INRS
Institut national
de la recherche
scientifique

FINAL REPORT

High-Resolution Urban Land Cover Mapping from Satellite Imagery Using Deep Learning

March 2025

Authors:

Kaushik Roy Saeid Homayouni

INRS Reference Number: R2277

ISBN : 978-2-925559-10-8 (version numérique)

Dépôt légal - Bibliothèque et Archives nationales du Québec, 2025
Dépôt légal - Bibliothèque et Archives Canada, 2025

Contents

1 Research Objectives

Under rapid urbanization and climate changes have increased global urban flood risks. The land cover is converted from natural to man-made with high percentages of impervious surfaces, which increases the surface runoff and induces flash floods after a major precipitation in urban areas, especially in a metropolitan region such as the Greater Toronto Area (GTA). To understand and analyze the intensity and dynamics of the urban flash flood risks, it needs to quantify the special distribution of the land surface imperviousness and vegetation land cover in urban watersheds.

In the previous CCMEO program related to FHIMP (Flood Hazards Identification and Mapping Program), a small-scale test (the sub-watershed Don Valley of GTA) was undertaken about the impacts of urban land cover on flash flood risk distribution and changes using the land cover information derived from low-resolution Landsat data. From the test, it was found that high-resolution urban land cover information with more details on the imperviousness/vegetation distribution is needed for more accurate urban flood risk assessments. However, due to a lack of in-house personal resources, information gaps exist in the high-resolution (HR) urban land cover map information for further flood risk analyses for major Canadian cities. This contract work will fill the information gap in the HR urban land cover map of GTA, which is planned to be generated from HR satellite imagery. Deep learning technology is required since the map's coverage is large, and a huge volume of imagery data processing is needed.

This work has the following objectives:

1. To preprocess the HR multi-spectral image data, geo-register the HR image data with Landsat image data and collect related existing HR geospatial data useful for the mapping.

2. To determine which classification methods are useful and efficient for this map generation.

3. To create the HR land cover map of GTA using the HR satellite imagery.

4. To validate the accuracies of the land cover classes in the map product.

2 Summary

Accurate mapping of land cover and land use at very high resolution (VHR) is crucial for studying urban development and human-environment interactions. Deep learning techniques, particularly semantic segmentation models, have emerged as powerful tools for this task. However, their widespread application is hindered by the substantial demand for annotated VHR datasets. Nonetheless, their effectiveness is often constrained by the extensive volume of labeled VHR imagery required for training.. Existing studies have mostly used low to medium-resolution imagery and fewer bands, resulting in limited downstream applicability. To our knowledge, this is the first attempt at studying urban areas in Canada at such spatial resolution using self-supervised deep learning techniques. The objective is to classify VHR multispectral imagery into eight urban land cover categories. The main challenges are preparing analysis-ready data, class imbalance, and a limited amount of labeled data. To address these challenges, we introduce an innovative deep learning framework designed to improve spectral-spatial consistency while leveraging the wealth of available unlabeled data for more effective learning and easily apply pre-trained representations to downstream tasks. We perform super-resolution using deep learning pansharpening, then latent feature extraction without labels and knowledge distillation using a small amount of labeled data. The proposed workflow is applied to Worldview 3 imagery over 80,000 patches of size 256x256 at 1m spatial resolution. The methodology was applied to two unet variants, a simple Unet and an attention-gated Unet with a Resnet50 encoder. The results show that while the simple Unet could not adequately capture the complexity of the data, unlike the complex model, self-supervised pre-training improves the overall accuracy(OA) of the prediction in both cases. For simple Unet, the accuracy was improved from 69% to 74%, and for complex unet, the OA improved from 80% to 88%. In conclusion, we display the effectiveness of multi-view self-supervised semantic segmentation on multispectral VHR images and create a land cover product for future research.

3 Introduction

Urban land cover classification (ULC) is crucial for policymakers and planners of all countries. In Canada, where diverse climates and terrains pose unique challenges, accurate land cover classification aids in understanding urban sprawl, assessing environmental impacts, planning for resilient infrastructure, and and reduction of urban natural disasters. The rapid urbanization observed in many Canadian cities increases risks of natural hazards such as flooding, landslides, heat island effects, and wildfires. Creating high resolution ULC is labor-intensive, time-consuming, and costly, making it challenging and impractical for large areas, especially Canadian urban settings. In Canadian cities like Toronto, Vancouver, and Montreal, collecting ground truth data requires extensive fieldwork, often involving teams of researchers and technicians with high labour costs. Remote sensing is a viable alternative widely used to create such products. The rapid development of sensor technology has significantly enhanced the availability and quality of very high-resolution (VHR) satellite imagery, which has proven to be an essential tool for monitoring urban environments. Satellites such as WorldView-2/3/4, and GaoFen-1/2 provide imagery with a ground sampling distance (GSD) of less than 5 meters, capturing intricate spatial details that are valuable for urban land use and land cover classification. The ability to distinguish fine-scale features in urban landscapes has led to increased research interest in leveraging VHR imagery for detailed classification tasks. Effectively distinguishing urban land cover in VHR imagery is challenging due to the heterogenous nature of urban landscapes. Approaches to classification generally fall into pixels or object based analysis categories. In pixel-based classification, each pixel is analyzed individually, with labels assigned based only on its spectral characteristics, without considering spatial relationships with neighboring pixels. While this approach has been widely used, its effectiveness is often compromised in high-resolution imagery due to the "salt-and-pepper" effect, where variations in spectral responses at the pixel level lead to fragmented and noisy classification results. This issue arises because individual pixels do not always correspond to distinct real-world objects, especially in urban settings where materials, shadows, and mixed pixels introduce spectral inconsistencies.

