Taylor & Francis
Taylor & Francis Group

Canadian Journal of Remote Sensing
Journal canadien de télédétection

ISSN: 0703-8992 (Print) 1712-7971 (Online) Journal homepage: www.tandfonline.com/journals/ujrs20

Self-Supervised Deep Learning for Urban Land Cover Classification from Very High Resolution Imagery

Apprentissage profond auto-supervisé pour la classification de la couverture terrestre urbaine à partir d'images à très haute résolution spatiale

Kaushik Roy, Saeid Homayouni & Ying Zhang

To cite this article: Kaushik Roy, Saeid Homayouni & Ying Zhang (2025) Self-Supervised Deep Learning for Urban Land Cover Classification from Very High Resolution Imagery, Canadian Journal of Remote Sensing, 51:1, 2532528, DOI: 10.1080/07038992.2025.2532528

To link to this article: https://doi.org/10.1080/07038992.2025.2532528

Published online: 19 Aug 2025.

CANADIAN JOURNAL OF REMOTE SENSING
2025, VOL. 51, NO. 1, 2532528
https://doi.org/10.1080/07038992.2025.2532528

Taylor & Francis
Taylor & Francis Group

Self-Supervised Deep Learning for Urban Land Cover Classification from Very High Resolution Imagery

Apprentissage profond auto-supervisé pour la classification de la couverture terrestre urbaine à partir d'images à très haute résolution spatiale

Kaushik Roy[a], Saeid Homayouni[a], and Ying Zhang[b]

[a]Eau Terre Environment (ETE), Institut National de la Recherche Scientifique (INRS), Quebec City, Canada; [b]Department of Geology, Natural Resources Canada (NRCan), Ottawa, Canada

ABSTRACT

Accurate mapping of land cover and land use at very high spatial 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. Existing studies have primarily employed low- to medium-resolution imagery and a few bands, which limits their downstream applicability. To our knowledge, this is the first attempt to study urban areas in Canada at such spatial resolution using self-supervised deep learning techniques. The objective of this study is to classify Worldview 3 multispectral imagery into eight urban land cover categories. The primary challenges are preparing analysis-ready data, addressing class imbalance, and having a limited amount of labelled data. To address these challenges, we introduce an innovative deep learning framework designed to enhance spectral-spatial consistency while leveraging the wealth of available unlabelled data for more effective learning and easily applying 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 labelled data. The proposed workflow is applied to Worldview 3 imagery patches of size 256 x 256 at a 1m spatial resolution. The methodology was applied to two UNet variants: a simple UNet and an attention-gated UNet with a ResNet-50 encoder. The results show that while the simple UNet could not adequately capture the complexity of the data, unlike the complex model. Self-supervised pretraining improved 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 demonstrate the effectiveness of multi-view self-supervised semantic segmentation on multispectral Worldview 3 images, creating a land cover product for future research. The code for the proposed architecture is publicly available at https://github.com/kaushikCanada/landcover-ssl.

RÉSUMÉ

Une cartographie précise de la couverture et de l'utilisation du territoire à très haute résolution spatiale (THR) est essentielle à l'étude du développement urbain et des interactions entre l'homme et l'environnement. Les techniques d'apprentissage profond, en particulier les modèles de segmentation sémantique, se sont révélées être des outils puissants pour cette tâche. Cependant, leur application généralisée est entravée par le besoin de nombreux jeux de données annotées à THR. Les études existantes ont principalement utilisé des images à basse et moyenne résolution spatiale et un nombre réduit de bandes, ce qui limite leur applicabilité. À notre connaissance, il s'agit de la première étude des zones urbaines au Canada à une telle résolution spatiale à l'aide de techniques d'apprentissage profond auto-supervisé. L'objectif de cette étude est de classifier l'imagerie multispectrale Worldview 3 en huit classes composant le milieu urbain. Les principaux défis étaient la préparation de données pour l'analyse, et une solution au déséquilibre des classes et à la limite du nombre de données étiquetées. Pour y remédier, nous avons introduit un cadre innovant

ARTICLE HISTORY
Received 17 February 2025
Accepted 30 June 2025

KEYWORDS
urban land cover classification; semantic segmentation; self-supervised learning; very high resolution

CONTACT Kaushik Roy ✉ kaushik.roy@inrs.ca

d'apprentissage profond conçu pour améliorer la cohérence spectrale-spatiale, tout en optimisant l'usage des données non étiquetées disponibles pour permettre un apprentissage efficace et faciliter l'application des représentations pré-entraînées dans les étapes subséquentes. Nous avons produit une image en super-résolution par pansharpening à l'aide de l'apprentissage profond. Ensuite nous avons extrait les caractéristiques latentes à l'aide d'un faible nombre de données étiquetées. Le flux de travail proposé a été appliqué à des imagettes Worldview 3 de taille 256 x 256 à une résolution spatiale de 1 m. La méthodologie a été appliquée à deux variantes d'UNet : un UNet simple et un UNet à accès contrôlée composé d'un encodeur ResNet-50. Les résultats montrent que, si l'UNet simple ne pouvait pas saisir toute la complexité des données, contrairement au modèle complexe. Le pré-entraînement auto-supervisé a amélioré la précision globale (PG) de la prédiction dans les deux cas. Pour les UNet simples, la précision est passée de 69 % à 74 %, et pour les UNet complexes, la PG est passée de 80 % à 88 %. En conclusion, nous démontrons l'efficacité de la segmentation sémantique auto-supervisée multi-vues sur des images multispectrales Worldview 3, créant ainsi une carte de la couverture terrestre pour les recherches futures. Le code de l'architecture utilisée dans cette étude est disponible à https://github.com/kaushikCanada/landcover-ssl.

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 reduction of urban natural disasters (Conway et al., 2020). 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 reference data requires extensive fieldwork, often involving teams of researchers and technicians with high labor 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, making it an essential tool for monitoring urban environments. Satellites such as WorldView-2, WorldView-3, and WorldView-4, as well as GaoFen-1 and GaoFen-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 Very High Resolution (VHR) imagery for detailed classification tasks (Qin and Liu, 2022). Effectively distinguishing urban land cover in VHR imagery is challenging due to the heterogeneous nature of urban landscapes. Approaches to classification generally fall into two categories: pixel-based and object-based analysis. 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 (Saboori et al., 2022). 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 the research gap mentioned above, this study employs a novel deep learning-based workflow that generates accurate land cover maps for urban areas in Canada using Very High Resolution (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 and 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 a pretrained Resnet50 backbone, and d) Resnet50_AttUnet + SSL - model c but with SSL pretrained 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 approaches such as super-resolution and self-supervised learning (SSL) to Worldview 3 imagery of a Canadian urban area. 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.
- 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.

