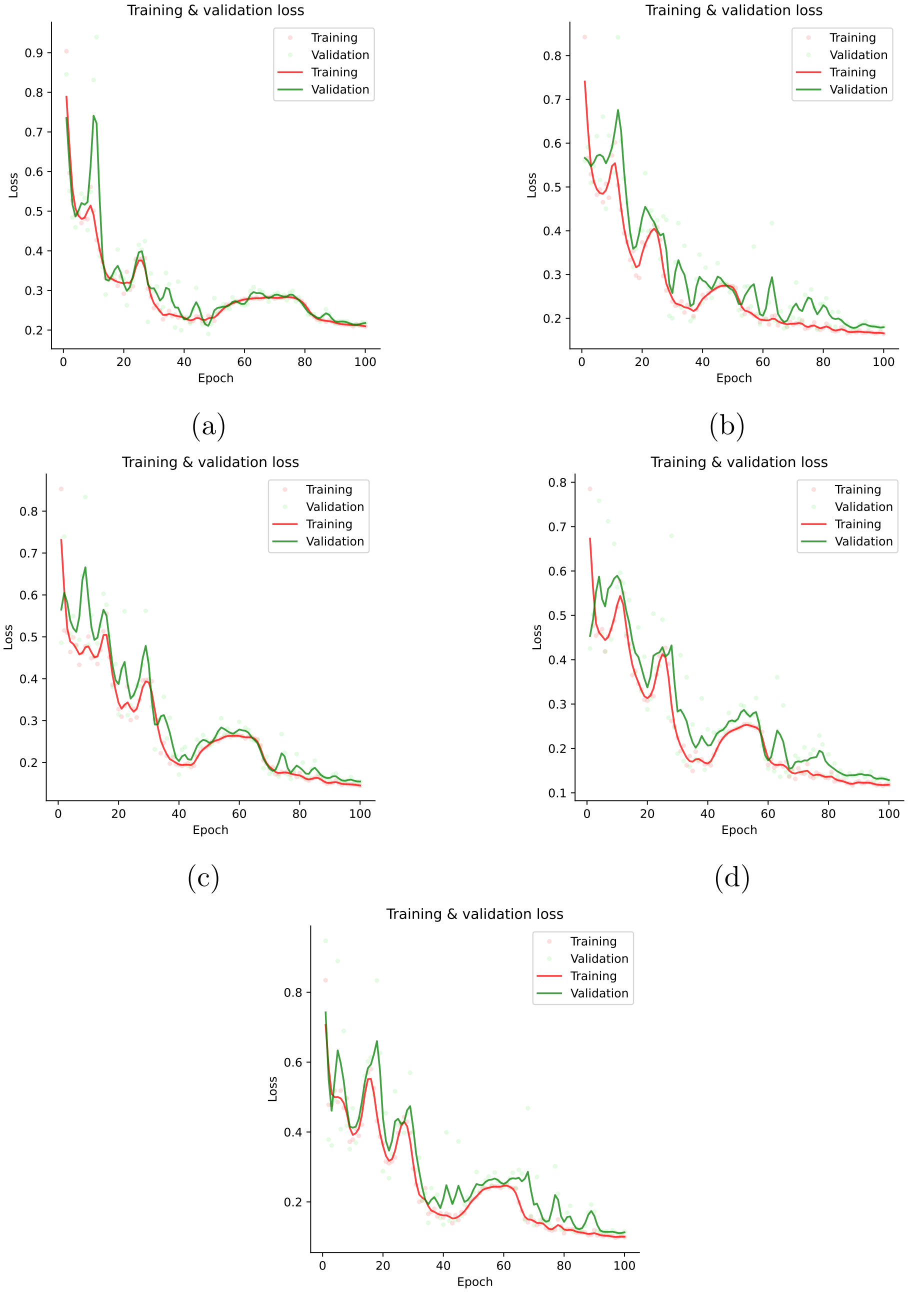
(b)

(c)

(d)

(e)

Figure 23. The training loss (in red) and validation loss (in green) for Seamless-PU model without ImageNet's pre-trained weights on (a) PU5, (b) PU6, (c) PU7, (d) PU8, (e) PU9 datasets.



(a)

(b)

(c)

(d)

(e)

Figure 24. The training loss (in red) and validation loss (in green) for Seamless-PU model with ImageNet's pre-trained weights on (a) PU5, (b) PU6, (c) PU7, (d) PU8, (e) PU9 datasets.

4.4.1.1 Conclusions. In this section, I presented the results of the proposed model for the homogeneous case, which is the basic building block for the heterogeneous case. I demonstrated that, even in this easier scenario (compared to the heterogeneous scenario), with or without transferring pre-learned knowledge (i.e. model weights) from similar or different image domains, NNPU-based baselines do not perform as well as the proposed approach and do not show competitive results. What is more, NNPU-based baselines degrade in heterogeneous cases and cannot be used as a baseline for the proposed model in such cases. For the proposed model, the domain-invariant feature learner accompanied with signals from PU data has shown promising results and suggests its potential compatibility for the heterogeneous setting. Therefore, building upon this, in the next section, I will present the results from the proposed model with consistency-learning module for the heterogeneous setting and show that, with only a fraction of labeled positive data, the model can handle this even more challenging scenario and perform much better than the state-of-the-art UDA models.

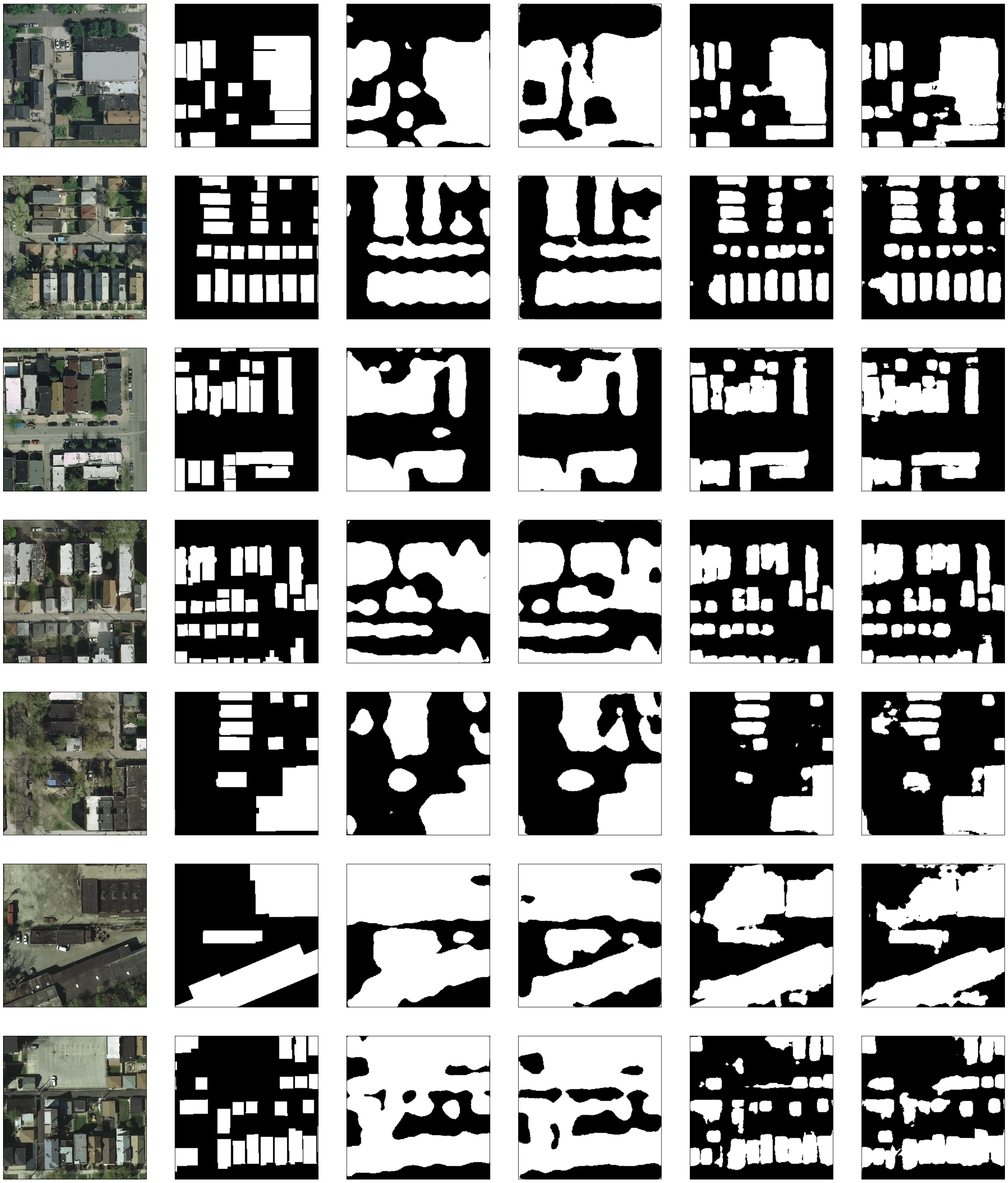


Figure 25. Left-to-right: Sample image patches from homogeneous PU Inrial Dataset, the corresponding ground truth, and predictions from four models trained on the PU5 case: (i) Target-PU without pre-trained weights, (ii) Target-FT without pre-trained weights, (iii) Seamless-U with pre-trained weights, and (iv) Seamless-PU with pre-trained weights.

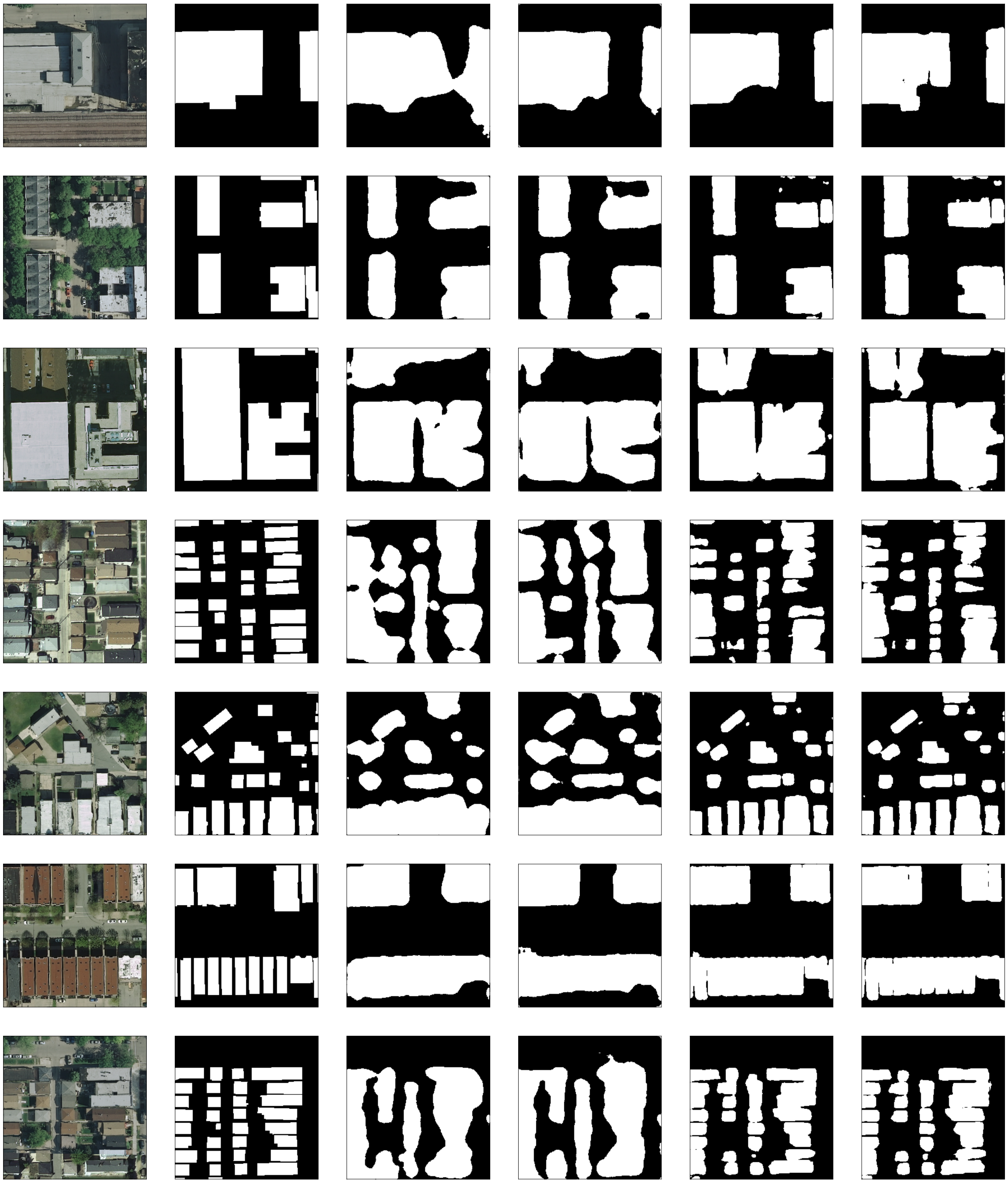


Figure 26. Left-to-right: Sample image patches from homogeneous PU Inrial Dataset, the corresponding ground truth, and predictions from four models trained on the PU9 case: (i) Target-PU without pre-trained weights, (ii) Target-FT without pre-trained weights, (iii) Seamless-U with pre-trained weights, and (iv) Seamless-PU with pre-trained weights.

4.4.2 Heterogeneous case. The performance of models in this section is evaluated using the heterogeneous dataset containing PN Massachusetts Buildings Dataset as the source domain and the complete PU datasets with different labeling frequencies for the positive class (i.e. PU5, PU6, PU7, PU8, and PU9) from Inria Image Dataset created in Chapter III as the target domain. PixMatch (Melas-Kyriazi & Manrai, 2021) is considered as the baseline with its recommended hyper-parameters setting from the original paper. PixMatch has different variations based on the type of the perturbation that is used in it. The extent to which each perturbation contributes is controlled with its coefficient λ_i. Within the original paper, it is not clear what is the value for λ for *CutMix*, *Fourier*, and *Fourier+CutMix* perturbations. Therefore, I use the same best value of λ that is mentioned for *Augmentation* perturbation, which is equal to 0.15. For the proposed Sealmess-PU model, all the hyper-parameters are the same as before in the homogeneous section.