To address some these limitations, object-based image analysis (OBIA) could be a more effective alternative for classifying VHR imagery. Rather than assigning labels to individual pixels, object-based methods first segment the image into meaningful groups of pixels, known as objects, using methods such as Multi-Resolution Segmentation, Watershed, or Quadtree-Segmentation. These objects, which incorporate both spectral and spatial properties, are then classified as distinct land cover types. By considering the relationships between neighboring pixels, object-based methods help mitigate the salt-and-pepper effect and produce more coherent classification results. However, the accuracy of object-based analysis classification depends on the quality of the segmentation output, as poorly segmented objects can introduce classification errors. In addition, object-based approaches require the careful selection of relevant features, such as segment shape, texture, and contextual attributes, which guide the classification. In complex urban environments, manually selecting features that effectively distinguish all land cover types can be challenging. As a result, there is an increasing need for automated feature extraction techniques that can learn and adapt to different land cover classes directly from the data, reducing the reliance on manual feature engineering. The integration of deep learning methods with object-based classification holds promise for improving the accuracy and scalability of urban land cover mapping, offering a more data-driven approach to feature representation and classification.

Various artificial intelligence methods have been widely applied to pattern recognition and computer vision tasks, demonstrating their effectiveness in classifying remote sensing imagery. Techniques such as boosted trees or K-means have been successfully utilized to analyze and categorize land cover features in satellite images, offering robust solutions for automated classification . Hinton et al. introduced deep learning theory in 2015, marking a significant shift from traditional machine learning approaches like support vector classifiers, and Random Forest (RF). Unlike these conventional methods, which often rely on manually crafted features, deep learning models excel at automatically extracting meaningful features directly from large datasets. This capability allows them to identify discriminative patterns as part of the training process, eliminating the need for predefined feature engineering. Recent advancements in deep network architectures have produced state-of-the-art results across many computer vision applications by capturing complex spatial patterns in images. CNNs, in particular, have revolutionized image classification by using deep convolutional layers with sparse connections to learn high-level feature representations, significantly improving performance compared to conventional handcrafted feature-based methods. This shift toward automated feature learning has shown widespread success of deep learning in visual recognition tasks. Early convolutional neural network (CNN) models only helped to assign a single class label to an entire image, requiring large-scale datasets for effective training. To improve performance, researchers introduced increasingly complex and deeper architectures, such as VGG, GoogLeNet, ResNet, and EfficientNet. While these models excelled at image-level classification, they were not designed for pixel-wise predictions. To address this, various methods were later introduced to generate classification outputs at the pixel level, producing dense prediction maps that match the dimensions of the input image. One of the most influential developments in this area was the Fully Convolutional Network (FCN), introduced by Long et al.. This approach introduced deconvolution layers in place of the fully connected layers typically found in conventional CNN architectures, allowing for classification at the pixel level. Despite its effectiveness, FCN has notable limitations, including difficulties in capturing fine details and generating smooth object boundaries. These drawbacks arise from the downsampling process, which leads to the loss of spatial information. Nevertheless, the core structure of FCN, which follows an encoder-decoder paradigm with downsampling and upsampling operations, has provided the foundation for numerous modern semantic segmentation architectures. Models such as U-Net, FPN have built upon this framework, incorporating additional mechanisms to enhance spatial accuracy and boundary delineation. Fully convolutional models have been extensively applied to semantic segmentation tasks for classifying very high-resolution (VHR) imagery. Numerous applications in remote sensing have benefited from semantic segmentation, including road and building footprint extraction, land cover classification, and change detection. Despite its effectiveness, deep learning-based segmentation models often require substantial amounts of annotated data to produce reliable results. However, generating pixel-wise labeled datasets for multiple classes in remotely sensed imagery is a labor-intensive and costly process. Given the challenge of acquiring sufficient labeled data, researchers have explored various strategies to reduce the reliance on extensive training datasets, as limited training samples can significantly impact the accuracy and generalization of deep learning models.

To address these challenges, self-supervised learning (SSL) has become a promising technique for training deep learning models without relying on large labeled datasets. SSL enables a model to extract meaningful features from vast amounts of unlabeled data by solving a predefined pretext task—an auxiliary learning objective designed to guide the initial training process. The learned representations serve as a strong foundation for downstream supervised tasks, reducing the dependence on extensive manual annotations. This approach has gained significant attention in computer vision, often achieving performance levels comparable to or even exceeding those of fully supervised methods. Currently, different kind of self-supervised learning methods are Momentum Contrast, SimCLR, SwAV, BYOL, Siamese, and Barlow Twins. Methods based on contrastive learning and distillation function by enhancing similarity, which depends on efficiently generating positive samples (corresponding images) and negative samples (non-corresponding images) during the pre-training phase. Distinguishing negative samples in remote sensing imagery presents a significant challenge due to the complexity of feature representations at both low and high levels. Variations in spectral and spatial characteristics make it difficult to define clear negative pairs, adding to the complexity of contrastive learning approaches. In contrast, Barlow Twins eliminates the need for explicit negative sample selection, making it a more suitable choice for remote sensing segmentation. Despite the promise of self-supervised learning (SSL) in this domain, its application remains relatively unexplored, with most existing studies focusing on low-resolution RGB datasets and tasks such as scene classification . Some datasets are available for pixel-level classification and have been used in the literature to create Urban LULC products . However, these datasets are not based in Canada, and any model trained on these datasets may not produce accurate or usable land cover maps.

To fill the research gap mentioned above, this study applies a novel deep learning-based workflow that creates accurate land cover maps for urban areas in Canada using VHR data. The study exploits the availability of panchromatic and multispectral bands and fuses them using deep learning pansharpening. Following this, a novel analysis-ready dataset is created with only 10% pixels labeled with one of eight urban land cover classes. Four models are compared – a) Unet – a simple UNet model, b) Unet+SSL - an Unet with a pretrained SSL backbone, and c) Resnet50_AttUnet – an attention-gated Unet with not pretrained Resnet50 backbone, and d) Resnet50_AttUnet+SSL - model c but with SSL pre trained Resnet50 backbone. The research demonstrates that Resnet50_AttUnet+SSL outperforms models a,b, and c for urban land cover classification. This is one of the first studies to apply modern deep learning approches such as super-resolution and self-supervised learning (SSL) to VHR imagery of Canadian urban areas. The key contributions of this paper are as follows:

- A novel satellite imagery dataset unique to Canadian urban database is used for generation of high resolution land cover information for two majoy urban areas.
- The issue of scarce labeled data in remote sensing is tackled by generating multiple invariant views using data augmentations and pre-training a normal Unet encoder and a Resnet50 encoder in a customized Barlow Twins strategy.
- The pretrained SSL encoders are frozen, and two U-Net variants, a basic U-Net, and an Attention Gated Unet, are trained as decoders. The study shows SSL pre-training improves baseline accuracy in both variants of U-Net.