Existing research

Recent advancements in remote sensing have sparked growing interest in applying self-supervised learning techniques to extract meaningful representations from satellite imagery. For instance, generative adversarial networks in Lin et al., (2017) 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 et al., 2020) introduced a novel pretext task that utilizes high-frequency channel information to predict RGB values, thereby facilitating feature learning in a self-supervised manner. Further extending this paradigm, (Moore and Hester, 2023) 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 labeled samples. In another study, (Wang et al., 2022) 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 et al., 2022) systematically analyzed the impact of various data augmentation techniques in contrastive self-supervised learning applied to remote sensing datasets. Meanwhile, (Ayush et al., 2020) proposed a self-supervised pre-training strategy that leverages the relationship between satellite imagery and geo-tagged sound recordings simultaneously. 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.

Among all the available techniques, contrastive learning is the most widely used 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 et al., (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 et al., 2018). This method was conceptually influenced by word2vec (Mikolov et al., 2013) and bore similarities to Contrastive Predictive Coding (CPC) (van den Oord et al., 2018). However, a significant challenge in contrastive learning arises from the tendency of models to collapse into trivial solutions when they are solely optimized 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 (Chen et al., 2020) and MoCo (He et al., 2019) 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 a practical time. BYOL (Grill et al., 2020) and SimSiam (Chen and He, 2020) are two other classes of contrastive SSL techniques that learn latent representations through knowledge distillation and have been successfully applied in remote sensing. Barlow Twins model (Zbontar et al., 2021) is a recent proposal in computer vision SSL and uses a novel loss function to reduce redundancy between learned representations. Although

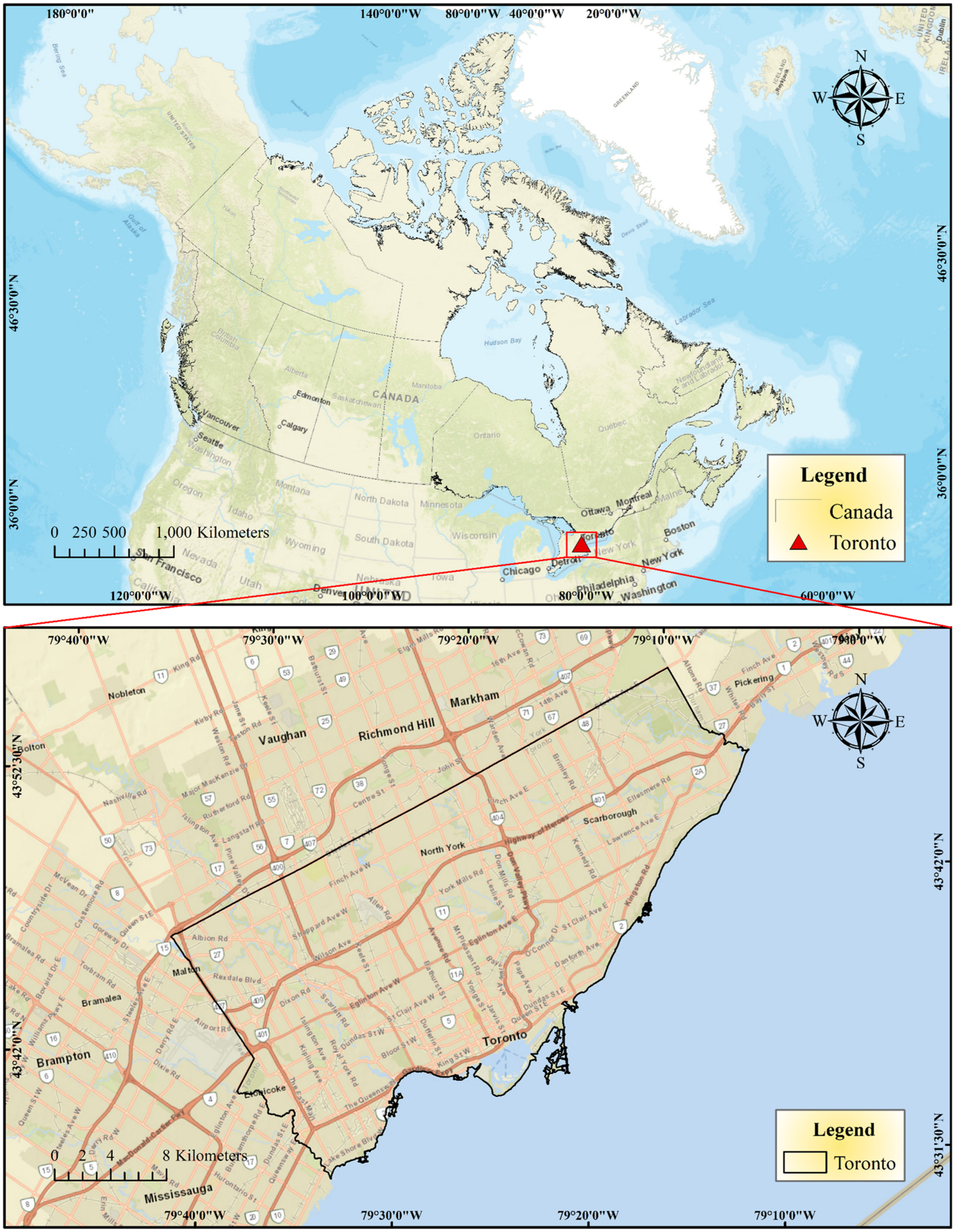


Figure 1. Map of the Study Area.

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.

Study area

The study focuses on the Greater Toronto Area (GTA) (Figure 1). Over the past five decades, GTA has experienced substantial growth due to international and domestic migration, leading to its development as a major multicultural urban center characterized by significant social, cultural, and economic diversity. 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 characterized by detached homes and cul-de-sacs. The Greater Toronto Area (GTA), includes the city of Toronto, and is home to near about 6.5 million people. Toronto itself has a population of 2,794,356, ranking as the most populous city in Canada and the fourth-largest in North America (Statistics Canada, 2021).

Remote sensing data

This study utilizes high-resolution (VHR) remote sensing imagery to classify urban land cover, with a focus on high-resolution satellite data. The imagery used 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.