Table 4 shows the performance drop of the proposed Sealmess-PU model when trained in a heterogeneous scenario. The performance drop is not very far in the case of PU5 with the highest amount of labeled positive data. However, for the PU9 case, which is the most interesting case, the model performs poorly. This situation requires some considerations within the proposed model to make up for the challenges caused by the heterogeneous scenario.

According to Table 5, for RS imageries, the PixMatch model in its vanilla form performs as well as when accompanied by augmentation perturbation and even better than when the other perturbations are incorporated. In addition, the performance gap among different perturbations is relatively large. These observations do not follow the results in the PixMatch paper for urban scene

Table 4. Performance drop of PUDA model (Seamless-PU) without the consistency loss on heterogeneous PU data from Inria Image Dataset. The model uses a warm start using the ImageNet pre-trained weights.

Acc	IoU	F1	PU Dataset
83.58	68.85	81.51	PU5
79.64	62.28	76.70	PU6
79.12	60.35	75.22	PU7
72.68	53.04	69.26	PU8
65.31	34.06	50.66	PU9

datasets. One explanation for this could be that the type of the data determines which perturbation technique is most effective. This hypothesis is based on the results in Balestriero, Bottou, and LeCun (2022) where they discover that not all data augmentation techniques affect the different classes within the dataset in the same way, and thus there could be model performance drop for some classes with some specific data augmentation while the rest of the classes experience model performance gains. In addition, the performance of PixMatch model depends heavily on the network architecture used. For example, changing the network architecture from DeepLabV2 with ResNet-101 backbone to UNET with VGG16 backbone degrades the PixMatch(*Augmentations*) model's performance (i.e. IoU) from 33.67 to 01.79.

As shown in Table 5, the proposed Sealmess-PU model outperforms the PixMatch baseline model. I also investigate the effect of model architecture and the type of backbone on the performance of Sealmess-PU model. Therefore, I change the architecture to DeepLabV2 and the backbone to ResNet-101. Since PixMatch model also uses an exponential weighted moving average component in its learning process, I use such component too in the changes to the original proposed Sealmess-PU model to make everything the same and the comparison fair. Although this setting for Sealmess-PU model still outperforms the PixMatch baseline model, it results in performance improvements only on PU5, PU6, and PU7 datasets, whereas it results in performance degradations on PU8 and PU9 datasets. The reasons behind such performance changes can be due to (i) the difference in the number of learnable parameters between the two architectures and/or (ii) the operations used in each of the two architectures. In terms of the former, DeepLabV2 model has 42610632 ($\sim$ 43 million) learnable parameters with 42500032 ($\sim$ 42.5 million) of these parameters being for the encoder part, whereas UNET model has 49698434 ($\sim$ 50 million) learnable parameters with 14723136 ($\sim$ 15 million) of these parameters being for the encoder part, and each decoder having 17487649 ($\sim$ 17.5 million) learnable parameters. The lightweight characteristics of the UNET model for the target domain is an advantage when facing limited number of labeled data. For the latter, my hypothesis is that operations such as Atrous Convolution may depend on the supervision that comes from the labeled data, which causes model's performance degradations (this hypothesis needs to be further investigated in future). Finally, Fig. 27 shows a sample image patches from heterogeneous PU Inrial Dataset, their corresponding ground truth, and predictions from PixMatch(*Augmentations*), Seamless-PU with UNET trained on PU5 dataset, Seamless-PU with DeepLabV2 trained on PU5 dataset, Seamless-PU with UNET trained on PU9 dataset, and Seamless-PU with DeepLabV2 trained on PU9 dataset.

Table 5. Performance of unsupervised domain adaptation model (PixMatch) and the proposed PUDA model (Seamless-PU). All utilize a warm start using the ImageNet pre-trained weights.

Method	Architecture	Backbone	Acc	IoU	F1	PU Dataset
PixMatch(*Fourier*)	UNET	VGG	57.82	01.78	03.50	N/A
PixMatch(*CutMix*)			57.82	01.87	03.67	N/A
PixMatch(*Augmentations*)			57.82	01.79	03.51	N/A
PixMatch(*Fourier* + *CutMix*)			57.81	01.28	02.52	N/A
PixMatch	DeepLabV2	ResNet-101	66.82	33.19	49.56	N/A
PixMatch(*Fourier*)			54.21	23.15	37.51	N/A
PixMatch(*CutMix*)			54.14	20.54	34.00	N/A
PixMatch(*Augmentations*)			63.83	33.67	50.15	N/A
PixMatch(*Fourier* + *CutMix*)			54.20	22.58	36.75	N/A
Seamless-PU	UNET	VGG16	83.08	67.36	80.45	PU5
			81.06	64.40	78.29	PU6
			79.57	59.62	74.65	PU7
			77.66	57.52	72.99	PU8
			77.09	60.33	75.20	PU9
	DeepLabV2	ResNet-101	84.35	70.46	82.63	PU5
			84.37	70.08	82.37	PU6
			83.67	67.53	80.55	PU7
			74.82	46.47	63.33	PU8
			70.39	38.00	54.90	PU9

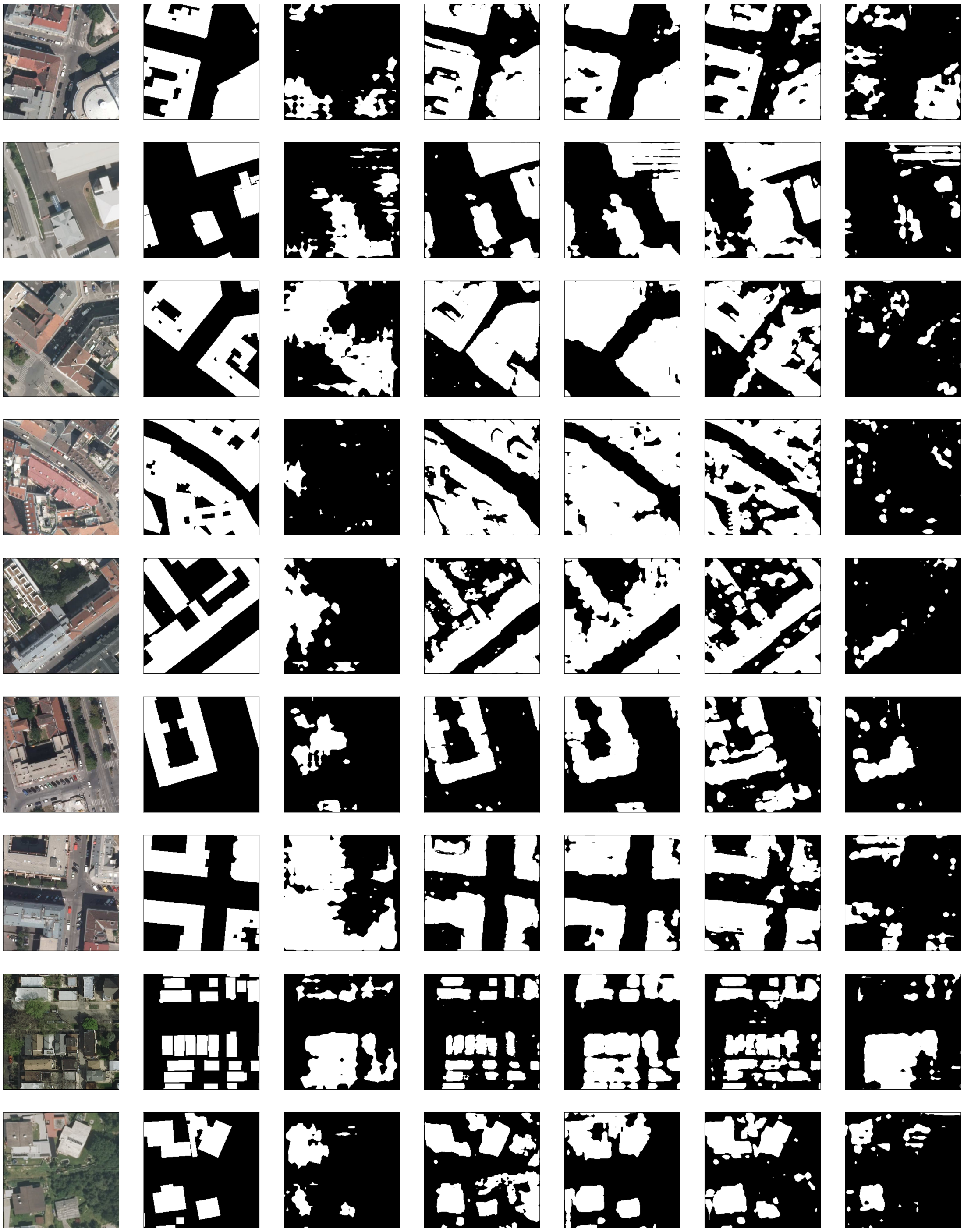


Figure 27. Left-to-right: Sample image patches from heterogeneous PU Inrial Dataset, the corresponding ground truth, and predictions from (i) PixMatch(*Augmentations*) , (ii) Seamless-PU with UNET trained on PU5, (iii) Seamless-PU with DeepLabV2 trained on PU5, (iv) Seamless-PU with UNET trained on PU9, (v) Seamless-PU with DeepLabV2 trained on PU9.

4.4.2.1 Ablation Study. I tested the effect of consistency-training module through its amount of contribution to the proposed Seamless-PU model using $\lambda_{consistency}$. Therefore, using different values for $\lambda_{consistency}$, I investigate different levels of trade-off for focusing on the heterogeneity of the data within the learning process. Table 6 shows the results for $\lambda_{consistency} =$ 0.05, 0.10, 0.15, 0.20, 0.25, 0.5, and 1.0 for both UNET and DeepLabV2 architectures for PU9 dataset. The behavior of $\lambda_{consistency}$ is not consistent across the two architectures. $\lambda_{consistency} = 0.1$ shows the best result for DeepLabV2 which is still lower than the worst performing value for $\lambda_{consistency}$ for UNET. The best case for UNET case is when consistency training module fully contributes to the learning process(i.e. $\lambda_{consistency} = 1.0$).

Table 6. The effect of the degree of the magnitude that the consistency loss is incorporated, which is shown for the case of PU9 dataset.

		$\lambda_{consistency}$						
Architecture Backbone	Performance Metric	0.05	0.10	0.15	0.20	0.25	0.50	1.00
DeepLabV2 ResNet-101	Acc	70.33	71.77	72.26	70.35	72.00	67.44	70.39
	IoU	37.37	42.24	41.79	36.21	41.10	27.78	38.00
	F1	54.32	59.30	58.85	53.04	58.16	43.33	54.90
UNET VGG16	Acc	72.06	72.55	75.59	72.68	69.58	72.41	77.09
	IoU	46.40	44.78	49.89	44.35	42.47	42.89	60.33
	F1	63.32	61.75	66.45	61.34	59.54	59.87	75.20

4.4.2.2 ***Conclusions.*** In this section, I presented the results of the proposed positive and unlabeled domain adaptation model when taking into account the heterogeneity of the data within the training process. I demonstrated that, even in the hard case of heterogeneous scenario, the proposed approach performs well and outperforms the UDA models that target such scenarios. The incorporation of consistency training loss has shown promising results. However, it yielded relatively weaker results for more complex architectures for PU datasets with relatively lower amount of labeled data for the positive class. Therefore, there is a need to further investigate the relationship between the consistency training module and the complexity of the model architecture.

4.5 Conclusion

The proposed model in both homogenous and heterogenous settings outperforms the state-of-the-art baselines while (i) it does not impose additional computational burdens such as style-transfer modules in adversarial approaches, (ii) it leverages relatively light-weight architecture and backbone, (iii) it outperforms the stand-alone PU learning and has potentials for multi-domain learning, (iv) it performs twice as well as UDA models with only a fraction of labeled data from the positive class, and (v) it allows PU learning to experience performance improvement by taking advantage of the other available fully labeled datasets through transfer learning. Finally, this work can be considered as multi-source multi target domain adaptation without the hassle of introducing different networks/encoders for each domain (Isobe et al., 2021) and/or an extra domain alignment module that generates intermediate adapted domains explicitly (S. Zhao, Li, Xu, & Keutzer, 2020).

CHAPTER V

DEEP ENSEMBLE POSITIVE AND UNLABELED LEARNING

The performance of machine learning models may degrade if they fail to capture the underlying structure within data (Dong et al., 2020a). Such failure is more probable when labeled data are not available, such as in the case of positive and unlabeled (PU) data. Therefore, this can cause degradation in the performance of PU models, compared to fully-supervised models. Furthermore, defining the hypothesis space and selecting algorithms that search the defined hypothesis space can introduce inductive biases (Utgoff, 1986, 2012), which can result in machine learning models failing to learn the problem, and thus failing to generalize. One approach to address these two problems is ensemble learning (Dong et al., 2020a; Opitz & Maclin, 1999), which aims to leverage the collective predictive power of multiple different models in order to achieve a better predictive performance compared to any of the participating individual models, alone (Sagi & Rokach, 2018b).

Recently, ensemble learning has gained attention in PU learning research in different areas—for example Nguyen et al. (2012), P. Yang et al. (2014), P. Yang et al. (2016), Claesen et al. (2015), Jowkar and Mansoori (2016), and Basile et al. (2019). However, despite the effectiveness and popularity of ensemble learning in remote sensing research—see Du et al. (2012) for a survey on such methods—little research has been done on ensemble PU learning in remote sensing (e.g. R. Liu et al., 2018, is one of the few).

Furthermore, there is a need to investigate the fusion of ensemble learning within deep learning methodology in remote sensing research since the effectiveness and power of such fusion has been shown in deep learning research (Fort, Hu, &

Lakshminarayanan, 2019) even in a limited data regime (Brigato & Iocchi, 2021). Research such as Ekim and Sertel (2021) has realized such needs in remote sensing research and investigated some native ensemble approaches developed within deep learning framework for supervised RS image classification. However, there is a dearth of research in the area of ensemble PU learning for RS imageries, in general, and RS image segmentation, in particular. Therefore, in this chapter, I investigate the performance of deep learning-based ensemble methodologies for PU learning of RS imageries, and then I propose an ensemble PU learning model, which outperforms all the other models. Finally, I discuss the limitations of the proposed model and future research opportunities.