The remainder of the paper is divided into several sections, where Sect. 2 presents the literature review of the recent developments in the self-supervised segmentation approaches. Sections 3 and 4 provide an overview, while Sections 5 and 6 detail the proposed methodology and experimental results, respectively. Lastly, Section 7 offers the concluding remarks.

4 Literature Review

Recent advancements in remote sensing have seen growing interest in the application of self-supervised learning techniques to extract meaningful representations from satellite imagery. For instance, generative adversarial networks in leveraged multiple-layer feature-matching to derive scale-specific spatial characteristics from VHR data, which were then employed for land cover classification tasks. Similarly, Walter introduced a novel pretext task in which high-frequency channel information was used to predict RGB values, facilitating feature learning in a self-supervised manner. Further extending this paradigm, Moore incorporated nominal pretext tasks such as image colorization, coordinates prediction, and sample discrimination—to extract domain-specific representations from Pleiades Neo imagery. The effectiveness of the pretraining was evaluated through transfer learning, demonstrating improved performance on LULC with very few labelled samples. In another study, Wang modified the MoCov2 framework by integrating a location-aware cost function to enhance spectral feature extraction from satellite images. Unlike conventional contrastive learning approaches that rely on standard data augmentation techniques to generate positive pairs, their method utilized geographic location information from frequently observed satellite paths to establish meaningful positive samples. Additionally, research has shown that successive pre-training on natural image datasets and on remote sensing imagery can enhance model accuracy on downstream tasks. For example, Jung systematically analyzed the impact of various data augmentation techniques in contrastive self-supervised learning applied to remote sensing datasets. Meanwhile, Ayush proposed a self-supervised pre-training strategy that simultaneously leveraged the relationship between satellite imagery and geo-tagged sound recordings. By incorporating both visual and auditory information, this approach facilitated a more robust pre-training process, enabling improved feature learning from multimodal remote sensing data.

Of all the available techniques, contrastive learning is the most popular approach. Contrastive learning methods are designed to train models by distinguishing between similar and dissimilar samples in the representation space. Specifically, these approaches encourage representations of semantically related inputs, such as different

Of all the available techniques, contrastive learning is the most popular approach. Contrastive learning methods are designed to train models by distinguishing between similar and dissimilar samples in the representation space. Specifically, these approaches encourage representations of semantically related inputs, such as different augmented versions of the same image, to be mapped closely together. Due to this fundamental principle, contrastive learning typically employs a Siamese-style architecture, where paired inputs undergo simultaneous processing to learn meaningful feature representations. Although self-supervised contrastive learning has gained prominence in recent years, the application of contrastive loss in remote sensing has a longer history. A notable early example is the work of (Cheng, Yang, Yao, Guo, and Han, 2018), which incorporated a supervised contrastive regularization term into convolutional neural network (CNN) features to enhance remote sensing scene classification. The first application of contrastive learning in a self-supervised remote sensing context was introduced by Jean et al. in their Tile2Vec framework (Jean, Wang, Samar, Azzari, Lobell, and Ermon, 2018). This method was conceptually influenced by word2vec (Mikolov, Chen, Corrado, and Dean, 2013) and bore similarities to Contrastive Predictive Coding (CPC) (Oord, Li, and Vinyals, 2018). However, a significant challenge in contrastive learning arises from the tendency of models to collapse into trivial solutions when solely optimizing for similarity between paired inputs. To mitigate this issue, researchers have explored various strategies, including alternative contrastive loss formulations, negative sampling techniques, and architectural modifications that promote more diverse and informative feature representations. SimCLR and MoCo use many negative samples in a single batch to learn proper representations. Although these models have been applied to remote sensing, they require high-performance computing clusters to train in practical time. BYOL and SimSiam are another class of contrastive SSL techniques that learn latent representations from knowledge distillation and have been successfully applied in remote sensing. Barlow Twins model is a recent proposal in computer vision SSL and uses a novel loss function to reduce redundancy between learned representations. Although this model has been applied in the medical field, it is well suited for remote sensing applications and does not require large batch sizes or high-performance clusters to train. Hence, exploring the effect of Barlow Twins training in remote sensing is interesting.

5 Study area

The study focuses on the Greater Montreal Area (GMA) and the Greater Toronto Area (GTA) (Figure 1), as they encompass the largest urban populations in Canada. These metropolitan regions contain extensive impervious surfaces and exhibit diverse land cover types with varying building densities. Over the past five decades, both areas have experienced substantial growth due to international and domestic migration, leading to their development as major multicultural urban centers characterized by significant social, cultural, and economic diversity. Furthermore, both Montreal and Toronto have been recognized as influential global cities according to Globalization and World Cities Research Network (GaWC), highlighting their role in shaping international cultural, political, and economic dynamics.

The urban built-up landscapes of the GTA and the GMA are characterized by their dynamic and diverse urban forms, shaped by distinct histories, population densities, and developmental patterns. In the GTA, the built-up environment features a mix of high-density urban cores, such as Toronto's downtown with its iconic skyline dominated by skyscrapers and mixed-use developments, and expansive low-density suburban neighborhoods with detached homes and cul-de-sacs. The transportation infrastructure is a defining feature, with major highways like the 401 and a network of commuter rail systems supporting urban sprawl. Industrial zones and business parks are interspersed throughout the suburban fringes, reflecting the area's role as a hub for finance, technology, and manufacturing. Meanwhile, green spaces, such as the Don Valley and Rouge National Urban Park, weave through the urban fabric, offering a counterbalance to the dense built-up areas. In contrast, the GMA displays a unique blend of European-inspired urbanism and modern development. Montreal's core is known for its historic architecture, including narrow streets, stone buildings, and iconic triplexes with external staircases, particularly in neighborhoods like Plateau Mont-Royal. The central business district features high-rise office towers, yet urban planning has retained an emphasis on maintaining sightlines to landmarks like Mount Royal. Beyond the island, suburban development is characterized by medium-density housing and industrial parks connected by an extensive highway network and commuter rail services. Both regions have seen significant growth in mixed-use developments and high-rise condominiums in response to population pressures, signaling a shift toward vertical urbanization. Despite differences, both urban areas are marked by the challenges of balancing growth with sustainable development, preserving green spaces, and managing infrastructure demands in highly urbanized environments.