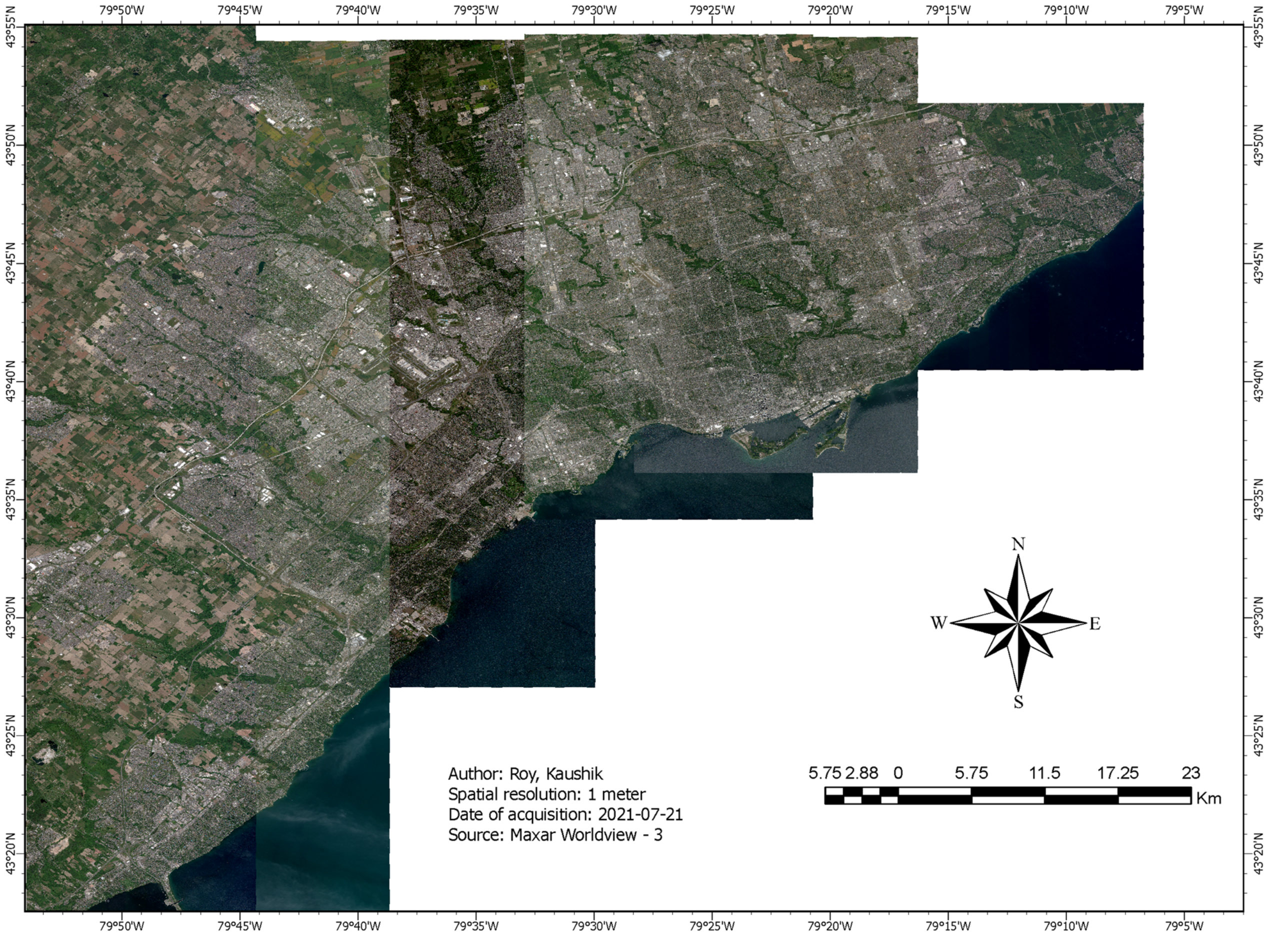


Figure 2. Worldview 3 data for study area.

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 mapping urban land. 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 (Figure 2).

Methodology

This paper proposes a deep learning-based end-to-end workflow (Figure 3) to achieve the research goals.

The proposed methodology (Figure 3) begins by combining a one-band panchromatic image with an 8-band multispectral image through deep learning pansharpening using LambdaPNN (Section 5.1). Following this, an analysis-ready dataset is created using ground reference labels with 8 classes (Section 5.2). Multiple views are generated using data augmentation and fed into a self-supervised pretraining using the Barlow Twins approach (Section 5.3), which generates a latent representation. Finally, supervised finetuning (Section 5.4) is used to compare different deep learning models and report results.

Deep learning pansharpening

Pansharpening is a technique used to enhance the spatial resolution of multispectral images by fuzing them with a higher-resolution panchromatic image. Traditionally, non-machine learning-based methods like component substitution, multi-resolution analysis, and optimization have been taxonomy mainstays since Wald (Wald et al., 1997) introduced this specific preprocessing technique. Recent research has focused on developing robust algorithms that utilize universal approximators, such as convolutional neural networks (CNNs), for image fusion tasks. These approaches demonstrate superior capabilities in capturing intricate, non-linear relationships, providing a high signal-to-noise ratio, and producing visually appealing, fused images.

The original pan-sharpening neural network PNN (Masi et al., 2016) inspired by SRCNN (Dong et al., 2015), has since been used as a basis for various innovative solutions like PanNet, DiCNN, SRPNN (Deng et al., 2022). The main challenge in this field is the quality assessment of the final output and the design of metrics for it. It suffers from a major problem: lack of ground reference data. In this study, a state-of-the-art deep learning model LambdaPNN (Figure 4) (Ciotola et al., 2023) is applied on the data. LambdaPNN uses ResNet blocks and CBAM modules to map low to high-resolution information. To visually compare with the DL output, the Gram-Schmidt pan-sharpening method (Maurer, 2013) was applied to the Blue, Green, Red, and NIR1 bands. The Gram-Schmidt pan-sharpening method Eq.(1) creates a synthetic panchromatic image P_{syn} by linearly combining selected multispectral bands:

$$P_{\text{syn}} = \sum_{i=1}^{N} w_i M_i \tag{1}$$

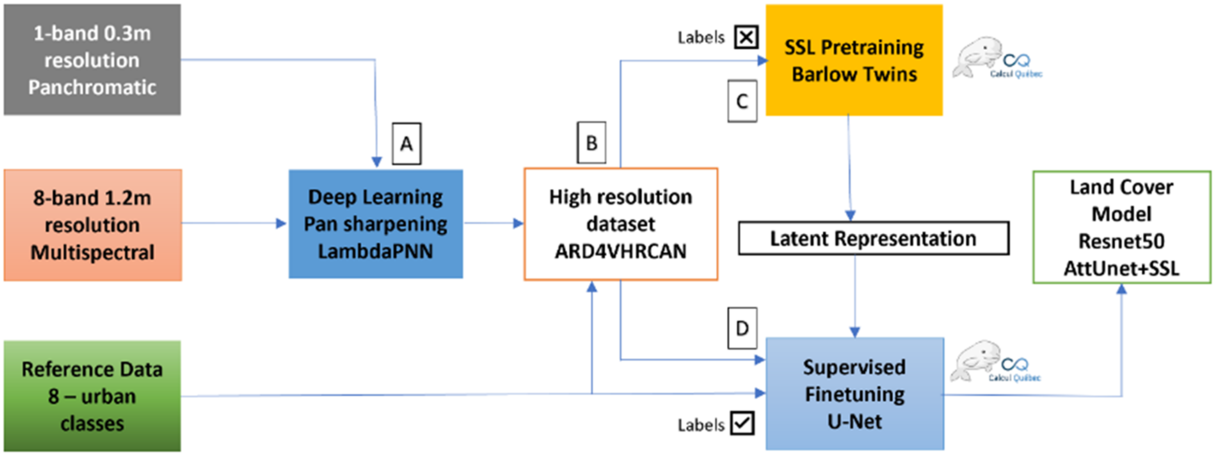

1-band 0.3m resolution Panchromatic
8-band 1.2m resolution Multispectral
Reference Data 8 – urban classes
A
Deep Learning Pan sharpening LambdaPNN
B
High resolution dataset ARD4VHRCAN
Labels ☒
C
SSL Pretraining Barlow Twins
Calcul Québec
Latent Representation
D
Supervised Finetuning U-Net
Labels ☑
Calcul Québec
Land Cover Model Resnet50 AttUnet+SSL

Figure 3. Complete Self Supervised Learning Workflow.

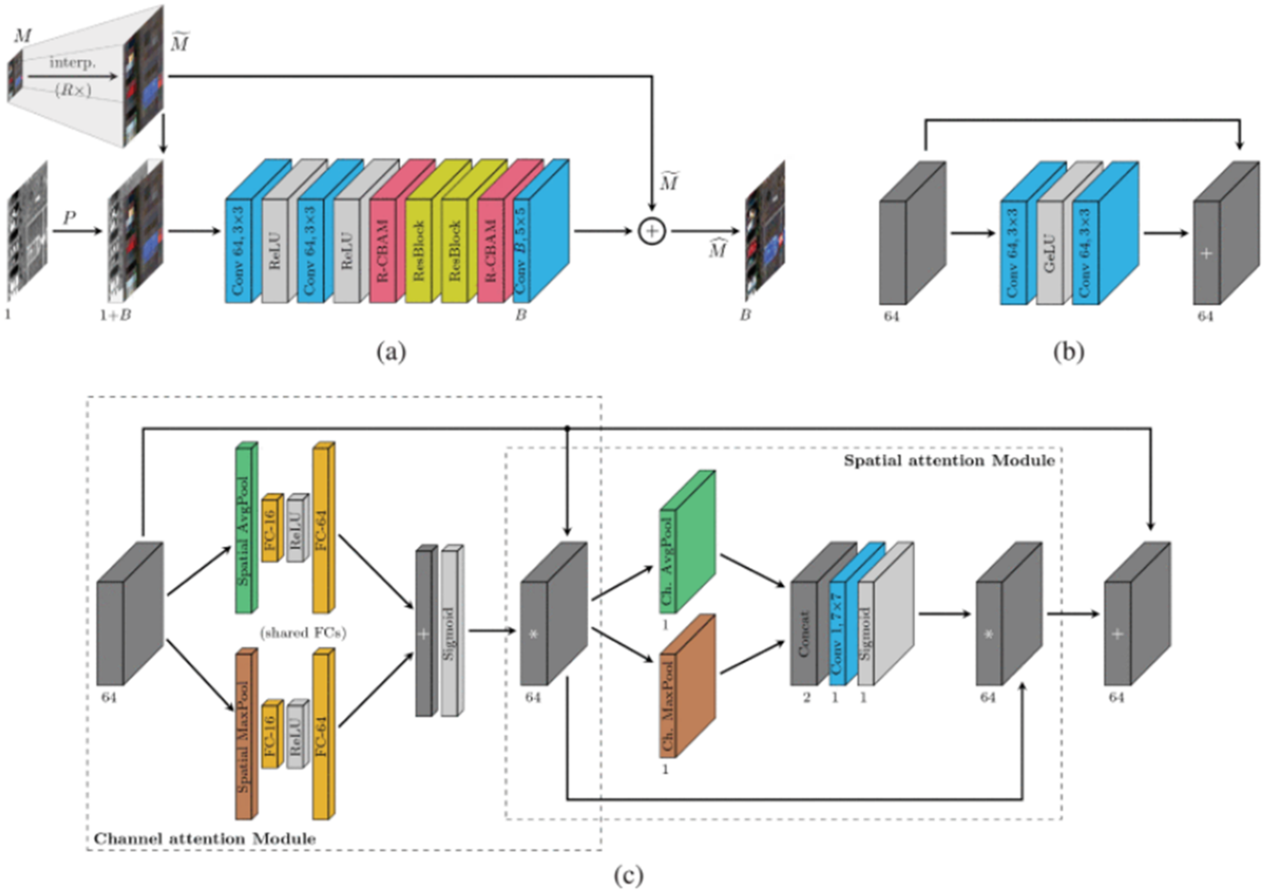


Figure 4. LambdaPNN network (Ciotola et al., 2023).

The multispectral bands are orthogonalized relative to P_{syn}, the PAN replaces P_{syn}, and an inverse transformation reconstructs the pan-sharpened output.