5.1 Background

Despite the great performance of deep learning models in computer vision tasks, due to a positive inductive bias through utilization of convolutional neural networks (CNNs) (Cohen & Shashua, 2016; LeCun, Bengio, et al., 1995), the convergence of such models to a global minimum may never happen (G. Huang et al., 2017). The impossibility of converging to a global minimum is due to many other inductive biases such as: increasing the number of model parameters and model complexity (Wasay & Idreos, 2020), incomplete knowledge of inductive bias of convolutional operations (Cohen & Shashua, 2016; Wasay & Idreos, 2020), and optimization techniques (Dauphin et al., 2014). Although researchers have sought to understand such inductive biases, such as pooling schemes in CNNs (Cohen & Shashua, 2016) and Stochastic Gradient Descent (SGD) (Dauphin et al., 2014), there are many more parameters embedded in training deep (CNN-based) learning models, making their inductive bias hard to quantify. As an example of such research, it has been shown that as the number of model parameters increases, the number of local minima grows exponentially (Cohen & Shashua, 2016; Kawaguchi, 2016). However, the good news is that not all local minima have an adverse affect and deep learning models can still be generalizable (Keskar, Mudigere, Nocedal, Smelyanskiy, & Tang, 2016). For this reason, deep learning models still perform well even with converging to local optima. In addition, models that converge to different local optima with similar error rates can have different prediction errors, and thus an ensemble of diverse collection of such models converging to different local optima can result in a reduction in the final error rates (Cohen & Shashua, 2016; Fort et al., 2019). In such situations, the ensemble learner outperforms each of the participating individual models (Ekim & Sertel, 2021).

Deep ensemble learning approaches can be categorized into two main streams: (i) those that attempt to learn a single model through ensembling the model parameters in the training phase, and (ii) those that attempt to learn an ensemble of models and take advantage of multi-modal optimization.

There are different approaches considered for ensembling model parameters such as Dropout (G. E. Hinton, Srivastava, Krizhevsky, Sutskever, & Salakhutdinov, 2012), Snapshot Ensemble (G. Huang et al., 2017), Fast Geometric Ensemble (Garipov, Izmailov, Podoprikhin, Vetrov, & Wilson, 2018), Stochastic Weight Averaging (Izmailov, Podoprikhin, Garipov, Vetrov, & Wilson, 2018), and Exponential Moving Average (Tarvainen & Valpola, 2017). Dropout, by G. E. Hinton et al. (2012); Srivastava, Hinton, Krizhevsky, Sutskever, and Salakhutdinov (2014), is originally introduced as a regularization technique to prevent model overfitting, which is common in deep neural networks. Helmbold and Long (2017) show that dropout can result in a better model complexity/performance balance since the penalty by dropout can grow exponentially as the number of hidden layers increases while the penalty by other traditional regularizers, such as L_2-norm, grow linearly. In addition, they show that dropout is scale invariance with respect to, for example, model parameters resulting in the possibility of finding any local optima. Finally, Warde-Farley, Goodfellow, Courville, and Bengio (2013) introduce dropout as an ensemble learning technique, which can be similar to bagging or boosting—such behavior of dropout has been then further studied in different research such as Z. Zhang, Dalca, and Sabuncu (2019). As far as the other approaches, G. Huang et al. (2017) propose a single training multi-model learning approach called Snapshot Ensemble (SE). The authors show that, using a cyclic cosine annealing schedule for learning rate, a model can scape multiple local optima while visiting them properly along its optimization path and saving the corresponding model parameters at each local optimum along the way. Therefore, at the end of a single training, there are multiple saved models, each of which correspond to a different local optimum. Then, an ensemble of these models is used to produce the final predictions. Garipov et al. (2018) propose Fast Geometric Ensemble (FGE), which is based on the same logic as SE. However, it utilizes a linear piecewise cyclical learning rate instead. They also show that deep neural networks' local optima are connected with a path such that the train and test loss stay low while moving from one local optimum to another, and thus such paths can be utilized for an ensemble. Izmailov et al. (2018) propose Stochastic Weight Averaging (SWA) that is an improvement on FGE in that SWA can approximate FGE while it only requires one model at the test time. In comparison to FGE, which averages different models' predictions, SWA averages the weights of such models. More specifically, first, they train a model in a conventional manner using either the full or a proportion of the number of epochs to create the initial model parameters. Then, the training continues using a cyclical learning rate for traversing multiple local optima for each of the corresponding model parameters that are captured—for constant learning rate, model parameters are captured at every epoch. At the end, the average of all these captured model parameters constitutes the final model parameters. Ekim and Sertel (2021) investigate all three SE, FGE, and SWA for image classification of RS imagery, and show that SWA outperforms the other two and the non-ensemble baseline. Finally, Tarvainen and Valpola (2017) propose an improvement on Temporal Ensemble (Laine & Aila, 2016), in which model weights are averaged using the Exponential Moving Average (EMA) method over epochs during the training phase.

The model ensemble approaches can be categorized into Knowledge Distillation (discussed in Chapter IV) (Tarvainen & Valpola, 2017), TreeNets (S. Lee, Purushwalkam, Cogswell, Crandall, & Batra, 2015), AdaBoosts (Mosca & Magoulas, 2017), and MotherNets (Wasay, Hentschel, Liao, Chen, & Idreos, 2020). S. Lee et al. (2015) propose TreeNets as a spectrum of ensembles ranging between single models and non-parameter sharing ensembles of multiple models. Within the spectrum, it is possible to have models that share some layers at the beginning of the network and then diverge from each other to create different classification heads resulting in different outputs to be averaged for the final output. Mosca and Magoulas (2017) propose using AdaBoost for CNN-based models such that the learned model parameters of the previous round of sub-training are transferred to be used as a part of the extended model in the next round. MotherNets, by Wasay et al. (2020), try to reduce the training time for an ensemble of individual models. MotherNets capture the structural similarity between a cluster of networks such that the maximum common core structure of models in each cluster is identified as the MotherNet of that cluster. After training the MotherNet, the individual models within the cluster inherit the model parameters from the MotherNet and will be further trained.

Finally, the experiments by Fort et al. (2019) show that models with different random initializations explore the weight space better than other approaches, such as bayesian networks, and thus deep ensembles of different random initializations are able to capture different modes of the space of solutions. Therefore, a deep ensemble with each of its different models initialized differently can capture a multi-modal landscape solution due to each of its members discovering a different local optimum in the solution space, and thus such deep ensemble can have a better prediction performance.

All in all, each of the aforementioned venues of deep ensemble learning has their advantages and disadvantages considering the balance among final performance, training time, and the number of added parameters. In addition, most of these approaches are developed for supervised settings and/or for image classification task. Therefore, I evaluate some of these approaches against a non-ensemble model for semantic segmentation with PU data. The rest of this chapter is organized as follows. In § 5.2, I first introduce the selected ensemble approaches and the baseline against which I evaluate them, then I introduce the proposed approach. In § 5.3, I assess the results, and, finally, I summarize and discuss future work in § 5.4.

5.2 Methodologies

In this section, different approaches of deep ensemble learning are investigated against the performance of a non-ensemble baseline model in a PU learning scenario. Target-PU model (introduced in Chapter IV) is chosen as the baseline. The baseline uses random weights from a Gaussian distribution and does not use a warm start from ImageNet pre-trained weights, since using such weights results in a negative transfer effect—see Chapter IV. Thus, for a fair and better comparison, the ensemble models do not incorporate ImageNet pre-trained weights either. The selected ensemble models for this chapter are dropout, EMA, SWA, feature ensemble, model ensemble, TreeNet, and contextual ensemble. The structure of all of these models are the same as the baseline, except for the contextual ensemble, which has a smaller depth than the baseline model. All models utilize a UNET architecture with VGG16 as the backbone. Finally, based on the lessons learned from these models, I propose a multi-scale ensemble approach that performs the best among these models.

5.2.1 Dropout.

Inspired by the results in Bartolome, Zhang, and Ramaswami (2018), I consider dropout layers in the expansive path of the UNET network such that a dropout with probability $p = 0.1$ is applied after each concatenation of an up-sample and its corresponding feature map from the contracting path.

5.2.2 EMA.

The model's temporal ensemble is done using exponential moving average of model weights over each epoch in the training stage. As shown in equation 5.1, the averaged model weights at time t ($\bar{\theta}_t$) are calculated from the average value of model parameters over $t-1$ times ($\bar{\theta}_{t-1}$) and the current state of values at time t (θ_t). The value for γ is chosen to be equal to 0.999, which converts the simple average to a weighted average.

$$\bar{\theta}_t = \gamma\bar{\theta}_{t-1} + (1-\gamma)\theta_t \tag{5.1}$$

5.2.3 SWA. The SWA trains two models in the training phase. The first model scans the weight space to find different local optima using a cyclical learning rate scheduler—for convenience, I call this model cyclical. After a warm-up stage for the cyclical model, its weights are used as the starting point for the second model which I call the SWA model. The cyclical model continues its contribution to the SWA model until the end of the training phase. The SWA model maintains an exponential moving average (equation 5.2) of its weights and the weights from the cyclical model. The contribution of the cyclical model's weights is controlled and decreased over time by adjusting the parameter α from value 1 to a value very close to 0.

$$\theta_{SWA} = (1-\alpha)\theta_{SWA} + \alpha\theta_{cyclical} \tag{5.2}$$

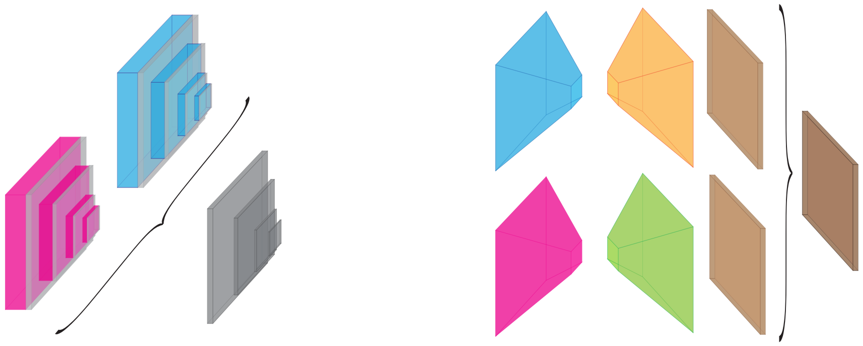
5.2.4 Feature ensemble. The idea is that two (or multiple feature generator) contribute to the same decoder. This means that the ensemble of models happens at feature level instead of the model's output level. Since UNET takes advantage of multiple mid-level features to be used directly in the expansive path, I implement the feature ensemble at the last and all mid feature maps (Fig 28a). The ensemble of feature maps is done using a 1×1 convolution operation.

5.2.5 Model ensemble. In this approach of ensemble, multiple random initializations are used for training multiple models, and then their outputs are collectively used in order to ensure diversity (Fig. 28b). This kind of ensemble is expensive in terms of training time. However, it may result in the best performance, as suggested by Fort et al. (2019). I investigate two 2-model ensembles and one 3-model ensemble with (i) a random initialization strategy, and (ii) an average method for constructing the final output of the models' ensemble from the output of individual models.

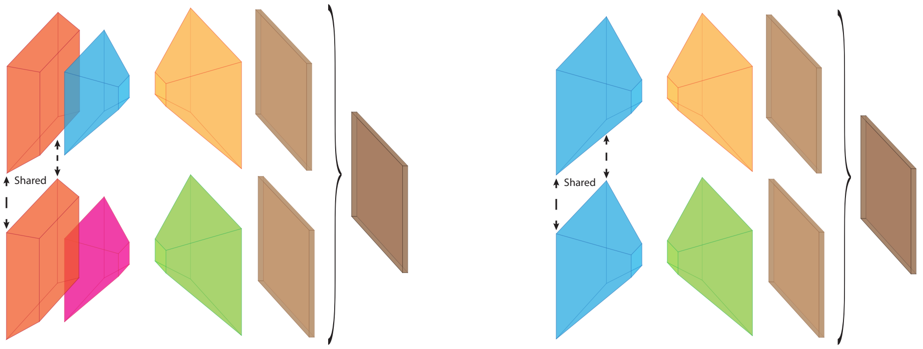
5.2.6 TreeNet. TreeNets are a spectrum between a non-ensemble single model and model-ensemble. I investigate two different ways of implementing TreeNets: one that consider the divergence to happen at encoder level (Fig. 28c), and the other one that consider the divergence to happen at decoder level (Fig. 28d).

5.2.7 Contextual ensemble. Different research such as Marmanis et al. (2016), Z. Zhou, Siddiquee, Tajbakhsh, and Liang (2019), Ma, Li, Zhang, Tang, and Guo (2021), and L. Chen et al. (2021) have been trying to provide ensemble networks by modifying the structure of UNET network. Most recently, Q. Zhou et al. (2022) propose a UNET-based ensemble for semantic segmentation called contextual ensemble network (Fig. 28e) which fully explores contextual features by modifying the expansive module of the UNET network. At each level within the expansive module, a stack of multi-scale feature representations from previous stages are concatenated to the stack of upsampled features and feature maps copied from the contracting module. This process aims to combine feature maps created with different receptive fields and thus allows harvesting and leveraging multi-scale context clues.

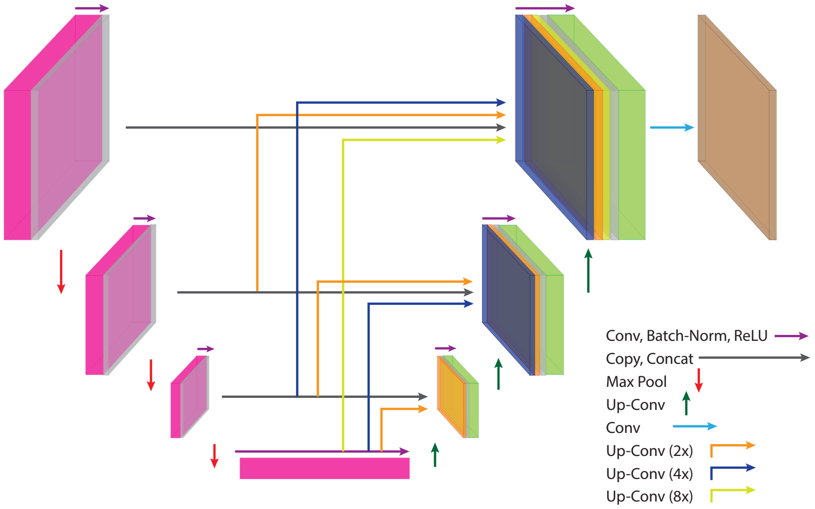
5.2.8 Multi-scale ensemble. The idea of feature sharing and ensemble of multiple predictions is not new. Feature Pyramid Network (FPN)



(a) Feature ensemble. (b) Model ensemble.