5.1 GTA

The Greater Toronto Area (GTA), which includes the city of Toronto, is home to near about 6.5 million people, making Canada's largest metropolitan region and a crucial economic hub. Located in Southern Ontario along the northern shore of Lake Ontario, Toronto is a highly urbanized and diverse city, serving as the provincial capital of Ontario. According to the most recent census data, Toronto itself has a population of 2,794,356, ranking as the highest populated city in Canada and the fourth biggest in North America (Statistics Canada 2021). The region experiences a humid continental climate with noticeable seasonal variations. Historical climate records indicate that between 1991 and 2020, the average daily temperature in January was 3.5°C, while July had an average daily temperature of 22.5°C. Annual precipitation levels in the area were recorded at

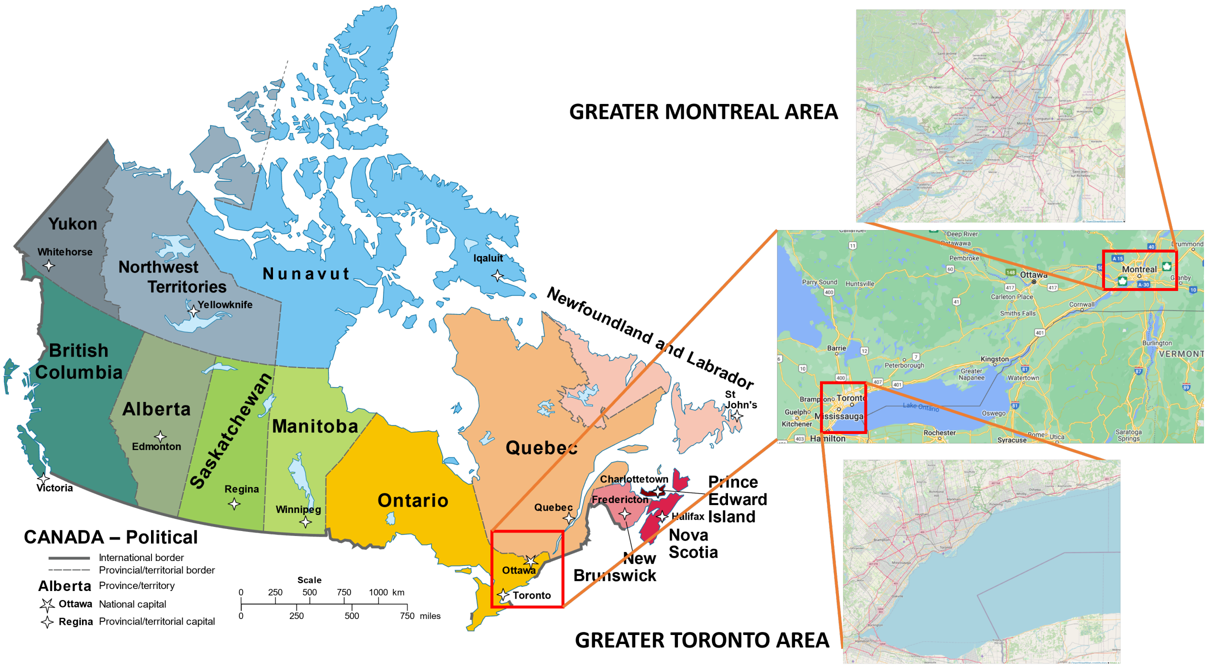


Figure 1: Map of the Study Area

approximately 822.7 mm (Environment and Climate Change Canada). Geographically, Toronto is positioned on the northwestern edge of Lake Ontario, where its proximity to the lake influences its local climate. The GTA features a varied topography, with the highest point reaching approximately 561 meters above sea level in the Oak Ridges Moraine near King City, while the lowest elevation, around 74 meters, is found along Toronto's waterfront. The land cover in the region is predominantly urban, characterized by a dense concentration of roads, residential neighborhoods, commercial districts, and industrial areas. These impervious surfaces define the city's landscape, reflecting the extensive development and infrastructure that support its growing population.

5.2 GMA

The Greater Montreal Area (GMA), home to approximately 4.5 million residents, is the second most populous city in Canada and the easternmost urban center considered in this study. Situated in the province of Quebec, Montreal is located on the Island of Montreal at the meeting point of the Saint Lawrence and Ottawa Rivers. The region experiences a diverse climate characterized by significant seasonal variations, with temperatures ranging from an average low of -15.8°C during colder seasons to a high of 27.8°C in warmer ones, accompanied by substantial precipitation throughout the year. The study area, positioned in southwestern Quebec, encompasses the Greater Montreal Area (GMA) which spans 4,300 km^2 and had a recorded population of 4,100,100, according to Statistics Canada (2011). The GMA's maximum elevation is approximately 233 meters, found at the summit of Mount Royal, which serves as a focal point for the city's design and cultural identity. The minimum elevation is about 6 meters along the popular St. Lawrence River, which defines much of the region's geography and urban infrastructure. Major land cover types in the GMA include urban impervious surfaces, interspersed with significant green spaces such as Mount Royal Park, agricultural lands on the periphery, and wetlands along the St. Lawrence River.