Analysis ready dataset preparation

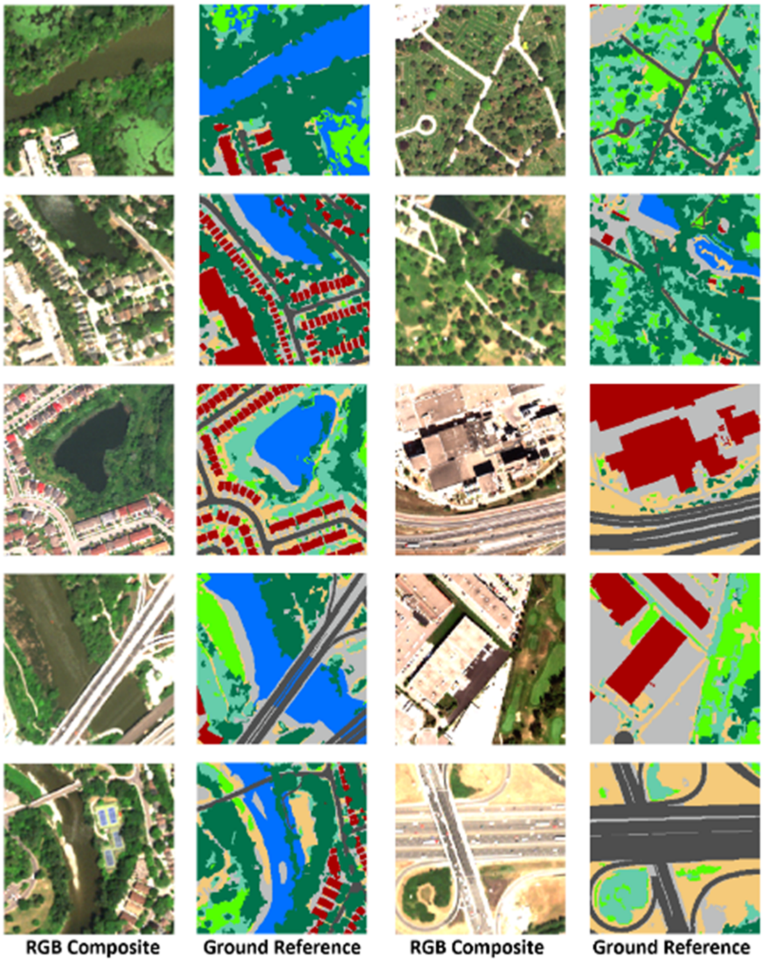
To conduct land cover classification, an analysis-ready dataset was prepared using WorldView-3 imagery acquired over the City of Toronto in July 2021. The imagery was pansharped and resampled to a final spatial resolution of 1 m to maintain consistency across all processing steps and to match the granularity of the target land cover classes. The target labels for this study were derived from the City of Toronto's 2022 Urban Land Cover dataset (City of Toronto, 2022), which offers a high-quality manually annotated vector layer consisting of eight semantic classes: other impervious surfaces, road, bare soil, trees, shrubs, grass, water, and building. The use of NDVI Eq.(2) (Rouse et al., 1974) and NDWI Eq.(3) (McFeeters, 1996) are widely recognized in remote sensing literature as an effective strategy to augment raw spectral bands for land cover classification tasks (Yaloveha et al., 2023), particularly when dealing with complex urban mosaics. The data was preprocessed to generate 256 × 256 non-overlapping patches with 10 input channels per patch. The split ratio was set at 70% for training, 15% for validation, and 15% for testing. The data preparation pipeline ensured that the imagery was precisely georegistered, spectrally enriched with derived indices, and accurately annotated with high quality reference labels. Table 1 shows the steps of the above process.

Table 1. Processing steps to obtain analysis ready dataset (ARD).

```
Algorithm 1
1: Input: Pansharpened WorldView-3 multispectral imagery
2: Input: City of Toronto land cover data
3: Output: ARD ready for deep learning semantic segmentation
4: Rasterize the city of Toronto land cover data at 1 m to create masks
5: Create a composite of Worldview 3 data and land cover masks for
   labeled areas
6: Validate metadata to check extent, spatial reference, and cell size
7: Create non-overlapping patches of 256×256 with the stride of 256
8: Remove all patches with no data pixels
9: Add NDVI and NDWI spectral indices to the data
10: Create random data splits for training, validation, and test
11: Visualize patches to confirm and attach metadata file for
    dissemination
```

$$\mathrm{NDVI} = \frac{NIR - R}{NIR + R} \tag{2}$$

$$\mathrm{NDWI} = \frac{G - NIR}{G + NIR} \tag{3}$$



ID	Color	Landcover
1	⬜	Impervious
2	⬛	Road
3	🟨	Bare
4	🟩	Tree
5	🟩	Shrubs
6	🟩	Grass
7	🟥	Building
8	🟦	Water

Figure 5. Training data with land cover classes.

Self-Supervised learning

Images often contain a lot of information that is irrelevant for further downstream tasks. *A representation* of an image ideally extracts the relevant parts of it (Figure 5). The goal of *representation learning* is to learn an *encoder network* f_θ with learnable parameters θ that maps an input image x to a lower-dimensional representation (embedding) $y = f_\theta(x)$.

In a nutshell, such methods build an encoder by performing the following two distinct steps:

i. Formulate a supervised learning task by generating an output t corresponding to each input image x.
ii. Train the model using supervised learning to map inputs x to their associated targets t.

Given a dataset of images, we write

$$X = [x_1, \ldots, x_n] \tag{4}$$

for a randomly sampled batch of images. Every representation learning method trains an encoder network f_θ, where θ are the learnable parameters. This encoder network computes a representation

$$Y = [y_1, \ldots, y_n] = [f_\theta(x_1), \ldots, f_\theta(x_n)] = f_\theta(X) \tag{5}$$

of the images in X. Some methods additionally train a projection network g_ϕ, with parameters ϕ, that computes projections

$$Z = [z_1, \ldots, z_n] = [g_\phi(y_1), \ldots, g_\phi(y_n)] = g_\phi(Y) \tag{6}$$

of the representations in Y. Some methods also train a prediction network q_ψ, with parameters ψ, that computes a prediction based on z or y. Both projections and predictions are only used to train the network and after training the projection and prediction networks are discarded, and only the encoder f_θ is used for downstream tasks.

Multiple views of data using augmentation

The **invariance principle** in self-supervised learning posits that an effective model should learn representations that remain consistent under various transformations of the input data. This principle is critical in tasks such as image classification, object detection, and land cover segmentation, where the goal is to extract meaningful features robust to input changes. For this purpose, we use our unlabeled data. We artificially create noisy versions (views) of our data using different spatial, spectral, and geometric transformations.

Mathematically, the invariance principle can be expressed as follows:

$$Z(X) \approx Z(T(X)) \tag{7}$$

Table 2. Augmentations compared with original paper and ours.

View #	Name	Original	Ours
1	Random resized crop	Yes	Yes
2	Random rotation	Yes	Yes
3	Random horizontal flip	Yes	Yes
4	Random vertical flip	Yes	Yes
5	Random box blur	Yes	Yes
6	Random contrast	Yes	No
7	Random solarization effect	Yes	No
8	Random brightness change	Yes	No
9	Random color jitter	Yes	No
10	Random grayscale effect	Yes	No
11	Random view of angle	No	Yes
12	Random shuffling of channels	No	Yes
13	Random shuffling of patches	No	Yes

where:

- $Z(X)$ is the representation learned from the original input X,
- $T(X)$ represents a transformation (or augmentation) applied to X, such as rotation, scaling, cropping, or color jittering,
- The approximation symbol $\approx$ indicates that the representations for the original and augmented inputs should be similar.

Table 2 below shows a list of augmentations used in this study compared to the original Barlow twins implementation (Zbontar et al., 2021). Some operations are only applicable to 3-channel data, so we replaced them with more meaningful transformations for multispectral data. Figure 6 shows the original patch and its multiple views generated using different transformations.