(c) TreeNet ensemble at Encoder. (d) TreeNet ensemble at Decoder.



(e) Contextual ensemble.

Figure 28. Different deep ensemble architectures.

by T.-Y. Lin et al. (2017) proposes using feature maps at different scales for producing multiple predictions to be combined for the final prediction. Such final prediction performs better than a prediction based solely on a single feature map output of a feature generator network. FPN is originally for image classification with a brief extension idea for image segmentation within the original paper. Following on FPN, different research such as Tao, Sapra, and Catanzaro (2020) and Bousselham et al. (2021) have been trying to provide an ensemble learning framework for semantic segmentation. However, such models usually incorporate complex modules such as different variations of attention and transformer modules (Z. Liu et al., 2021; Vaswani et al., 2017). Although successful, these models are developed for fully supervised learning. However, as shown in Chapter IV in the case of UNET-VGG16 versus DeepLabV2-ResNet101, the high complexity of the learning model becomes harmful rather than beneficial as the proportion of labeled data from the positive class decreases (i.e. from PU5 to PU9). This is important since it is desirable to have a model that performs well with the least amount of available labeled data from the positive class. In addition such complex models require using pre-trained weights for optimal performance, whereas pre-trained weights cause negative transfer in the case of PU data (which is shown in Chapter IV). Considering these, I propose a model that is based on three elements: (i) the successful lightweight UNET structure for PU learning that is explored in Chapter IV, (ii) the multi-scale multi-prediction idea by FPN, and (iii) the multi local optima exploration idea by SWA model. The idea of using SWA model as the third element comes from the results by Fort et al. (2019) showing that mixing multiple ensemble approaches can result in even more improvements in the prediction performance.

As shown in Fig. 29, in addition to the final layer created by a series of upsamples using deconvolution operations, the two previous layers in the expansive module are used to create two additional prediction segmentation maps. These two additional layers are upsampled to create an output with the same size as the original output. The upsampling is done by a bilinear operation (rather than a deconvolution operation) in order to (i) keep the model complexity low and (ii) maintain the information at each layer without deforming it. The final prediction for each of these two additional upsampled layers is produced by applying a 1 × 1 convolution, which is the same way as for the original prediction output. Finally, all three prediction segmentation maps are averaged to produce the final prediction segmentation map.

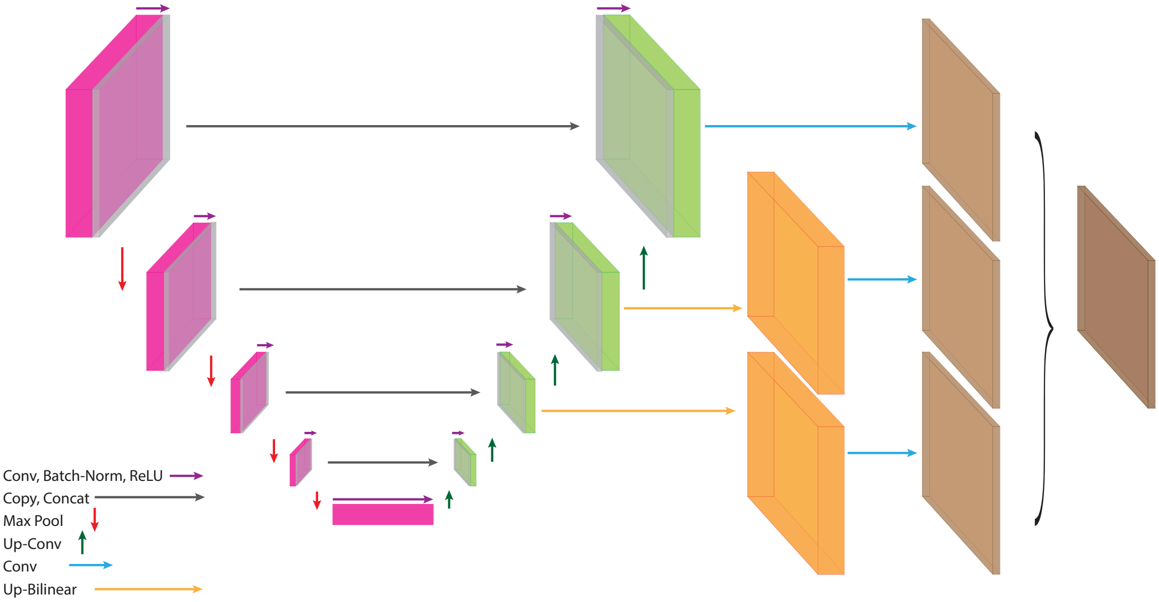


Figure 29. self-ensemble multi-scale UNET.

5.3 Results

The performance of models in this section is evaluated using the PU dataset from the homogeneous dataset created from Inria Image Dataset, as discussed in Chapter III. The loss function for all models is NNPU loss, and the same evaluation metrics as in Chapter IV are used here. Target-PU model (introduced in Chapter IV) is chosen as the non-ensemble baseline, and is initiated using random weights from a Gaussian distribution. Dropout, EMA, and SWA models are also initiated using random weights from a Gaussian distribution. Encoders in feature ensemble model are initiated using random weights from a Gaussian distribution and Xavier uniform distribution, respectively. This is the same for TreeNet models as well. In the case of model ensemble, there are two 2-model ensembles and a 3-model ensemble. In each of the 2-model ensembles, one model is initialized using random weights from a Gaussian distribution and the other model is initialized using random weights from either Xavier uniform distribution or Kaiming uniform distribution, respectively. In 3-model ensemble, Gaussian, Xavier uniform, and Kaiming uniform distributions are used for each of the three models' weights initializations. Finally, contextual ensemble and the proposed multi-scale model are initiated using random weights from a Gaussian distribution. The optimizer and learning rate scheduler hyper-parameters are the same as in Chapter IV for all models, except for the SWA model. For SWA model, I use the configuration set that the original paper (Izmailov et al., 2018) suggests.

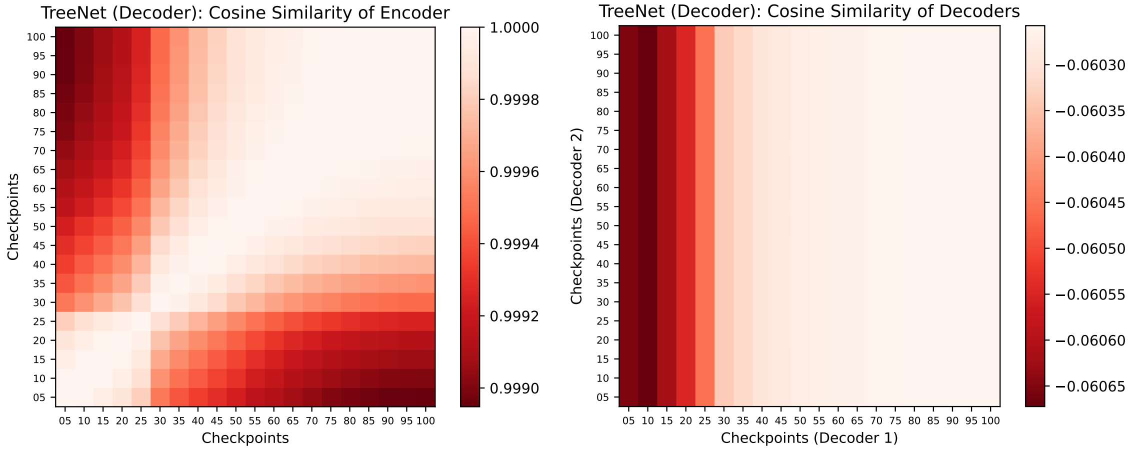
According to Table 7, excluding the proposed model, SWA model performs the best and outperforms the baseline. After that, model ensemble has the second place with outperforming the baseline in some cases such as PU9 dataset. Although Fort et al. (2019) show that using ensemble of models with different random initializations can result in better performances, it can be seen that such improvements are not always the case and not guaranteed for PU data. TreeNet models surprisingly perform very poorly, given that the model ensemble does well.

Since model ensemble is on the extreme end of the spectrum of TreeNets model architectures, I further investigate the performance gap between TreeNets and model ensemble by analyzing the changes in models' parameters during the training phase using the cosine similarity metric.

The cosine similarity of two vectors projected in a multi-dimensional space is the cosine of the angle between the two vectors. Therefore, it can be used to measure the alignment of the two vectors. According to Fort et al. (2019), cosine similarity measures weight space alignment along optimization trajectory, and thus each of the two aforementioned vectors contain the parameters (i.e. weights) of two models that are analyzed. I calculate the cosine similarity using the equation 5.3 at every five checkpoints during the training phase, where θ_i is the parameter set for model i. The value of the cosine similarity of two vectors ranges from -1 to 1, where -1, 0, and 1 indicate strongly opposite, independent, and similar vectors, respectively.

$$cos(\theta_1, \theta_2) = \frac{\theta_1^T \theta_2}{||\theta_1||||\theta_2||} \tag{5.3}$$

Although the shared part within the TreeNet models explores the weight space pretty well (Fig. 30a and Fig. 31a), it controls how and to what extent separated branches within the models would explore the weight space. As shown in Fig. 30b and Fig. 31b, if the shared part of the network in a TreeNet model reduces, there is a higher chance for the separate branches to explore different areas within the weight space. For example, when the separation point is at decoder level, since the feature map generation is done similar for the two networks, the decoders tend to end up exploring the same area within the weight space. Whereas, when the separation point is within the encoder, the two models can drift away from each other within the weight space, which gives them a higher chance of finding two local minima, such that their ensemble performs better. This is also the case when the models are completely separated. For example Fig. 32 compares two models with Xavier and Kaiming random weight initializations within the 3-model ensemble network. Although the two models start within the similar neighborhood within the weight space, they drift apart during the training phase, which again gives them a higher chance of finding two local minima, such that their ensemble performs better.



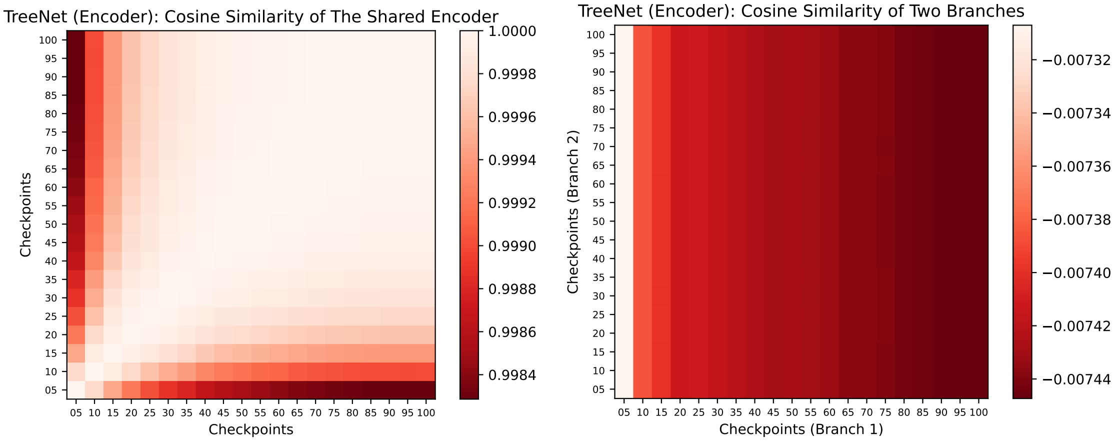
(a) The shared part.

(b) The unshared part.

Figure 30. Cosine similarity of model weights for TreeNet (Decoder) model over different checkpoints at different epochs during the training phase for PU9 dataset.

Table 7. The performance of different ensemble models compared to the performance of a single model.

		PU5			PU6			PU7			PU8			PU9		Best
Method	Acc	IoU	F1	Acc	IoU	F1	Acc	IoU	F1	Acc	IoU	F1	Acc	IoU	F1	Case
Baseline (non-ensemble)	77.69	63.45	77.57	81.06	67.46	80.49	81.52	66.95	80.13	80.75	66.37	79.69	81.38	66.59	79.87	PU6
Dropout	77.09	62.91	77.20	80.40	66.49	79.85	81.27	66.85	80.10	81.13	66.60	79.93	80.87	65.67	79.25	PU7
EMA	40.94	40.94	58.07	43.33	39.85	56.95	42.48	38.90	55.98	43.13	39.63	56.73	43.20	39.71	5682	PU5
SWA	77.61	63.75	77.79	41.85	41.85	58.91	40.50	40.49	57.54	82.54	68.88	81.49	83.13	69.50	81.95	PU9
Feature Ensemble	57.69	07.84	14.53	50.81	27.69	43.34	40.55	40.44	57.48	42.13	40.32	57.35	48.76	31.48	47.85	PU7
2-Model Ensemble (Xavier)	76.72	62.67	76.98	79.84	66.28	79.64	81.32	66.84	80.05	80.51	66.39	79.67	81.20	66.93	80.12	PU9
2-Model Ensemble (Kaiming)	75.89	61.76	76.29	80.34	66.59	79.88	80.57	65.90	79.38	80.35	66.12	79.48	81.24	66.71	79.95	PU9
3-Model Ensemble	77.03	62.90	77.16	80.87	67.22	80.33	81.46	66.92	80.11	80.96	66.83	79.98	81.92	67.67	80.64	PU9
TreeNet (Encoder)	50.12	28.65	44.51	57.95	01.21	02.39	48.18	31.70	48.09	41.61	40.91	57.94	57.94	03.62	06.98	PU8
TreeNet (Decoder)	52.00	25.02	40.01	51.89	25.12	40.13	53.29	22.16	36.26	49.21	30.62	46.83	49.51	30.20	46.36	PU8
Contextual Ensemble	77.18	60.04	74.98	77.16	58.89	74.02	78.38	59.33	74.37	75.00	54.87	70.76	78.55	60.19	74.91	PU9
Proposed Model	79.58	65.83	79.34	82.77	69.90	82.21	82.00	68.05	80.91	82.78	69.17	81.68	83.93	70.18	82.42	PU9



(a) The shared part.

(b) The unshared part.