6 Data

This study employs very high-resolution (VHR) remote sensing imagery to classify urban land cover with a focus on high-resolution satellite data. The imagery utilized in this research is obtained from WorldView-3, an advanced satellite originally launched by DigitalGlobe, which is now a part of Maxar Technologies. WorldView-3, which became operational following its launch on August 13, 2014, represents a significant advancement in commercial satellite imagery technology. This cutting-edge Earth observation platform provides VHR imagery that has revolutionized the classification and analysis of urban land cover. The exceptional spatial and spectral properties of WorldView-3 make it an invaluable data source for researchers and urban planners alike, offering unprecedented capabilities for detailed urban mapping and monitoring.

WorldView-3 sets a new standard for commercial satellites in terms of spatial resolution. The panchromatic band offers an impressive 0.31 m resolution at nadir, increasing slightly to 0.34 m at 20° off-nadir. The multispectral bands provide 1.24 m resolution at nadir and 1.38 m at 20° off-nadir. The spectral capabilities of WorldView-3 are equally impressive and play a crucial role in urban land mapping. The satellite features a panchromatic band covering the 450-800 nm range, providing high-resolution grayscale imagery. The multispectral sensor includes eight bands: Coastal (400-450 nm), Blue (450-510 nm), Green (510-580 nm), Yellow (585-625 nm), Red (630-690 nm), Red Edge (705-745 nm), Near-IR1 (770-895 nm), and Near-IR2 (860-1040 nm). This comprehensive spectral coverage allows for sophisticated differentiation between urban materials and land cover types.

7 Results and discussion

7.1 Experimental setup

In our experimental setup, we leveraged a high-performance computing environment to facilitate the training and evaluation of our deep learning models. The hardware configuration consisted of an Intel Xeon E5 10 cores CPU, 128 GB RAM, and two **NVIDIA RTX 6000** graphics processing units (GPUs), each equipped with 24 GB of VRAM, providing ample computational power for parallel processing of large-scale geospatial datasets.

Our software stack was built upon the PyTorch deep learning framework, specifically utilizing **PyTorch Lightning** for streamlined model training and experiment management. To address geospatial data's unique challenges, we incorporated **TorchGeo** (Stewart, Robinson, Corley, Ortiz, Ferres, and Banerjee, 2022), a specialized library for processing and analyzing Earth observation datasets within the PyTorch ecosystem. Additionally, we employed **Kornia**, a computer vision library, to implement advanced image augmentation techniques and geometric transformations on GPUs instead of CPUs.

7.2 Land cover classification

The proposed framework generates a segmentation mask for a given multispectral image. Moreover, the quantitative results of both U-Net models with and without the Barlow Twins based pre-training are presented in Tables **??** and **??**.

It is observed that initially, the simple Unet models could not learn the complexity of the data. Unet and Unet+SSL models predict impervious classes as either road, water, or building. A more complex model such as Resnet50+AttUnet can better classify the impervious area mostly due to the Resnet50 encoder and the attention-gating mechanism. However, the output still has some noise, but after doing SSL pre-training, Resnet50AttUnet+SSL can properly delineate the impervious area compared to the true area. For road classification, the Unet model completely misses the intersection and the side roads, as shown below. Unet+SSL can improve the result by correctly classifying the side roads and a part of the road further away from the intersection but not the actual intersection itself. Resnet50AttUnet can predict the intersection better, but it cannot properly delineate the road. This is accomplished with SSL training as Resnet50AttUnet+SSL can catch the finer details and extract precise road networks, along with more prominent buildings in the background. For the bare class, the Unet model cannot differentiate between grass and bare soil, as shown below. Hence, it overclassifies grass and shrubs while incorrectly predicting that the impervious running track around the bare field is the road. Here, SSL training helps the model Unet+SSL differentiate between grass and bare soil while identifying the running track as impervious instead of road. Resnet50AttUnet is shown to remove the misclassified grass pixels completely, but it overestimates the amount of bare soil while completely missing the thin running track. SSL training helps here, and Resnet50AttUnet+SSL can correctly estimate the amount of bare soil and the prominent running track. The example for tree classification shown below is a classic example of the benefits of SSL pre-training. This is because of the cloud cover present in the Worldview 3 imagery. Although providers try to remove as many clouds as possible before dissemination, clouds frequently occur in optical remote sensing data, and Worldview 3 is no exception. As the study shows, the vanilla Unet model fails to recognize the tree canopy beneath the cloud and considers it bare soil. Unet+SSL can understand that it is probably vegetation but cannot distinguish between trees and grass. Renet50AttUnet, on the other hand, can understand the tree canopy correctly but fails in places where the cloud casts a shadow. Finally, the Resnet50AttUnet+SSL model can delineate the trees and the fine roads going through them. While the initial Unets struggle with grasses and shrubs, our methodology eventually enables the model to learn to differentiate between the two. For water classification, the model correctly identifies water pixels for the most part, but later models are also able to get finer details, such as impervious embankments and boat wharves. Classifying buildings in densely populated or commercial areas has been challenging, even with high-resolution imagery. The spectral signature of buildings varies a lot in commercial zones like downtown Toronto. The example below demonstrates the proposed methodology's usefulness in extracting building footprints. Unet model considers buildings to be impervious surfaces and the shadows to be water pixels. Tall buildings cast many shadows, and a simple encoder is not enough to represent this. Unet+SSL improves by detecting more buildings, but it still overestimates the footprint, and there are still some spurious water pixels instead of shadows. Resnet50AttUnet can delineate the buildings much better, but some shadow pixels are still considered to be water. However, the final model, Resnet50AttUnet+SSL, can finally eliminate the shadow pixels and properly classify them as either buildings or roads.