Barlow Twins model on unlabeled data

In this work, we create a new SSL4EO model using a new paradigm called Barlow Twins (Figure 7). The central idea behind this framework is the principle of *redundancy reduction* (Zbontar et al., 2021). This principle states that reducing redundancy is crucial for organizing sensory messages in the brain. To implement this redundancy reduction principle, our approach takes a batch of images X and creates two

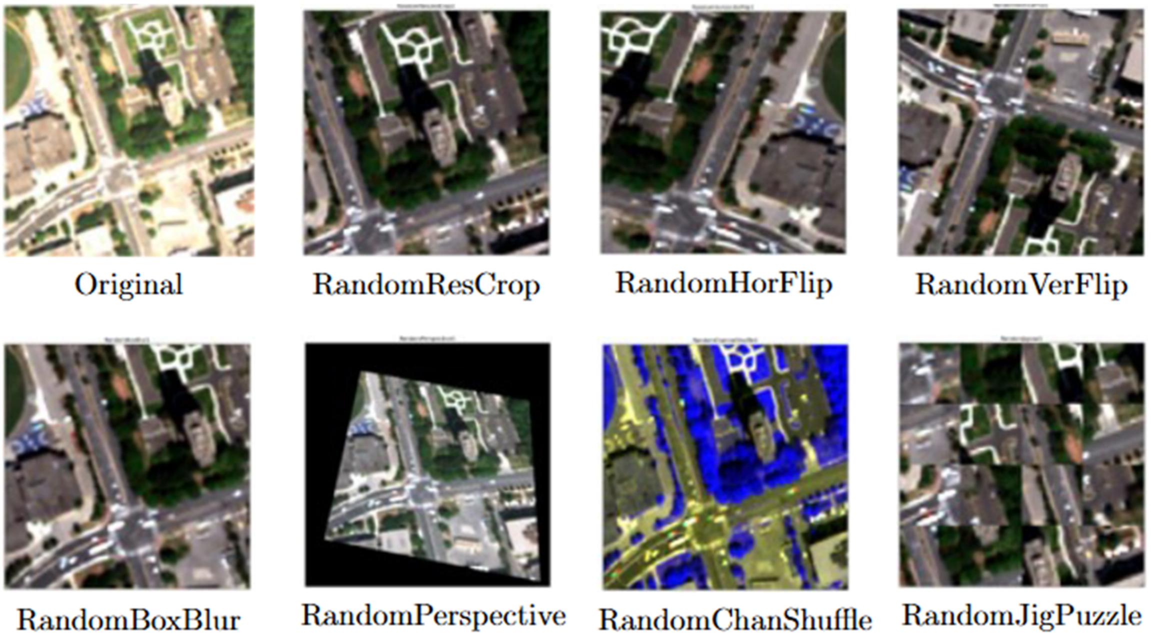


Figure 6. Visualizing multiple views generated from data augmentation.

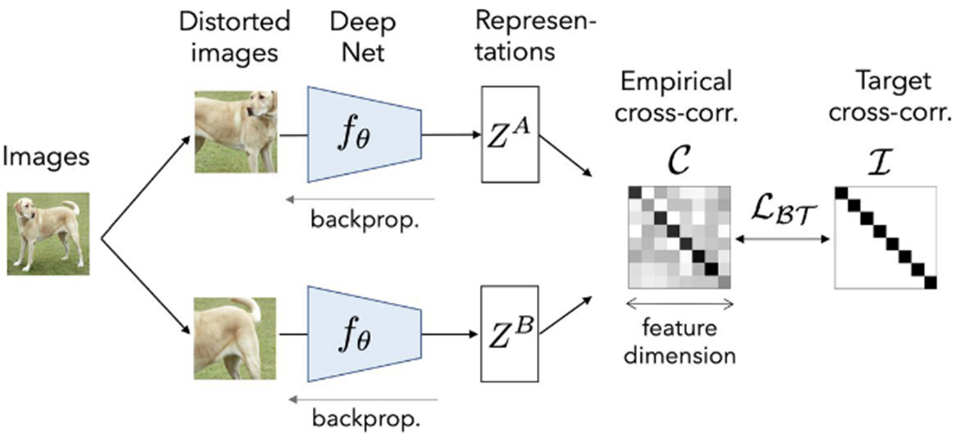


Figure 7. Barlow Twins strategy (Zbontar et al., 2021).

views $X^{(1)} = t^{(1)}(X)$ and $X^{(2)} = t^{(2)}(X)$ of these images, where $t^{(1)}, t^{(2)} \sim \mathcal{T}$ are transformations randomly sampled from $\mathcal{T}$ for every image of the batch. A Siamese encoder f_θ computes representations

$$Y^{(1)} = f_\theta(X^{(1)}) \quad \text{and} \quad Y^{(2)} = f_\theta(X^{(2)}),$$

which are fed into a Siamese projector g_ϕ to compute projections

$$Z^{(1)} = \left[z_1^{(1)}, \ldots, z_n^{(1)}\right] = g_\phi(Y^{(1)}), \quad Z^{(2)} = \left[z_1^{(2)}, \ldots, z_n^{(2)}\right] = g_\phi(Y^{(2)})$$

for both views. Figure 7 below shows the Barlow Twins architecture.

Following this we regularize the cross-correlation matrix between the projections of both views. The cross-correlation matrix is calculated as

$$C = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{z_i^{(1)} - \mu^{(1)}}{\sigma^{(1)}}\right)\left(\frac{z_i^{(2)} - \mu^{(2)}}{\sigma^{(2)}}\right)^{\top}, \tag{8}$$

where $\mu^{(j)}$ and $\sigma^{(j)}$ are the mean and standard deviation over the batch of projections of the j-th view, calculated as

$$\mu^{(j)} = \frac{1}{n}\sum_{i=1}^{n} z_i^{(j)}, \tag{9}$$

$$\sigma^{(j)} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(z_i^{(j)} - \mu^{(j)}\right)^2}. \tag{10}$$

The key steps in implementing the Barlow Twins are as follows:

1. **Data Augmentation:** Apply random augmentations (e.g., random cropping, color jittering) to the input images to create two distorted views.
2. **Embedding Network:** Use a neural network (e.g., ResNet) to encode the distorted views into feature vectors.
3. **Cross-Correlation Matrix:** Compute the cross-correlation matrix between the feature vectors of the two views.
4. **Loss Function:** Minimize the Barlow Twins loss, which consists of two terms: the invariance term that encourages the diagonal elements of the cross-correlation matrix to be close to 1, and the redundancy reduction term that encourages the off-diagonal elements to be close to 0.

The Barlow Twins loss function is defined as:

$$\mathcal{L}_{\mathrm{BT}} = \sum_{k=1}^{d}(1 - C[k,k])^2 + \lambda\sum_{k=1}^{d}\sum_{k' \neq k} C[k,k']^2, \tag{11}$$

where d is the number of dimensions of the projection and $\lambda > 0$ is a hyperparameter. The first term promotes invariance about the applied transformations, and the second term decorrelates the learned embeddings, i.e., reduces redundancy. By using this loss, the encoder f_θ is encouraged to predict embeddings that are decorrelated and thereby non-redundant.