Figure 31. Cosine similarity of model weights for TreeNet (Encoder) model over different checkpoints at different epochs during the training phase for PU9 dataset.

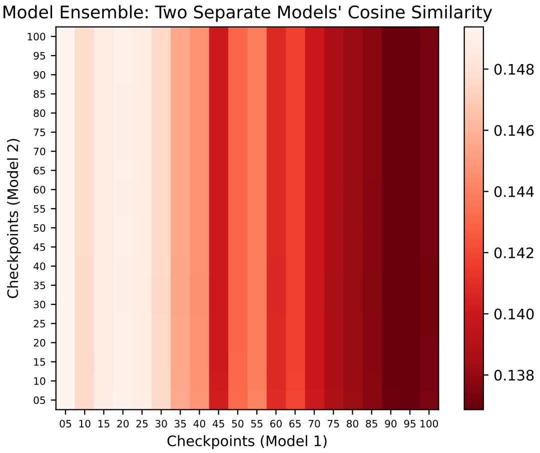


Figure 32. Cosine similarity of model weights for ModelNet model over different checkpoints at different epochs during the training phase for PU9 dataset.

The contextual ensemble performs poorly since there are so many convolution operations added to the expansive module of the network. I investigated other options that decrease such complexity such as bilinear upsampling. This resulted in improved model performance; however, the baseline model still shows superior performance. I also tried the proposed multi-scale ensemble model without weight sharing using three different models with Gaussian, Xavier uniform, and Kaiming uniform distributions for each of the three models' weights initializations. However, this approach resulted in a minor decrease in prediction performance of the model. For example, the IoU for PU9 drops from 70.18% to 70.03%. Fig. 33 shows a sample of image patches from homogeneous PU Inrial Dataset and the corresponding ground truth as well as predictions from the baseline model and the proposed model trained on PU9 dataset.

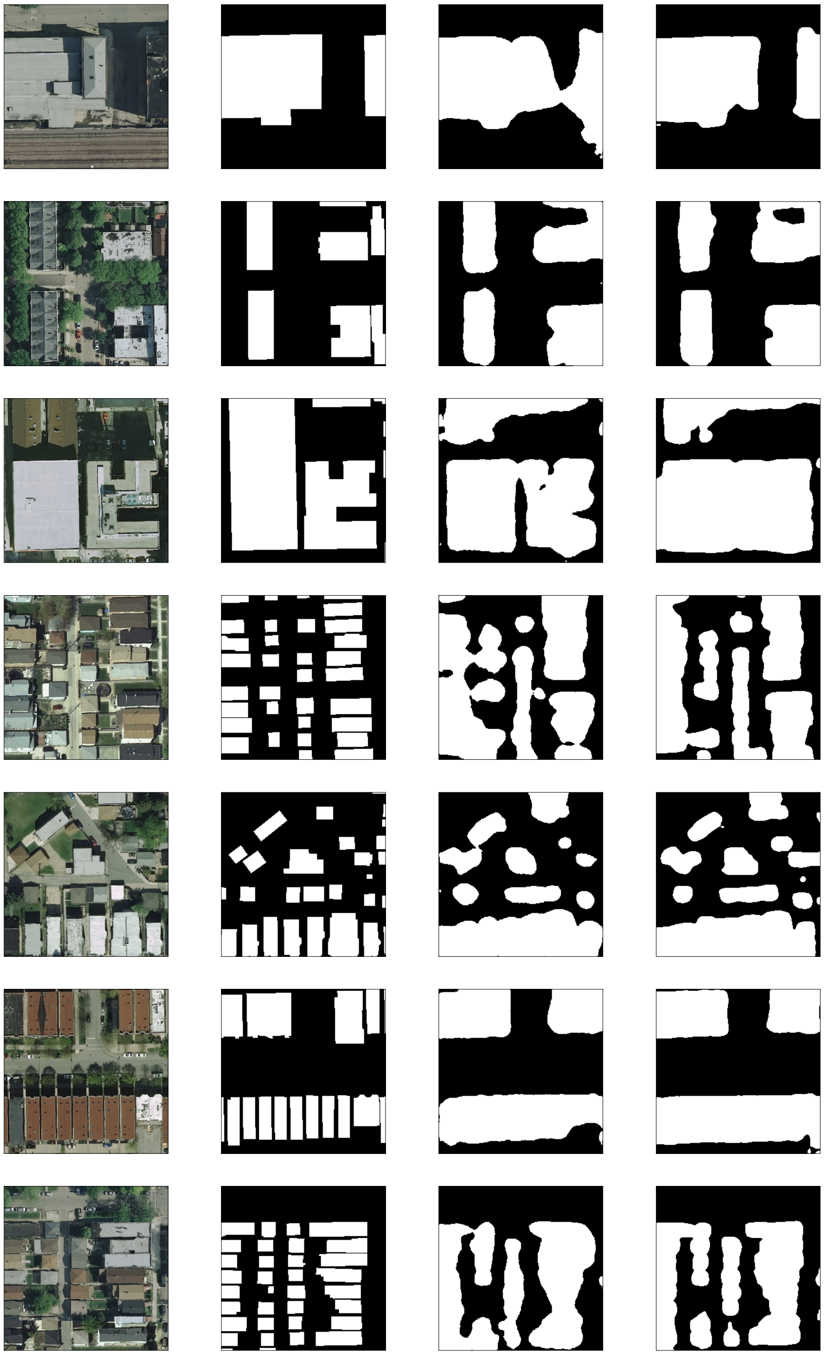


Figure 33. Left-to-right: Sample image patches from homogeneous PU Inrial Dataset, the corresponding ground truth, and predictions from the baseline model and the proposed model trained on PU9 dataset.

Finally, since the proposed ensemble model is an end-to-end learning and compatible with the network structure and the learning schema proposed for PU domain adaptation, it is possible to create a mixture of both in order to further extend the performance improvement. As shown in Table 8, the mixture of the proposed model for ensemble and transfer learning performs the best for PU5, PU7, and PU9 datasets, whereas its performance of PU6 and PU8 are weaker. What this has all shown is the superior performance of the proposed model on my most challenging dataset, that is PU9 dataset with the minimum amount of available labeled data from positive class. However, further research is needed to investigate the unsatisfying performance of the model on PU6 and PU8 datasets.

5.4 Conclusion

In this chapter, I investigated most of the available ensemble approaches developed for deep learning frameworks. However, they are mostly developed for image classification tasks and supervised learning framework. I demonstrated that these methods—except for SWA and model ensemble—do not perform well in the case of PU data for semantic segmentation of RS imageries. Next, I proposed a lightweight end-to-end ensemble approach that outperforms all the other models and shows promising results for creating a mixture with the proposed domain adaptation model for PU learning.

Table 8. The mixture of Ensemble and Transfer PU models.

Dataset	Performance Metric	Seamless-PU	Ensemble Seamless-PU
PU5	Acc	88.85	89.54
	IoU	76.63	77.78
	F1	86.71	87.44
PU6	Acc	89.21	84.78
	IoU	77.21	68.91
	F1	87.09	81.53
PU7	Acc	89.19	89.54
	IoU	76.54	77.08
	F1	86.66	87.00
PU8	Acc	88.92	88.42
	IoU	76.41	75.84
	F1	86.58	86.17
PU9	Acc	88.58	89.00
	IoU	76.01	76.63
	F1	86.21	86.67

CHAPTER VI

CONCLUSIONS & FUTURE WORK

Remotely-sensed (RS) imageries are key elements in research in many fields of study. The large amount of RS imageries with high spatial and temporal resolution that are produced frequently require machine learning algorithms for automatic information extraction. The labor-intensive and time-consuming nature of creating a set of representative labeled training samples, along with frequent applications with interests in identifying only one specific landcover or object from RS imageries, call for incorporating positive and unlabeled (PU) learning methodology in remote sensing research. The research presented in this dissertation is among the first that addresses the problem of PU learning in geospatial domain, in general, and in remote sensing, in particular. As the occurrence of PU data is inevitable and much of a reality in geospatial applications, this research aims to connect the research in machine learning and remote sensing communities. In this dissertation, I investigated the two research questions:

RQ1. *How can Transfer Learning be incorporated in the context of positive and unlabeled learning for semantic segmentation of satellite imagery?*

RQ2. *How can Ensemble Learning be incorporated in the context of positive and unlabeled learning for semantic segmentation of satellite imagery?*

In summary, my research demonstrated that the naive mixture of existing PU learning models with transfer learning paradigms does not result in acceptable model performance. To address this problem, I developed a method that can overcome the negative transfer effect for PU learning. In addition, I showed that simply adopting methods developed in computer vision, including UDA models, may not perform as well as they do in other categories, such as urban scene semantic segmentation. The proposed model shows promising results with limited amount of labeled data from the positive class. Finally, the proposed model can be seen as multi-target domain since each city is a separate domain within the dataset. This makes the proposed model more powerful since it can cope with the changes in landcover structures at different geographic locations—for example, the change in material used for rooftop of the buildings at different locations. Also, since there is no constraint on the number of source domains, the proposed model can easily be extended to the multi-source domain case as well.

Next, to address the ensemble PU learning framework, first I investigated the performance of the available ensemble learning methodologies proposed for deep learning. Most of these approaches are developed for fully supervised settings and address image classification task rather than semantic segmentation. I demonstrated that, again, simply adopting such models may not produce competitive results in comparison to non-ensemble PU models. By learning from the behavior of such models, I proposed an ensemble PU learning model, which outperforms all discussed models and shows promising results. The proposed deep ensemble PU learning is general and end-to-end, and thus it can be easily incorporated with other deep learning methods, such as the proposed deep transfer PU learning.

In summary, the results from my research delivers an expansion of the category of positive and unlabeled learning methods. Such deliverable is a valuable tool to be utilized for different real-world problems needing RS imagery segmentation, for which collecting a comprehensively labeled dataset is costly.

6.1 Future Work

The results of my dissertation research, illuminated two areas of future work. First, I would like to investigate relaxing the SCAR assumption in generating PU data for research, and thus relaxing this assumption in the PU loss function. This is because it is more realistic that users are biased towards buildings/objects that they can recognize better when creating distinct patches of labeled pixels within an image. Second, it is also more realistic that a group of classes together constitute the positive class which falls under multi-positive and unlabeled (multi-PU) learning. For example, this is the case in remote sensing for agriculture, in which multi-class methods are preferred to identify a set of crops of interest. There is some research in this area in computer vision literature such as Xu, Xu, Xu, and Tao (2017), Shu, Lin, Yan, and Li (2020), and Teisseyre (2021) that address multi-PU learning. Therefore, in the future, I to extend my research further in the context of RS imageries in order to extend the proposed models in this dissertation to the multi-PU case.

REFERENCES CITED

Abuduweili, A., Li, X., Shi, H., Xu, C.-Z., & Dou, D. (2021). Adaptive consistency regularization for semi-supervised transfer learning. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 6923–6932).

Ahn, J., & Kwak, S. (2018). Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 4981–4990).

Alonso, I., Sabater, A., Ferstl, D., Montesano, L., & Murillo, A. C. (2021). Semi-supervised semantic segmentation with pixel-level contrastive learning from a class-wise memory bank. In *Proceedings of the ieee/cvf international conference on computer vision* (pp. 8219–8228).

Asgarian, A., Sobhani, P., Zhang, J. C., Mihailescu, M., Sibilia, A., Ashraf, A. B., & Taati, B. (2018). A hybrid instance-based transfer learning method. *Machine Learning for Health (ML4H) Workshop at NeurIPS*.

Balestriero, R., Bottou, L., & LeCun, Y. (2022). The effects of regularization and data augmentation are class dependent. *arXiv preprint arXiv:2204.03632*.

Bartolome, C., Zhang, Y., & Ramaswami, A. (2018). Deepcell: Automating cell nuclei detection with.

Bashath, S., Perera, N., Tripathi, S., Manjang, K., Dehmer, M., & Streib, F. E. (2022a). A data-centric review of deep transfer learning with applications to text data. *Information Sciences*, *585*, 498–528.

Bashath, S., Perera, N., Tripathi, S., Manjang, K., Dehmer, M., & Streib, F. E. (2022b). A data-centric review of deep transfer learning with applications to text data. *Information Sciences*, *585*, 498–528.

Basile, T. M. A., Di Mauro, N., Esposito, F., Ferilli, S., & Vergari, A. (2019). Ensembles of density estimators for positive-unlabeled learning. *Journal of Intelligent Information Systems*, *53*(2), 199–217.

Bekker, J., & Davis, J. (2020). Learning from positive and unlabeled data: A survey. *Machine Learning*, *109*(4), 719–760.

Bekker, J., Robberechts, P., & Davis, J. (2019). Beyond the selected completely at random assumption for learning from positive and unlabeled data. In *Joint european conference on machine learning and knowledge discovery in databases* (pp. 71–85).

Benediktsson, J. A., Chanussot, J., & Fauvel, M. (2007). Multiple classifier systems in remote sensing: from basics to recent developments. In *International workshop on multiple classifier systems* (pp. 501–512).

Berthelot, D., Carlini, N., Cubuk, E. D., Kurakin, A., Sohn, K., Zhang, H., & Raffel, C. (2019). Remixmatch: Semi-supervised learning with distribution alignment and augmentation anchoring. *arXiv preprint arXiv:1911.09785*.

Bhat, S., & Culotta, A. (2017). Identifying leading indicators of product recalls from online reviews using positive unlabeled learning and domain adaptation. In *Proceedings of the international aaai conference on web and social media* (Vol. 11, pp. 480–483).

Bousselham, W., Thibault, G., Pagano, L., Machireddy, A., Gray, J., Chang, Y. H., & Song, X. (2021). Efficient self-ensemble framework for semantic segmentation. *arXiv preprint arXiv:2111.13280*.

Brigato, L., & Iocchi, L. (2021). On the effectiveness of neural ensembles for image classification with small datasets. *arXiv preprint arXiv:2111.14493*.

Brill, F., Schlaffer, S., Martinis, S., Schröter, K., & Kreibich, H. (2021). Extrapolating satellite-based flood masks by one-class classification—a test case in houston. *Remote Sensing*, *13*(11), 2042.

Chakraborty, S., & Roy, M. (2020). A multi-level weighted transformation based neuro-fuzzy domain adaptation technique using stacked auto-encoder for land-cover classification. *International Journal of Remote Sensing*, *41*(17), 6831–6857.