8 Conclusion

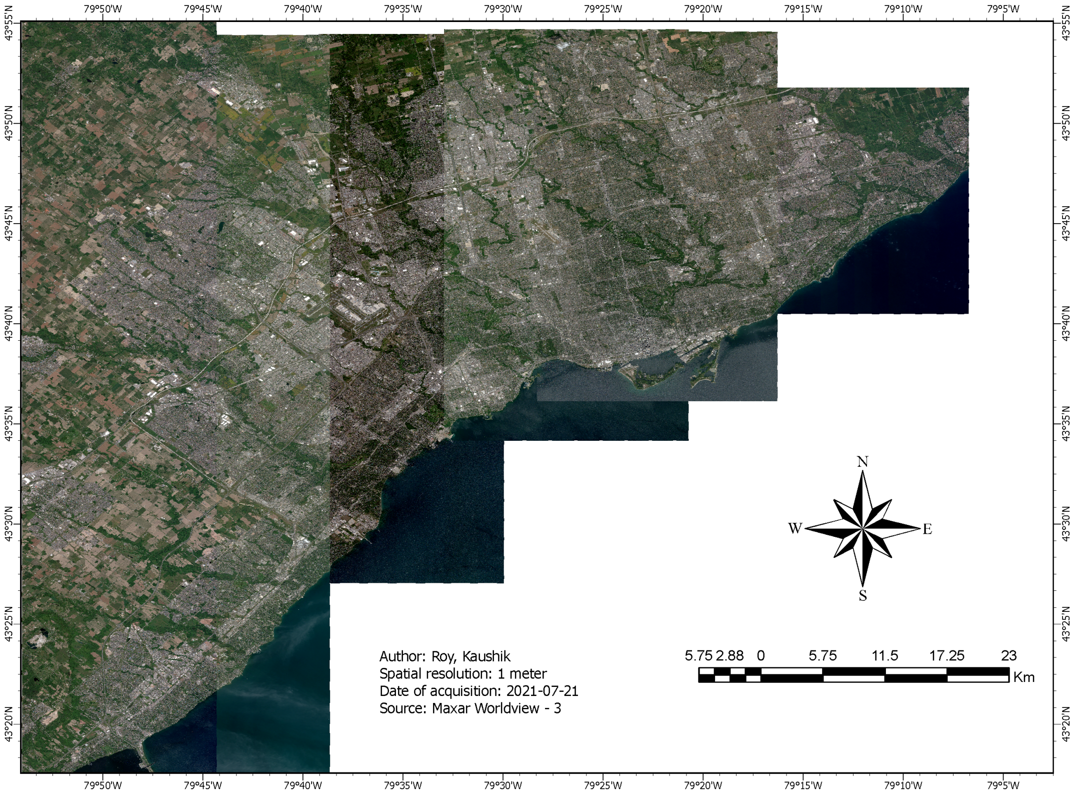
This study investigated the application of Barlow Twins, a self-supervised learning (SSL) framework, to enhance high-resolution land cover classification accuracies when labeled data is limited. By leveraging Barlow Twins' capacity for learning meaningful feature representations from vast amounts of unlabeled remote sensing data, we have demonstrated significant improvements in classification performance, specifically in scenarios where labeled samples are scarce. Adopting this method addresses a critical challenge in land cover classification: the costly and labor-intensive nature of acquiring labeled datasets for remote sensing tasks. Our approach employed a three-stage training pipeline, with deep learning pansharpening Barlow Twins pre-training on a large unlabeled dataset, followed by fine-tuning a small subset of labeled data for supervised classification. This pipeline allowed the model to learn useful representations of diverse land cover types during the pre-training stage, thereby improving the effectiveness of the supervised fine-tuning phase. Our experiments revealed that the model trained with SSL using Barlow Twins outperformed conventional supervised models' overall accuracy and F1 score. This is likely due to the design of Barlow Twins, which enforces redundancy reduction, encouraging the model to capture complementary features without relying on specific supervised signals. This characteristic is especially beneficial for remote sensing imagery, where variations in seasonal, atmospheric, and lighting conditions demand robust feature learning. In addition to improving classification performance, the Barlow Twins framework is highly adaptable and computationally efficient, making it a viable approach for land cover mapping projects where resources are constrained. Efficient usage of unlabeled data opens the possibility of applying land cover classification to new regions and time periods with minimal manual labeling.

In conclusion, Barlow Twins-based SSL offers a promising path forward in land cover classification with limited labeled data. This study is a foundation for future work exploring hybrid SSL and supervised methods tailored to various remote sensing domains. Expanding this work to include other modalities of remote sensing data, such as multispectral or SAR, can further validate the scalability and versatility of SSL frameworks in geospatial applications. The insights gained from our findings encourage continued exploration into self-supervised techniques, which hold the potential to revolutionize remote sensing analytics by overcoming the traditional reliance on large labeled datasets.

9 Repository

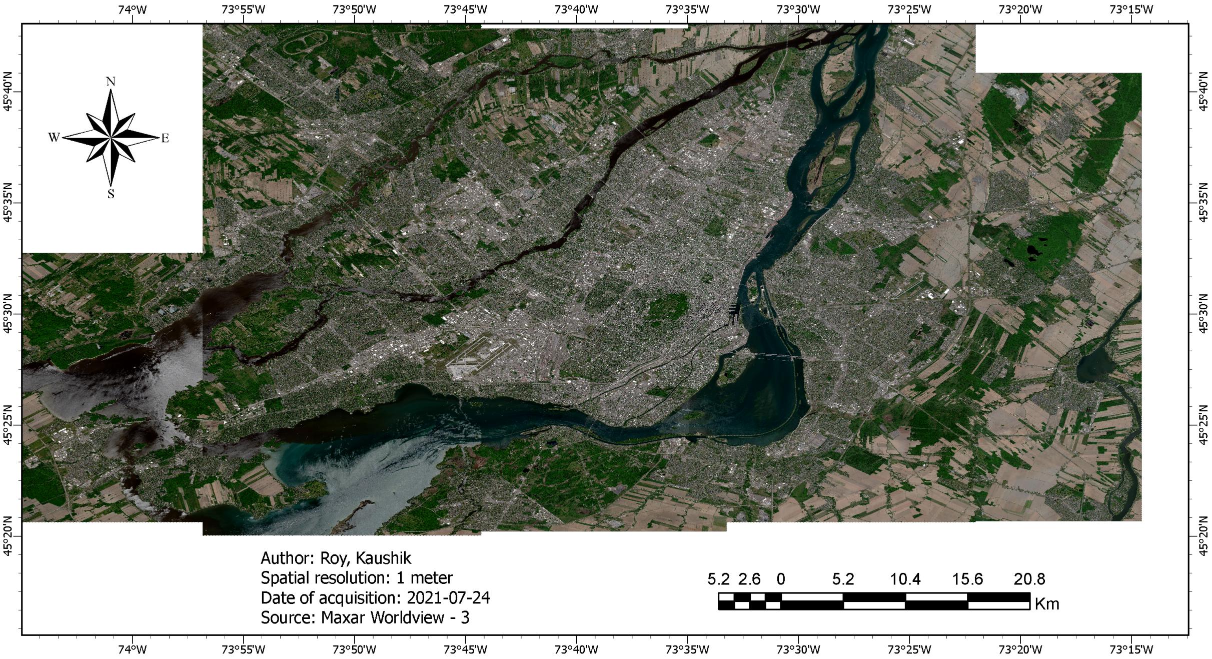
For more details, visit our repository at https://github.com/kaushikCanada/landcover-ssl: GitHub Repository.