Downstream supervised learning

Attention Gated U-net

The Attention Gated U-Net (Oktay and Schlemper, 2018) (Figure 8) is an extension of the traditional U-Net architecture that incorporates attention mechanisms to improve segmentation performance. The key components of the Attention Gated U-Net are:

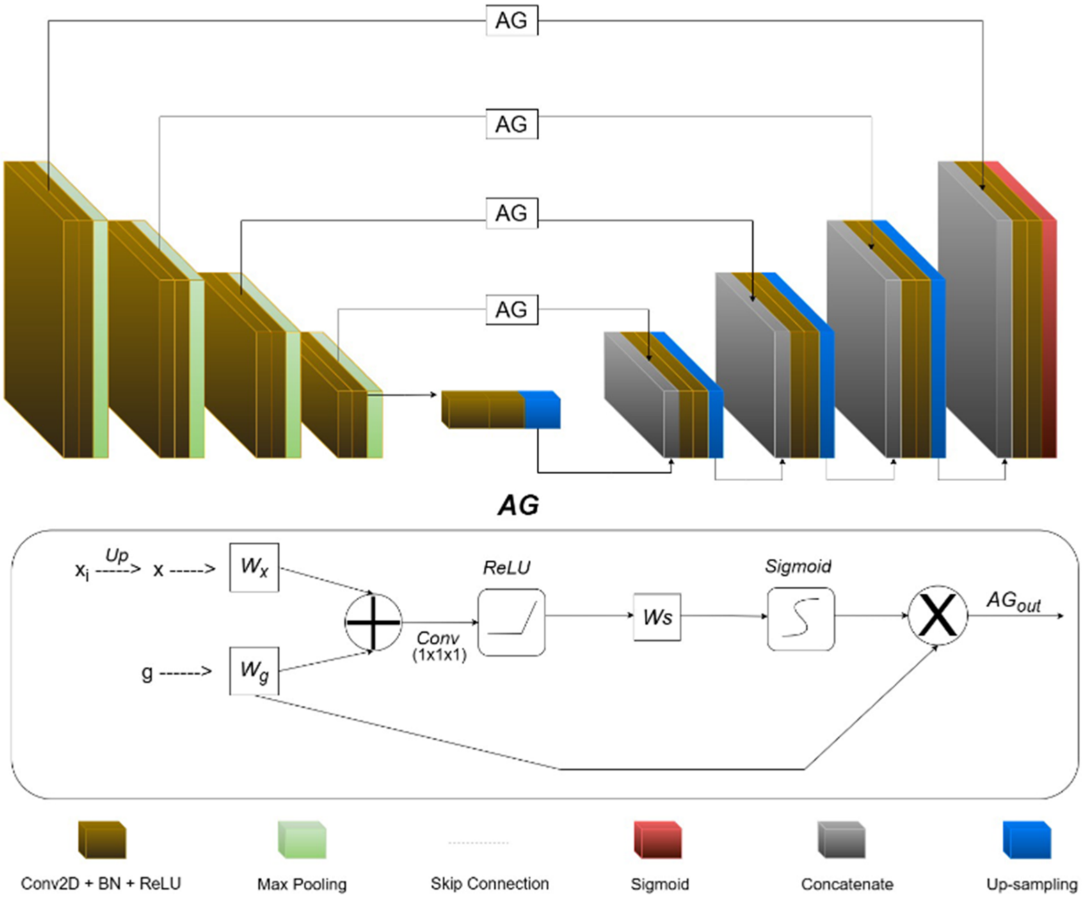


Figure 8. Attention Gated Unet (Oktay and Schlemper, 2018).

1. **Encoder:** The encoder part of the U-Net, which consists of convolutional layers followed by max-pooling layers, is used to extract features from the input image.
2. **Attention Gates:** Attention gates are introduced at each skip connection between the encoder and decoder. These gates compute a gating signal that modulates the feature maps from the encoder, focusing on the relevant regions of the image.
3. **Decoder:** The decoder part of the U-Net, which consists of up-convolutional layers, is used to reconstruct the segmentation map from the encoded features.

The attention gate α is computed as:

$$\alpha = \sigma(W_g * g + W_x * x + b) \tag{12}$$

where g is the gating signal from the decoder, x is the feature map from the encoder, W_g and W_x are learnable weights, b is a bias term, and σ is the sigmoid function. The modulated feature map $\tilde{x}$ is then computed as:

$$\tilde{x} = \alpha \cdot x \tag{13}$$

Finetuning U-Nets on labeled data

Fine-tuning the model is achieved through supervised learning. The pretraining process was guided by the Barlow Twins loss function (Equation 11). However, the U-Net models are subsequently fine-tuned using a segmentation loss function, L, which is computed as the mean of the binary cross-entropy loss.

The binary cross-entropy loss for semantic segmentation is defined as:

$$L_{BCE} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C}\left(y_{i,c}\log(p_{i,c}) + (1-y_{i,c})\log(1-p_{i,c})\right) \tag{14}$$

where:

- N is the total number of pixels,
- C is the number of classes (land cover types),
- $y_{i,c}$ is the true binary label for pixel i belonging to class c,
- $p_{i,c}$ is the predicted probability of pixel i being in class c.

Mean IoU for semantic segmentation is defined as:

$$\mathrm{IoU} = \frac{1}{C}\sum_{c=1}^{C}\frac{TP_c}{TP_c + FP_c + FN_c} \tag{15}$$

where:

- C is the number of classes (land cover types),
- TP_c is the number of true positives for class c,
- FP_c is the number of false positives for class c,
- FN_c is the number of false negatives for class c.

The F1 score for semantic segmentation is defined as:

$$F1 = 2 \cdot \frac{\mathrm{Precision} \cdot \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}} \tag{16}$$

where:

- Precision is defined as $\frac{TP}{TP + FP}$,
- Recall is defined as $\frac{TP}{TP + FN}$.