Chawla, N. V., Japkowicz, N., & Kotcz, A. (2004). Special issue on learning from imbalanced data sets. *ACM SIGKDD explorations newsletter*, *6*(1), 1–6.

Chen, H., Liu, F., Wang, Y., Zhao, L., & Wu, H. (2020). A variational approach for learning from positive and unlabeled data. *34th Conference on Neural Information Processing Systems (NeurIPS), Vancouver, Canada.*.

Chen, J., & Liu, X. (2014). Transfer learning with one-class data. *Pattern Recognition Letters*, *37*, 32–40.

Chen, L., Dou, X., Peng, J., Li, W., Sun, B., & Li, H. (2021). Efcnet: Ensemble full convolutional network for semantic segmentation of high-resolution remote sensing images. *IEEE Geoscience and Remote Sensing Letters*, *19*, 1–5.

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., & Yuille, A. L. (2017). Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, *40*(4), 834–848.

Chen, S., Jia, X., He, J., Shi, Y., & Liu, J. (2021). Semi-supervised domain adaptation based on dual-level domain mixing for semantic segmentation. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 11018–11027).

Chen, X., Gong, C., & Yang, J. (2021). Cost-sensitive positive and unlabeled learning. *Information Sciences*, *558*, 229–245.

Chen, X., Liu, F., Tu, E., Cao, L., & Yang, J. (2018). Deep-pumr: Deep positive and unlabeled learning with manifold regularization. In *International conference on neural information processing* (pp. 12–20).

Chen, Y., Nasrabadi, N. M., & Tran, T. D. (2012). Hyperspectral image classification via kernel sparse representation. *IEEE Transactions on Geoscience and Remote sensing*, *51*(1), 217–231.

Chen, Y.-C., Lin, Y.-Y., Yang, M.-H., & Huang, J.-B. (2019). Crdoco: Pixel-level domain transfer with cross-domain consistency. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 1791–1800).

Chiaroni, F., Rahal, M.-C., Hueber, N., & Dufaux, F. (2018). Learning with a generative adversarial network from a positive unlabeled dataset for image classification. In *2018 25th ieee international conference on image processing (icip)* (pp. 1368–1372).

Claesen, M., De Smet, F., Suykens, J. A., & De Moor, B. (2015). A robust ensemble approach to learn from positive and unlabeled data using svm base models. *Neurocomputing*, *160*, 73–84.

Cohen, N., & Shashua, A. (2016). Inductive bias of deep convolutional networks through pooling geometry. *arXiv preprint arXiv:1605.06743*.

Dauphin, Y. N., Pascanu, R., Gulcehre, C., Cho, K., Ganguli, S., & Bengio, Y. (2014). Identifying and attacking the saddle point problem in high-dimensional non-convex optimization. *Advances in neural information processing systems*, *27*.

Day, O., & Khoshgoftaar, T. M. (2017). A survey on heterogeneous transfer learning. *Journal of Big Data*, *4*(1), 1–42.

Deng, X., Li, W., Liu, X., Guo, Q., & Newsam, S. (2018). One-class remote sensing classification: one-class vs. binary classifiers. *International Journal of Remote Sensing*, *39*(6), 1890–1910.

Desloires, J., Ienco, D., Botrel, A., & Ranc, N. (2022). Positive unlabelled learning for satellite images' time series analysis: An application to cereal and forest mapping. *Remote Sensing*, *14*(1), 140.

Devassy, B. R., & Antony, J. K. (2021). Histopathological image classification using ensemble transfer learning. In *International conference on machine learning and big data analytics* (pp. 203–212).

DeVries, T., & Taylor, G. W. (2017). Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*.

Dey, V., Zhang, Y., & Zhong, M. (2010). *A review on image segmentation techniques with remote sensing perspective* (Vol. 38). na Vienna, Austria.

Dhurandhar, A., & Gurumoorthy, K. S. (2020). Classifier invariant approach to learn from positive-unlabeled data. In *2020 ieee international conference on data mining (icdm)* (pp. 102–111).

Djerriri, K., Benyelles, Z., Attaf, D., & Cheriguene, R. S. (2019). Extraction of built-up areas from remote sensing imagery using one-class classification. In L. Bruzzone & F. Bovolo (Eds.), *Image and signal processing for remote sensing xxv* (Vol. 11155, pp. 610 – 616). SPIE. Retrieved from `https://doi.org/10.1117/12.2535598` doi: 10.1117/12.2535598

Dong, X., Yu, Z., Cao, W., Shi, Y., & Ma, Q. (2020a). A survey on ensemble learning. *Frontiers of Computer Science*, *14*(2), 241–258.

Dong, X., Yu, Z., Cao, W., Shi, Y., & Ma, Q. (2020b). A survey on ensemble learning. *Frontiers of Computer Science*, *14*(2), 241–258.

Doshi, K., & Yilmaz, Y. (2020). Road damage detection using deep ensemble learning. In *2020 ieee international conference on big data (big data)* (pp. 5540–5544).

Du, P., Xia, J., Zhang, W., Tan, K., Liu, Y., & Liu, S. (2012). Multiple classifier system for remote sensing image classification: A review. *Sensors*, *12*(4), 4764–4792.

Du Plessis, M., Niu, G., & Sugiyama, M. (2015). Convex formulation for learning from positive and unlabeled data. In *International conference on machine learning* (pp. 1386–1394).

Du Plessis, M. C., Niu, G., & Sugiyama, M. (2014). Analysis of learning from positive and unlabeled data. *Advances in neural information processing systems*, *27*.

Du Plessis, M. C., Niu, G., & Sugiyama, M. (2016). Class-prior estimation for learning from positive and unlabeled data. In *Asian conference on machine learning* (pp. 221–236).

Ekim, B., & Sertel, E. (2021). Deep neural network ensembles for remote sensing land cover and land use classification. *International Journal of Digital Earth*, *14*(12), 1868–1881.

Elkan, C., & Noto, K. (2008). Learning classifiers from only positive and unlabeled data. In *Proceedings of the 14th acm sigkdd international conference on knowledge discovery and data mining* (pp. 213–220).

Fan, J., Xu, C., & Zhang, J. (2021). An ensemble learning approach of multi-model for classifying house damage. In *2021 2nd international conference on big data & artificial intelligence & software engineering (icbase)* (pp. 145–152).

Foody, G. M., Mathur, A., Sanchez-Hernandez, C., & Boyd, D. S. (2006). Training set size requirements for the classification of a specific class. *Remote Sensing of Environment*, *104*(1), 1–14.

Fort, S., Hu, H., & Lakshminarayanan, B. (2019). Deep ensembles: A loss landscape perspective. *arXiv preprint arXiv:1912.02757*.

French, G., Aila, T., Laine, S., Mackiewicz, M., & Finlayson, G. (2019). Semi-supervised semantic segmentation needs strong, high-dimensional perturbations.

French, G., & Mackiewicz, M. (2021). Colour augmentation for improved semi-supervised semantic segmentation. *arXiv preprint arXiv:2110.04487*.

Ganaie, M., Hu, M., et al. (2021). Ensemble deep learning: A review. *arXiv preprint arXiv:2104.02395*.

Ganin, Y., & Lempitsky, V. (2015). Unsupervised domain adaptation by backpropagation. In *International conference on machine learning* (pp. 1180–1189).

Garipov, T., Izmailov, P., Podoprikhin, D., Vetrov, D. P., & Wilson, A. G. (2018). Loss surfaces, mode connectivity, and fast ensembling of dnns. *Advances in neural information processing systems*, *31*.

Ghosh, R., Jia, X., & Kumar, V. (2021). Land cover mapping in limited labels scenario: A survey. *arXiv preprint arXiv:2103.02429*.

Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2015). Region-based convolutional networks for accurate object detection and segmentation. *IEEE transactions on pattern analysis and machine intelligence*, *38*(1), 142–158.

Gong, C., Shi, H., Liu, T., Zhang, C., Yang, J., & Tao, D. (2019). Loss decomposition and centroid estimation for positive and unlabeled learning. *IEEE transactions on pattern analysis and machine intelligence*, *43*(3), 918–932.

Gong, C., Wang, D., & Liu, Q. (2021). Alphamatch: Improving consistency for semi-supervised learning with alpha-divergence. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 13683–13692).

Gu, X., Zhang, C., Shen, Q., Han, J., Angelov, P. P., & Atkinson, P. M. (2022). A self-training hierarchical prototype-based ensemble framework for remote sensing scene classification. *Information Fusion*, *80*, 179–204.

Gui, R., Xu, X., Wang, L., Yang, R., & Pu, F. (2020). Eigenvalue statistical components-based pu-learning for polsar built-up areas extraction and cross-domain analysis. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *13*, 3192–3203.

Guo, T., Xu, C., Huang, J., Wang, Y., Shi, B., Xu, C., & Tao, D. (2020). On positive-unlabeled classification in gan. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 8385–8393).

Hammoudeh, Z., & Lowd, D. (2020). Learning from positive and unlabeled data with arbitrary positive shift. *Advances in Neural Information Processing Systems*, *33*, 13088–13099.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 770–778).

He, X., & Chen, Y. (2020). Transferring cnn ensemble for hyperspectral image classification. *IEEE Geoscience and Remote Sensing Letters*, *18*(5), 876–880.

He, X., Chen, Y., & Ghamisi, P. (2019). Heterogeneous transfer learning for hyperspectral image classification based on convolutional neural network. *IEEE Transactions on Geoscience and Remote Sensing*, *58*(5), 3246–3263.

Helmbold, D. P., & Long, P. M. (2017). Surprising properties of dropout in deep networks. In *Conference on learning theory* (pp. 1123–1146).

Hinton, G., Vinyals, O., Dean, J., et al. (2015). Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, *2*(7).

Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. R. (2012). Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580*.

Hoffman, J., Tzeng, E., Park, T., Zhu, J.-Y., Isola, P., Saenko, K., … Darrell, T. (2018). Cycada: Cycle-consistent adversarial domain adaptation. In *International conference on machine learning* (pp. 1989–1998).

Hoffman, J., Wang, D., Yu, F., & Darrell, T. (2016). Fcns in the wild: Pixel-level adversarial and constraint-based adaptation. *arXiv preprint arXiv:1612.02649*.

Hou, M., Chaib-Draa, B., Li, C., & Zhao, Q. (2017). Generative adversarial positive-unlabelled learning. *arXiv preprint arXiv:1711.08054*.

Hu, W., Le, R., Liu, B., Ji, F., Ma, J., Zhao, D., & Yan, R. (2021). Predictive adversarial learning from positive and unlabeled data. In *Proceedings of the aaai conference on artificial intelligence* (Vol. 35, pp. 7806–7814).

Huang, G., Li, Y., Pleiss, G., Liu, Z., Hopcroft, J. E., & Weinberger, K. Q. (2017). Snapshot ensembles: Train 1, get m for free. *arXiv preprint arXiv:1704.00109*.

Huang, X., & Zhang, L. (2012a). An svm ensemble approach combining spectral, structural, and semantic features for the classification of high-resolution remotely sensed imagery. *IEEE transactions on geoscience and remote sensing*, *51*(1), 257–272.

Huang, X., & Zhang, L. (2012b). An svm ensemble approach combining spectral, structural, and semantic features for the classification of high-resolution remotely sensed imagery. *IEEE transactions on geoscience and remote sensing*, *51*(1), 257–272.

Huang, Z., Wang, X., Wang, J., Liu, W., & Wang, J. (2018). Weakly-supervised semantic segmentation network with deep seeded region growing. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 7014–7023).

Hung, W.-C., Tsai, Y.-H., Liou, Y.-T., Lin, Y.-Y., & Yang, M.-H. (2018). Adversarial learning for semi-supervised semantic segmentation. *arXiv preprint arXiv:1802.07934*.

Iqbal, J., & Ali, M. (2020). Weakly-supervised domain adaptation for built-up region segmentation in aerial and satellite imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, *167*, 263–275.

Isobe, T., Jia, X., Chen, S., He, J., Shi, Y., Liu, J., … Wang, S. (2021). Multi-target domain adaptation with collaborative consistency learning. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 8187–8196).

Iyer, P., Sriram, A., & Lal, S. (2021). Deep learning ensemble method for classification of satellite hyperspectral images. *Remote Sensing Applications: Society and Environment*, *23*, 100580.

Izmailov, P., Podoprikhin, D., Garipov, T., Vetrov, D., & Wilson, A. G. (2018). Averaging weights leads to wider optima and better generalization. *arXiv preprint arXiv:1803.05407*.

Jain, S., Delano, J., Sharma, H., & Radivojac, P. (2020). Class prior estimation with biased positives and unlabeled examples. In *Proceedings of the aaai conference on artificial intelligence* (Vol. 34, pp. 4255–4263).

Jamali, A., Mahdianpari, M., Brisco, B., Granger, J., Mohammadimanesh, F., & Salehi, B. (2021). Comparing solo versus ensemble convolutional neural networks for wetland classification using multi-spectral satellite imagery. *Remote Sensing*, *13*(11), 2046.

Jaskie, K., & Spanias, A. (2019). Positive and unlabeled learning algorithms and applications: A survey. In *2019 10th international conference on information, intelligence, systems and applications (iisa)* (pp. 1–8).

Ji, S., Wang, D., & Luo, M. (2020). Generative adversarial network-based full-space domain adaptation for land cover classification from multiple-source remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, *59*(5), 3816–3828.

Jian, P., Chen, K., & Cheng, W. (2021). Gan-based one-class classification for remote-sensing image change detection. *IEEE Geoscience and Remote Sensing Letters*, *19*, 1–5.

Jowkar, G.-H., & Mansoori, E. G. (2016). Perceptron ensemble of graph-based positive-unlabeled learning for disease gene identification. *Computational biology and chemistry*, *64*, 263–270.

Kalluri, T., Varma, G., Chandraker, M., & Jawahar, C. (2019). Universal semi-supervised semantic segmentation. In *Proceedings of the ieee/cvf international conference on computer vision* (pp. 5259–5270).

Kandaswamy, C., Silva, L. M., Alexandre, L. A., & Santos, J. M. (2015). Deep transfer learning ensemble for classification. In *International work-conference on artificial neural networks* (pp. 335–348).

Karbalayghareh, A., Qian, X., & Dougherty, E. R. (2018). Optimal bayesian transfer learning. *IEEE Transactions on Signal Processing*, *66*(14), 3724–3739.

Kawaguchi, K. (2016). Deep learning without poor local minima. *Advances in neural information processing systems*, *29*.