10 Output



Author: Roy, Kaushik
Spatial resolution: 1 meter
Date of acquisition: 2021-07-21
Source: Maxar Worldview - 3

Figure 2: GTA Imagery used



Author: Roy, Kaushik
Spatial resolution: 1 meter
Date of acquisition: 2021-07-24
Source: Maxar Worldview - 3

Figure 3: GMA Imagery used

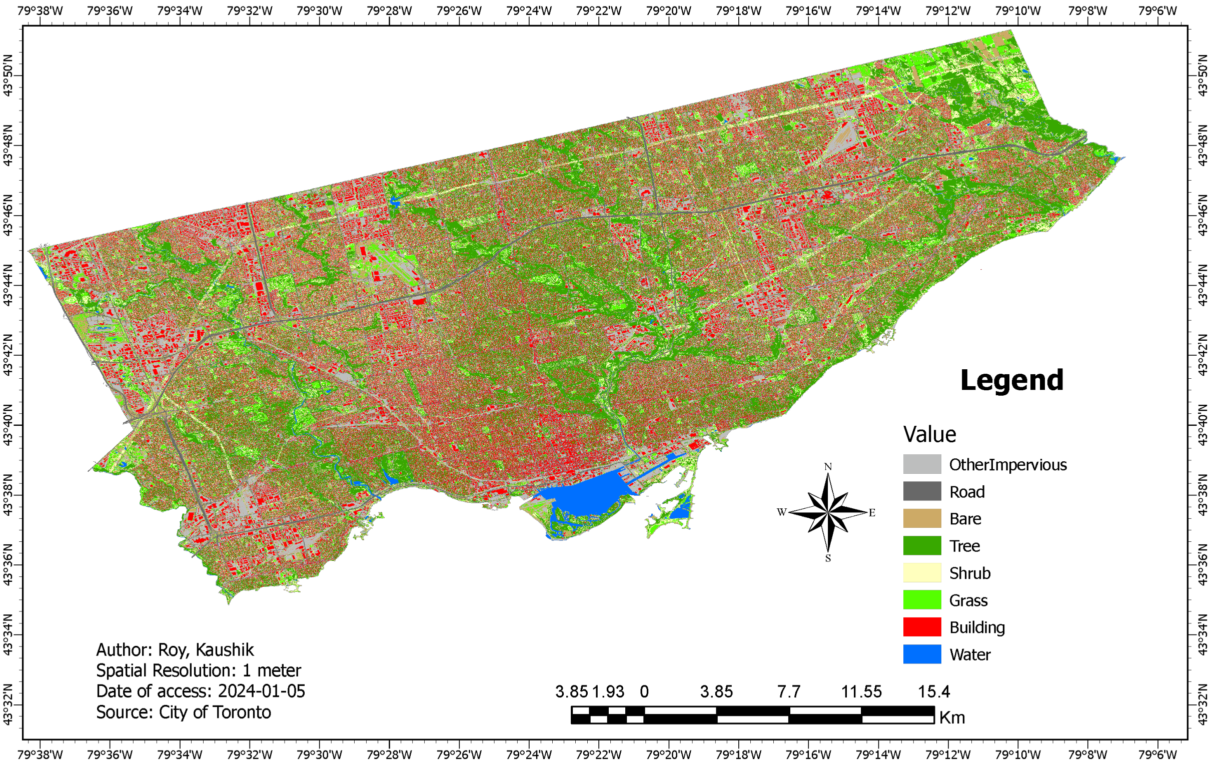


Figure 4: Ground Truth used

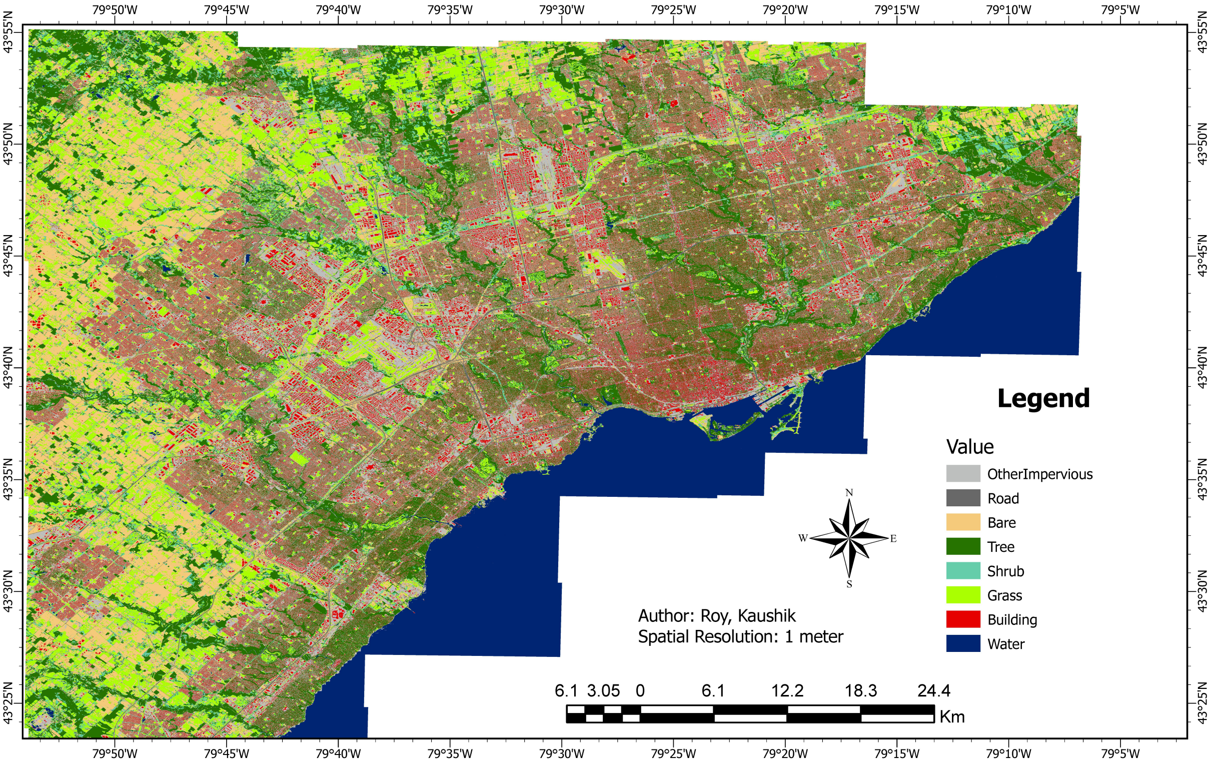


Figure 5: GTA Landcover obtained

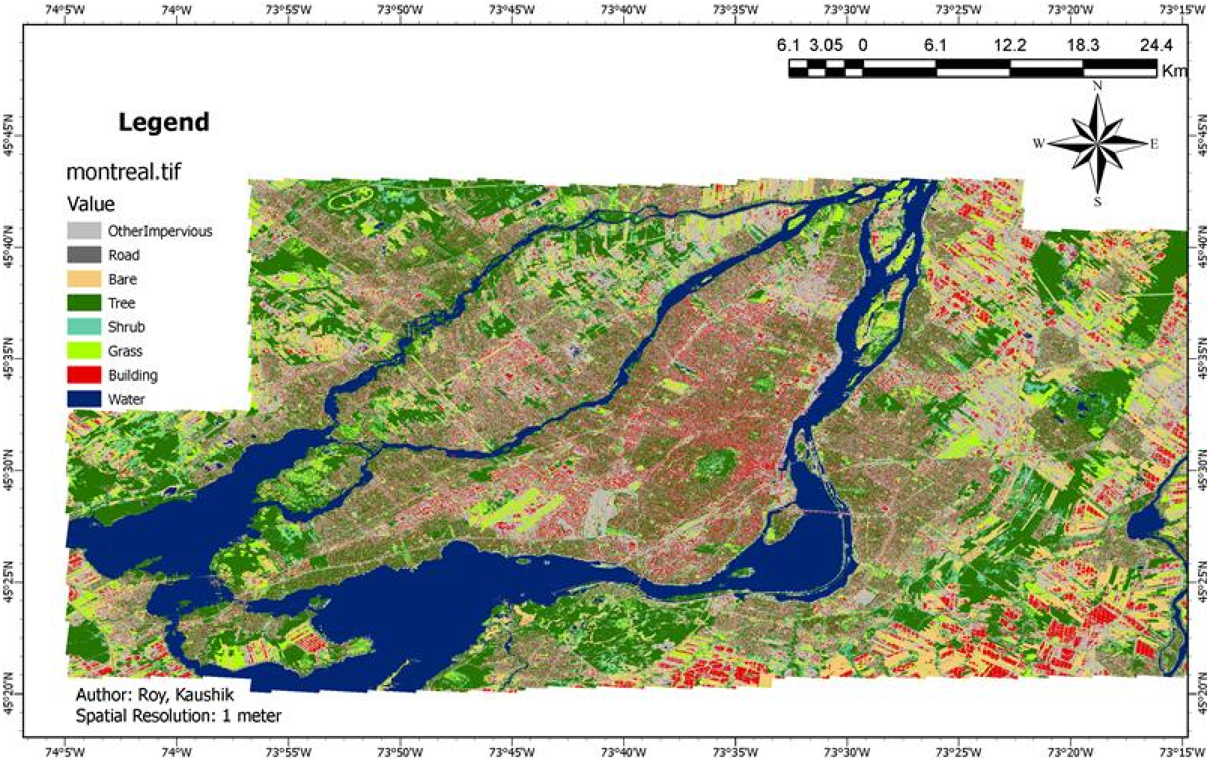


Figure 6: GMA Landcover obtained

References

Gong Cheng, Ceyuan Yang, Xiwen Yao, Lei Guo, and Junwei Han. When Deep Learning Meets Metric Learning: Remote Sensing Image Scene Classification via Learning Discriminative CNNs. *IEEE Transactions on Geoscience and Remote Sensing*, 56(5):2811–2821, May 2018. ISSN 0196-2892, 1558-0644. doi: 10.1109/TGRS.2017.2783902. URL `http://ieeexplore.ieee.org/document/8252784/`.

Neal Jean, Sherrie Wang, Anshul Samar, George Azzari, David Lobell, and Stefano Ermon. Tile2Vec: Unsupervised representation learning for spatially distributed data, 2018. URL `https://arxiv.org/abs/1805.02855`. Version Number: 2.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient Estimation of Word Representations in Vector Space, 2013. URL `https://arxiv.org/abs/1301.3781`. Version Number: 3.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation Learning with Contrastive Predictive Coding, 2018. URL `https://arxiv.org/abs/1807.03748`. Version Number: 2.

Adam J. Stewart, Caleb Robinson, Isaac A. Corley, Anthony Ortiz, Juan M. Lavista Ferres, and Arindam Banerjee. TorchGeo: deep learning with geospatial data. In *Proceedings of the 30th International Conference on Advances in Geographic Information Systems*, pages 1–12, Seattle Washington, November 2022. ACM. ISBN 978-1-4503-9529-8. doi: 10.1145/3557915.3560953. URL `https://dl.acm.org/doi/10.1145/3557915.3560953`.