Ke, Z., Qiu, D., Li, K., Yan, Q., & Lau, R. W. (2020). Guided collaborative training for pixel-wise semi-supervised learning. In *European conference on computer vision* (pp. 429–445).

Kemker, R., Salvaggio, C., & Kanan, C. (2018). Algorithms for semantic segmentation of multispectral remote sensing imagery using deep learning. *ISPRS journal of photogrammetry and remote sensing*, *145*, 60–77.

Keskar, N. S., Mudigere, D., Nocedal, J., Smelyanskiy, M., & Tang, P. T. P. (2016). On large-batch training for deep learning: Generalization gap and sharp minima. *arXiv preprint arXiv:1609.04836*.

Kim, T., & Kim, C. (2020). Attract, perturb, and explore: Learning a feature alignment network for semi-supervised domain adaptation. In *European conference on computer vision* (pp. 591–607).

Kiryo, R., Niu, G., Du Plessis, M. C., & Sugiyama, M. (2017). Positive-unlabeled learning with non-negative risk estimator. *Advances in neural information processing systems*, *30*.

Korzh, O., Joaristi, M., & Serra, E. (2018). Convolutional neural network ensemble fine-tuning for extended transfer learning. In *International conference on big data* (pp. 110–123).

Krupiński, M., Lewiński, S., & Malinowski, R. (2019). One class svm for building detection on sentinel-2 images. In *Photonics applications in astronomy, communications, industry, and high-energy physics experiments 2019* (Vol. 11176, p. 1117635).

Kumar, A., Kim, J., Lyndon, D., Fulham, M., & Feng, D. (2016). An ensemble of fine-tuned convolutional neural networks for medical image classification. *IEEE journal of biomedical and health informatics*, *21*(1), 31–40.

Laine, S., & Aila, T. (2016). Temporal ensembling for semi-supervised learning. *arXiv preprint arXiv:1610.02242*.

Łazęcka, M., Mielniczuk, J., & Teisseyre, P. (2021). Estimating the class prior for positive and unlabelled data via logistic regression. *Advances in Data Analysis and Classification*, *15*(4), 1039–1068.

LeCun, Y., Bengio, Y., et al. (1995). Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, *3361*(10), 1995.

Lee, J., Kim, E., Lee, S., Lee, J., & Yoon, S. (2019). Ficklenet: Weakly and semi-supervised semantic image segmentation using stochastic inference. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 5267–5276).

Lee, S., Purushwalkam, S., Cogswell, M., Crandall, D., & Batra, D. (2015). Why m heads are better than one: Training a diverse ensemble of deep networks. *arXiv preprint arXiv:1511.06314*.

Lee, W. S., & Liu, B. (2003). Learning with positive and unlabeled examples using weighted logistic regression. In *Icml* (Vol. 3, pp. 448–455).

Lei, L., Wang, X., Zhong, Y., Zhao, H., Hu, X., & Luo, C. (2021). Docc: Deep one-class crop classification via positive and unlabeled learning for multi-modal satellite imagery. *International Journal of Applied Earth Observation and Geoinformation*, *105*, 102598.

Li, W. (2013). *Geographic modeling with one-class data.* University of California, Merced.

Li, W., & Guo, Q. (2010). A maximum entropy approach to one-class classification of remote sensing imagery. *International Journal of Remote Sensing*, *31*(8), 2227–2235.

Li, W., & Guo, Q. (2013). A new accuracy assessment method for one-class remote sensing classification. *IEEE transactions on geoscience and remote sensing*, *52*(8), 4621–4632.

Li, W., Guo, Q., & Elkan, C. (2010). A positive and unlabeled learning algorithm for one-class classification of remote-sensing data. *IEEE transactions on geoscience and remote sensing*, *49*(2), 717–725.

Li, W., Guo, Q., & Elkan, C. (2011). Can we model the probability of presence of species without absence data? *Ecography*, *34*(6), 1096–1105.

Li, Y., Yuan, L., & Vasconcelos, N. (2019). Bidirectional learning for domain adaptation of semantic segmentation. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 6936–6945).

Li, Y., Zhang, H., Xue, X., Jiang, Y., & Shen, Q. (2018). Deep learning for remote sensing image classification: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *8*(6), e1264.

Lin, D., Dai, J., Jia, J., He, K., & Sun, J. (2016). Scribblesup: Scribble-supervised convolutional networks for semantic segmentation. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 3159–3167).

Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., & Belongie, S. (2017). Feature pyramid networks for object detection. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 2117–2125).

Liu, B., Dai, Y., Li, X., Lee, W. S., & Yu, P. S. (2003). Building text classifiers using positive and unlabeled examples. In *Third ieee international conference on data mining* (pp. 179–186).

Liu, B., Lee, W. S., Yu, P. S., & Li, X. (2002). Partially supervised classification of text documents. In *Icml* (Vol. 2, pp. 387–394).

Liu, B., Liu, C., Xiao, Y., Liu, L., Li, W., & Chen, X. (2022). Adaboost-based transfer learning method for positive and unlabelled learning problem. *Knowledge-Based Systems*, 108162.

Liu, B., Zhu, C., Wang, K., Liu, X., & Yu, W. (2011). A one-class-extraction framework for high resolution sar image classification. In *Proceedings of 2011 ieee cie international conference on radar* (Vol. 1, pp. 732–735).

Liu, M.-Y., Breuel, T., & Kautz, J. (2017). Unsupervised image-to-image translation networks. *Advances in neural information processing systems*, *30*.

Liu, R., Li, W., Liu, X., Lu, X., Li, T., & Guo, Q. (2018). An ensemble of classifiers based on positive and unlabeled data in one-class remote sensing classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *11*(2), 572–584.

Liu, X., Liu, H., Datta, P., Frey, J., & Koch, B. (2020). Mapping an invasive plant spartina alterniflora by combining an ensemble one-class classification algorithm with a phenological ndvi time-series analysis approach in middle coast of jiangsu, china. *Remote Sensing*, *12*(24), 4010.

Liu, X., Liu, H., Gong, H., Lin, Z., & Lv, S. (2017). Appling the one-class classification method of maxent to detect an invasive plant spartina alterniflora with time-series analysis. *Remote Sensing*, *9*(11), 1120.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., … Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the ieee/cvf international conference on computer vision* (pp. 10012–10022).

Loghmani, M. R., Vincze, M., & Tommasi, T. (2020). Positive-unlabeled learning for open set domain adaptation. *Pattern Recognition Letters*, *136*, 198–204.

Long, J., Shelhamer, E., & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 3431–3440).

Lu, D., & Weng, Q. (2007). A survey of image classification methods and techniques for improving classification performance. *International journal of Remote sensing*, *28*(5), 823–870.

Lucas, B., Pelletier, C., Schmidt, D., Webb, G. I., & Petitjean, F. (2021). A bayesian-inspired, deep learning-based, semi-supervised domain adaptation technique for land cover mapping. *Machine Learning*, 1–33.

Luo, Y., Zheng, L., Guan, T., Yu, J., & Yang, Y. (2019). Taking a closer look at domain shift: Category-level adversaries for semantics consistent domain adaptation. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 2507–2516).

Ma, S., Li, X., Zhang, Z., Tang, J., & Guo, F. (2021). Geu-net: Rethinking the information transmission in the skip connection of u-net architecture. In *2021 ieee international conference on bioinformatics and biomedicine (bibm)* (pp. 1020–1025).

Mack, B., Roscher, R., Stenzel, S., Feilhauer, H., Schmidtlein, S., & Waske, B. (2016). Mapping raised bogs with an iterative one-class classification approach. *ISPRS Journal of Photogrammetry and Remote Sensing*, *120*, 53–64.

Mack, B., & Waske, B. (2017). In-depth comparisons of maxent, biased svm and one-class svm for one-class classification of remote sensing data. *Remote sensing letters*, *8*(3), 290–299.

Maggiori, E., Tarabalka, Y., Charpiat, G., & Alliez, P. (2017). Can semantic labeling methods generalize to any city? the inria aerial image labeling benchmark. In *2017 ieee international geoscience and remote sensing symposium (igarss)* (pp. 3226–3229).

Marmanis, D., Wegner, J. D., Galliani, S., Schindler, K., Datcu, M., & Stilla, U. (2016). Semantic segmentation of aerial images with an ensemble of cnss. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 2016*, *3*, 473–480.

Maulik, U., & Chakraborty, D. (2017). Remote sensing image classification: A survey of support-vector-machine-based advanced techniques. *IEEE Geoscience and Remote Sensing Magazine*, *5*(1), 33–52.

McDonald, G. G., Costello, C., Bone, J., Cabral, R. B., Farabee, V., Hochberg, T., … Zahn, O. (2021). Satellites can reveal global extent of forced labor in the world's fishing fleet. *Proceedings of the National Academy of Sciences*, *118*(3).

Melas-Kyriazi, L., & Manrai, A. K. (2021). Pixmatch: Unsupervised domain adaptation via pixelwise consistency training. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 12435–12445).

Meng, Z., Xie, Y., & Sun, J. (2021). Short-term load forecasting by transfer learning based on positive and unlabeled learning. In *The 2nd international conference on computing and data science* (pp. 1–5).

Mignone, P., & Pio, G. (2018). Positive unlabeled link prediction via transfer learning for gene network reconstruction. In *International symposium on methodologies for intelligent systems* (pp. 13–23).

Mignone, P., Pio, G., D'Elia, D., & Ceci, M. (2020). Exploiting transfer learning for the reconstruction of the human gene regulatory network. *Bioinformatics*, *36*(5), 1553–1561.

Miller, H. J., & Goodchild, M. F. (2015). Data-driven geography. *GeoJournal*, *80*(4), 449–461.

Mittal, S., Tatarchenko, M., & Brox, T. (2019). Semi-supervised semantic segmentation with high- and low-level consistency. *IEEE transactions on pattern analysis and machine intelligence*, *43*(4), 1369–1379.

Mnih, V. (2013). *Machine learning for aerial image labeling* (Unpublished doctoral dissertation). University of Toronto.

Mosca, A., & Magoulas, G. D. (2017). Deep incremental boosting. *arXiv preprint arXiv:1708.03704*.

Musto, L., & Zinelli, A. (2020). Semantically adaptive image-to-image translation for domain adaptation of semantic segmentation. *arXiv preprint arXiv:2009.01166*.

Na, B., Kim, H., Song, K., Joo, W., Kim, Y.-Y., & Moon, I.-C. (2020). Deep generative positive-unlabeled learning under selection bias. In *Proceedings of the 29th acm international conference on information & knowledge management* (pp. 1155–1164).

Najjar, A., Kaneko, S., & Miyanaga, Y. (2017). Combining satellite imagery and open data to map road safety. In *Thirty-first aaai conference on artificial intelligence.*

Nam, H., Lee, H., Park, J., Yoon, W., & Yoo, D. (2021). Reducing domain gap by reducing style bias. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 8690–8699).

Nguyen, M. N., Li, X.-L., & Ng, S.-K. (2012). Ensemble based positive unlabeled learning for time series classification. In *International conference on database systems for advanced applications* (pp. 243–257).

Nigam, I., Huang, C., & Ramanan, D. (2018). Ensemble knowledge transfer for semantic segmentation. In *2018 ieee winter conference on applications of computer vision (wacv)* (pp. 1499–1508).

Niu, G., du Plessis, M. C., Sakai, T., Ma, Y., & Sugiyama, M. (2016). Theoretical comparisons of positive-unlabeled learning against positive-negative learning. *Advances in neural information processing systems*, *29*.

Noh, H., Hong, S., & Han, B. (2015). Learning deconvolution network for semantic segmentation. In *Proceedings of the ieee international conference on computer vision* (pp. 1520–1528).

Nozza, D., Fersini, E., & Messina, E. (2016). Deep learning and ensemble methods for domain adaptation. In *2016 ieee 28th international conference on tools with artificial intelligence (ictai)* (pp. 184–189).

Olsson, V., Tranheden, W., Pinto, J., & Svensson, L. (2021). Classmix: Segmentation-based data augmentation for semi-supervised learning. In *Proceedings of the ieee/cvf winter conference on applications of computer vision* (pp. 1369–1378).

Opitz, D., & Maclin, R. (1999). Popular ensemble methods: An empirical study. *Journal of artificial intelligence research*, *11*, 169–198.

Oquab, M., Bottou, L., Laptev, I., & Sivic, J. (2014). Learning and transferring mid-level image representations using convolutional neural networks. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 1717–1724).

Ouali, Y., Hudelot, C., & Tami, M. (2020). Semi-supervised semantic segmentation with cross-consistency training. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 12674–12684).

Pan, F., Shin, I., Rameau, F., Lee, S., & Kweon, I. S. (2020). Unsupervised intra-domain adaptation for semantic segmentation through self-supervision. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 3764–3773).

Papandreou, G., Chen, L.-C., Murphy, K. P., & Yuille, A. L. (2015). Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation. In *Proceedings of the ieee international conference on computer vision* (pp. 1742–1750).

Perini, L., Vercruyssen, V., & Davis, J. (2020). Class prior estimation in active positive and unlabeled learning. In *Proceedings of the 29th international joint conference on artificial intelligence and the 17th pacific rim international conference on artificial intelligence (ijcai-pricai 2020)* (pp. 2915–2921).

Pettorelli, N., Schulte to Bühne, H., C. Shapiro, A., & Glover-Kapfer, P. (2018). *Satellite remote sensing for conservation.* WWF Conservation Technology Series 1(4). Retrieved 03-19-2022, from `https://www.researchgate.net/profile/Aurelie-Shapiro/publication/324537528_Conservation_Technology_Series_Issue_4_SATELLITE_REMOTE_SENSING_FOR_CONSERVATION/links/5ad45106a6fdcc2935800fac/Conservation-Technology-Series-Issue-4-SATELLITE-REMOTE-SENSING-FOR-CONSERVATION.pdf?origin=publication_detail`

Phillips, S. J., Anderson, R. P., & Schapire, R. E. (2006). Maximum entropy modeling of species geographic distributions. *Ecological modelling*, *190*(3-4), 231–259.

Plazas, M., Ramos-Pollán, R., & Martínez, F. (2021). Ensemble-based approach for semisupervised learning in remote sensing. *Journal of Applied Remote Sensing*, *15*(3), 034509.

Rahman, A., Smith, D. V., & Timms, G. (2013). Multiple classifier system for automated quality assessment of marine sensor data. In *2013 ieee eighth international conference on intelligent sensors, sensor networks and information processing* (pp. 362–367).

Ran, Q., Zhang, M., Li, W., & Du, Q. (2016). Change detection with one-class sparse representation classifier. *Journal of Applied Remote Sensing*, *10*(4), 042006.

Rapinel, S., & Hubert-Moy, L. (2021). One-class classification of natural vegetation using remote sensing: A review. *Remote Sensing*, *13*(10), 1892.

Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 779–788).

Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, *28*, 91–99.

Rocchio, L., & Barsi, J. (n.d.). *Different bands comparisons among multiple satellites.* Retrieved 03-19-2022, from `https://landsat.gsfc.nasa.gov/about/technical-details/`

Ronneberger, O., Fischer, P., & Brox, T. (2015a). U-net: Convolutional networks for biomedical image segmentation. In *International conference on medical image computing and computer-assisted intervention* (pp. 234–241).

Ronneberger, O., Fischer, P., & Brox, T. (2015b). U-net: Convolutional networks for biomedical image segmentation. In *International conference on medical image computing and computer-assisted intervention* (pp. 234–241).

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., … Fei-Fei, L. (2015). ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, *115*(3), 211-252. doi: 10.1007/s11263-015-0816-y

Sagi, O., & Rokach, L. (2018a). Ensemble learning: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *8*(4), e1249.

Sagi, O., & Rokach, L. (2018b). Ensemble learning: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *8*(4), e1249.

Saito, K., Watanabe, K., Ushiku, Y., & Harada, T. (2018). Maximum classifier discrepancy for unsupervised domain adaptation. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 3723–3732).

Santara, A., Datta, J., Sarkar, S., Garg, A., Padia, K., & Mitra, P. (2019). Punch: Positive unlabelled classification based information retrieval in hyperspectral images. *arXiv preprint arXiv:1904.04547*.

Schölkopf, B., Platt, J. C., Shawe-Taylor, J., Smola, A. J., & Williamson, R. C. (2001). Estimating the support of a high-dimensional distribution. *Neural computation*, *13*(7), 1443–1471.

Shi, H., Pan, S., Yang, J., & Gong, C. (2018). Positive and unlabeled learning via loss decomposition and centroid estimation. In *Ijcai* (pp. 2689–2695).

Shi, L., Wang, Z., Pan, B., & Shi, Z. (2020). An end-to-end network for remote sensing imagery semantic segmentation via joint pixel-and representation-level domain adaptation. *IEEE Geoscience and Remote Sensing Letters*, *18*(11), 1896–1900.

Shu, S., Lin, Z., Yan, Y., & Li, L. (2020). Learning from multi-class positive and unlabeled data. In *2020 ieee international conference on data mining (icdm)* (pp. 1256–1261).

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Sohn, K., Berthelot, D., Carlini, N., Zhang, Z., Zhang, H., Raffel, C. A., … Li, C.-L. (2020). Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in Neural Information Processing Systems*, *33*, 596–608.

Song, B., Li, P., & Jia, X. (2020). New feature selection methods using sparse representation for one-class classification of remote sensing images. *IEEE Geoscience and Remote Sensing Letters*, *18*(10), 1761–1765.

Song, B., Li, P., Li, J., & Plaza, A. (2016). One-class classification of remote sensing images using kernel sparse representation. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *9*(4), 1613–1623.

Sonntag, J., Behrens, G., & Schmidt-Thieme, L. (2022). Positive-unlabeled domain adaptation. *arXiv preprint arXiv:2202.05695*.

Souly, N., Spampinato, C., & Shah, M. (2017). Semi supervised semantic segmentation using generative adversarial network. In *Proceedings of the ieee international conference on computer vision* (pp. 5688–5696).

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, *15*(1), 1929–1958.

Tan, C., Sun, F., Kong, T., Zhang, W., Yang, C., & Liu, C. (2018). A survey on deep transfer learning. In *International conference on artificial neural networks* (pp. 270–279).

Tao, A., Sapra, K., & Catanzaro, B. (2020). Hierarchical multi-scale attention for semantic segmentation. *arXiv preprint arXiv:2005.10821*.

Tarvainen, A., & Valpola, H. (2017). Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, *30*.

Tasar, O., Giros, A., Tarabalka, Y., Alliez, P., & Clerc, S. (2020). Daugnet: Unsupervised, multisource, multitarget, and life-long domain adaptation for semantic segmentation of satellite images. *IEEE Transactions on Geoscience and Remote Sensing*, *59*(2), 1067–1081.

Tasar, O., Happy, S., Tarabalka, Y., & Alliez, P. (2020a). Colormapgan: Unsupervised domain adaptation for semantic segmentation using color mapping generative adversarial networks. *IEEE Transactions on Geoscience and Remote Sensing*, *58*(10), 7178–7193.

Tasar, O., Happy, S., Tarabalka, Y., & Alliez, P. (2020b). Semi2i: Semantically consistent image-to-image translation for domain adaptation of remote sensing data. In *Igarss 2020-2020 ieee international geoscience and remote sensing symposium* (pp. 1837–1840).

Tasar, O., Tarabalka, Y., Giros, A., Alliez, P., & Clerc, S. (2020). Standardgan: Multi-source domain adaptation for semantic segmentation of very high resolution satellite images by data standardization. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition workshops* (pp. 192–193).

Teisseyre, P. (2021). Classifier chains for positive unlabelled multi-label learning. *Knowledge-Based Systems*, *213*, 106709.

Torrey, L., & Shavlik, J. (2010). Transfer learning. In *Handbook of research on machine learning applications and trends: algorithms, methods, and techniques* (pp. 242–264). IGI global.

Truong, T.-D., Duong, C. N., Le, N., Phung, S. L., Rainwater, C., & Luu, K. (2021). Bimal: Bijective maximum likelihood approach to domain adaptation in semantic scene segmentation. In *Proceedings of the ieee/cvf international conference on computer vision* (pp. 8548–8557).

Tuia, D., Persello, C., & Bruzzone, L. (2016). Domain adaptation for the classification of remote sensing data: An overview of recent advances. *IEEE geoscience and remote sensing magazine*, *4*(2), 41–57.

Utgoff, P. E. (1986). Shift of bias for inductive concept learning. *Machine learning: An artificial intelligence approach*, *2*, 107–148.

Utgoff, P. E. (2012). *Machine learning of inductive bias* (Vol. 15). Springer Science & Business Media.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., … Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

Vu, T.-H., Jain, H., Bucher, M., Cord, M., & Pérez, P. (2019). Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 2517–2526).

Wang, A. X., Tran, C., Desai, N., Lobell, D., & Ermon, S. (2018). Deep transfer learning for crop yield prediction with remote sensing data. In *Proceedings of the 1st acm sigcas conference on computing and sustainable societies* (pp. 1–5).

Wang, S., Hu, H., Lin, T., Liu, Y., Padmanabhan, A., & Soltani, K. (2015). Cybergis for data-intensive knowledge discovery. *SIGSPATIAL Special*, *6*(2), 26–33.

Wang, Z., Wei, Y., Feris, R., Xiong, J., Hwu, W.-M., Huang, T. S., & Shi, H. (2020). Alleviating semantic-level shift: A semi-supervised domain adaptation method for semantic segmentation. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition workshops* (pp. 936–937).

Warde-Farley, D., Goodfellow, I. J., Courville, A., & Bengio, Y. (2013). An empirical analysis of dropout in piecewise linear networks. *arXiv preprint arXiv:1312.6197*.

Wasay, A., Hentschel, B., Liao, Y., Chen, S., & Idreos, S. (2020). Mothernets: Rapid deep ensemble learning. *Proceedings of Machine Learning and Systems*, *2*, 199–215.

Wasay, A., & Idreos, S. (2020). More or less: When and how to build convolutional neural network ensembles. In *International conference on learning representations.*

Weiss, K., Khoshgoftaar, T. M., & Wang, D. (2016). A survey of transfer learning. *Journal of Big data*, *3*(1), 1–40.

Wu, B., Qiu, W., Jia, J., & Liu, N. (2020). Landslide susceptibility modeling using bagging-based positive-unlabeled learning. *IEEE Geoscience and Remote Sensing Letters*, *18*(5), 766–770.

Wurm, M., Stark, T., Zhu, X. X., Weigand, M., & Taubenböck, H. (2019). Semantic segmentation of slums in satellite images using transfer learning on fully convolutional neural networks. *ISPRS journal of photogrammetry and remote sensing*, *150*, 59–69.

Xia, J., & Yokoya, N. (2021). Building damage mapping with self-positive unlabeled learning. *Artificial Intelligence for Humanitarian Assistance and Disaster Response Workshop, NeurIPS*.

Xia, J., Yokoya, N., & Adriano, B. (2021). Building damage mapping with self-positiveunlabeled learning. *arXiv preprint arXiv:2111.02586*.

Xie, M., Jean, N., Burke, M., Lobell, D., & Ermon, S. (2016). Transfer learning from deep features for remote sensing and poverty mapping. In *Thirtieth aaai conference on artificial intelligence.*

Xu, Y., Xu, C., Xu, C., & Tao, D. (2017). Multi-positive and unlabeled learning. In *Ijcai* (pp. 3182–3188).

Yan, Z., Yu, X., Qin, Y., Wu, Y., Han, X., & Cui, S. (2021). Pixel-level intra-domain adaptation for semantic segmentation. In *Proceedings of the 29th acm international conference on multimedia* (pp. 404–413).

Yang, P., Humphrey, S. J., James, D. E., Yang, Y. H., & Jothi, R. (2016). Positive-unlabeled ensemble learning for kinase substrate prediction from dynamic phosphoproteomics data. *Bioinformatics*, *32*(2), 252–259.

Yang, P., Li, X., Chua, H.-N., Kwoh, C.-K., & Ng, S.-K. (2014). Ensemble positive unlabeled learning for disease gene identification. *PloS one*, *9*(5), e97079.

Yang, P., Liu, W., & Yang, J. Y. H. (2017). Positive unlabeled learning via wrapper-based adaptive sampling. In *Ijcai* (pp. 3273–3279).

Yang, W., Yin, X., Song, H., Liu, Y., & Xu, X. (2013). Extraction of built-up areas from fully polarimetric sar imagery via pu learning. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *7*(4), 1207–1216.

Yang, Y., Liang, K. J., & Carin, L. (2020). Object detection as a positive-unlabeled problem. *British Machine Vision Conference (BMVC).*

Yang, Y., Lv, H., & Chen, N. (2021). A survey on ensemble learning under the era of deep learning. *arXiv preprint arXiv:2101.08387*.

Yang, Y., & Soatto, S. (2020). Fda: Fourier domain adaptation for semantic segmentation. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 4085–4095).

Yu, Y., Liu, H., Fu, M., Chen, J., Wang, X., & Wang, K. (2021). A two-branch neural network for non-homogeneous dehazing via ensemble learning. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 193–202).

Yun, S., Han, D., Oh, S. J., Chun, S., Choe, J., & Yoo, Y. (2019). Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the ieee/cvf international conference on computer vision* (pp. 6023–6032).

Zhang, C., Hou, Y., & Zhang, Y. (2020). Learning from positive and unlabeled data without explicit estimation of class prior. In *Proceedings of the aaai conference on artificial intelligence* (Vol. 34, pp. 6762–6769).

Zhang, Y., Zong, R., Han, J., Zheng, H., Lou, Q., Zhang, D., & Wang, D. (2019). Transland: An adversarial transfer learning approach for migratable urban land usage classification using remote sensing. In *2019 ieee international conference on big data (big data)* (pp. 1567–1576).

Zhang, Z., Dalca, A. V., & Sabuncu, M. R. (2019). Confidence calibration for convolutional neural networks using structured dropout. *arXiv preprint arXiv:1906.09551*.

Zhao, D., Li, J., Yuan, B., & Shi, Z. (2021). V2rnet: An unsupervised semantic segmentation algorithm for remote sensing images via cross-domain transfer learning. In *2021 ieee international geoscience and remote sensing symposium igarss* (pp. 4676–4679).

Zhao, L., Li, Q., Zhang, Y., Du, X., Wang, H., & Shen, Y. (2019). Study on the potential of whitening transformation in improving single crop mapping accuracy. *Journal of Applied Remote Sensing*, *13*(3), 034512.

Zhao, S., Li, B., Xu, P., & Keutzer, K. (2020). Multi-source domain adaptation in the deep learning era: A systematic survey. *arXiv preprint arXiv:2002.12169*.

Zheng, P., Yuan, S., Wu, X., Li, J., & Lu, A. (2019). One-class adversarial nets for fraud detection. In *Proceedings of the aaai conference on artificial intelligence* (Vol. 33, pp. 1286–1293).

Zhou, Q., Wu, X., Zhang, S., Kang, B., Ge, Z., & Latecki, L. J. (2022). Contextual ensemble network for semantic segmentation. *Pattern Recognition*, *122*, 108290.

Zhou, Z., Siddiquee, M. M. R., Tajbakhsh, N., & Liang, J. (2019). Unet++: Redesigning skip connections to exploit multiscale features in image segmentation. *IEEE transactions on medical imaging*, *39*(6), 1856–1867.

Zhou, Z., Zhang, F., Xiao, H., Wang, F., Hong, X., Wu, K., & Zhang, J. (2021). A novel ground-based cloud image segmentation method by using deep transfer learning. *IEEE Geoscience and Remote Sensing Letters*, *19*, 1–5.

Zhou, Z.-H. (2021). Ensemble learning. In *Machine learning* (pp. 181–210). Springer.

Zhu, C., Liu, B., Yu, Q., Liu, X., & Yu, W. (2012). A spy positive and unlabeled learning classifier and its application in hr sar image scene interpretation. In *2012 ieee radar conference* (pp. 0516–0521).

Zhu, X. X., Tuia, D., Mou, L., Xia, G.-S., Zhang, L., Xu, F., & Fraundorfer, F. (2017). Deep learning in remote sensing: A comprehensive review and list of resources. *IEEE Geoscience and Remote Sensing Magazine*, *5*(4), 8–36.

Zou, M., & Zhong, Y. (2018). Transfer learning for classification of optical satellite image. *Sensing and Imaging*, *19*(1), 1–13.

Zou, Y., Yu, Z., Kumar, B. V., & Wang, J. (2018, September). Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In *Proceedings of the european conference on computer vision (eccv).*