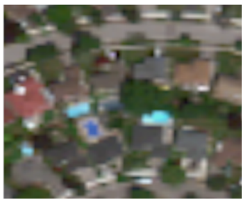
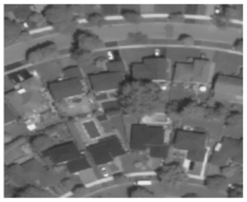
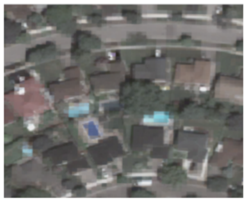
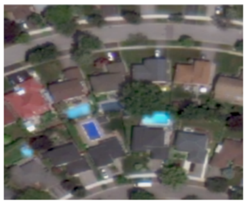
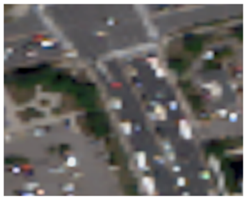
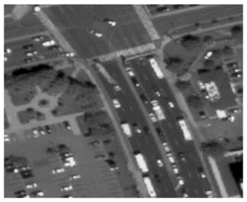
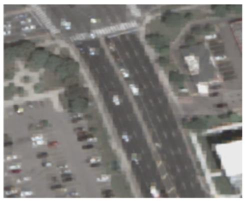
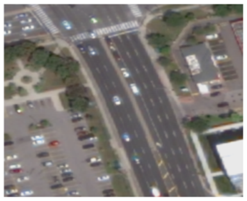
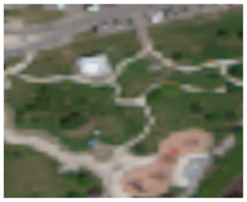
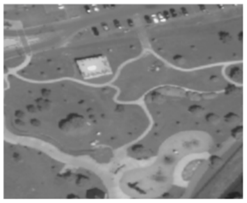
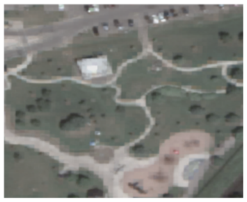
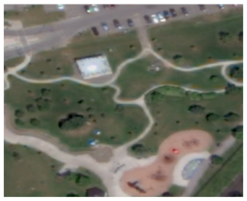
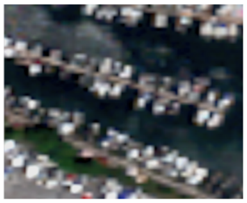
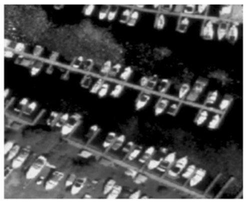
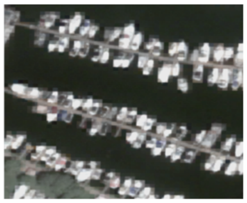
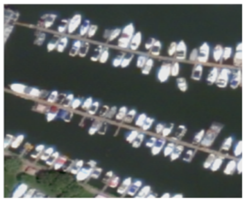
The overall accuracy (OA) is a common metric used in classification problems to measure the proportion of correctly classified samples among the total number of samples. It can be expressed as:

$$OA = \frac{\sum_{i=1}^{C} TP_i}{N} \tag{17}$$

where:

- TP_i is the number of true positives for class i,
- C is the total number of classes,
- N is the total number of samples.

Table 3. Qualitative comparison of deep learning and Gram-Schmidt pansharpening.

Multispectral	Panchromatic	Gram-Schmidt	Deep Learning
			
			
			
			

In this equation, the numerator represents the sum of correctly classified samples for all classes, and the denominator is the total number of samples.

Results and discussion

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 on 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 et al., 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.

Pansharpening

The results from the pansharpening phase are presented in Table 3. The multispectral and panchromatic data, as well as the outputs from Gram-Schmidt and deep learning-based pansharpening, are shown here. Visual inspection reveals the output from deep learning is more vibrant and sharper than classically pan-sharpened products. This is mainly because we can leverage information from all eight bands and apply a state-of-the-art model, such as LambdaPNN. The output shows that while fine objects, such as swimming pools, boats, and cars, appear hazy in the Gram-Schmidt output, these objects are clearer in the new deep learning pansharpened product.

Table 4. Comparison of U-Net and Attention U-Net with ResNet50 backbone.

Stage	Standard U-Net	Attention U-Net (ResNet50)
Encoder 1	64	64
Encoder 2	128	256
Encoder 3	256	512
Encoder 4	512	1024
Bottleneck	1024	2048
Decoder 1	512	1024
Decoder 2	256	512
Decoder 3	128	256
Decoder 4	64	64
Skip Conn.	Direct concat.	Attention gate + concat.
Attention	None	At each skip conn.

Land cover classification

The proposed framework generates a segmentation mask for a given multispectral image. Table 4 enumerates the encoder and decoder configurations for the two Unet variants used in experiments. The dataset consisted of 8,800 labeled image patches, each of size 256×256, with 10 spectral bands and corresponding single-channel land cover masks, comprising 8 classes. The dataset was randomly divided into training, validation, and test sets using a stratified approach to ensure class balance across splits. 70% of the data (6,160 images) was used for training, 15% (1,320 images) for validation, and 15% (1,320 images) for testing. All splits were mutually exclusive and fixed across all experiments. To optimize the model's performance, the Adam optimizer is used in all experiments, with a learning rate initialized at 1×10^{-3}. This learning rate is reduced whenever training stagnates, thereby enhancing segmentation performance. Moreover, early stopping is implemented to prevent overfitting, ensuring that training halts once the loss function stops improving. Tables 5 and 6 present the qualitative comparison of the segmentation performance. Moreover, the quantitative results of both U-Net models with and without the Barlow Twins-based pre-training are presented in Tables 7 and 8. The models were trained for 100 epochs, and inference was performed on the test set.

It is observed that initially, the simple Unet models (32 M parameters) were unable to 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 (43 M parameters) can better classify the impervious area mostly due to the Resnet50 encoder and the attention-gating mechanism. However, the output still contains some noise. Nevertheless, after performing 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 through SSL training, as Resnet50AttUnet + SSL can capture finer details and extract precise road networks, along with more

Table 5. Qualitative comparison of model outputs.

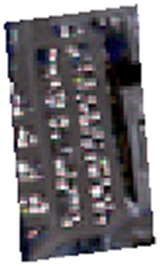
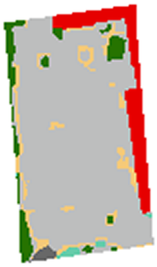
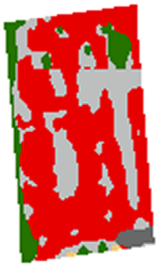
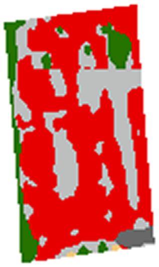
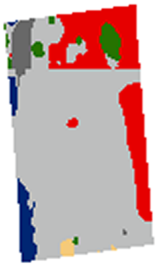
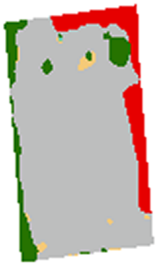
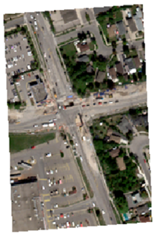
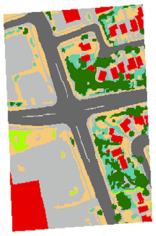
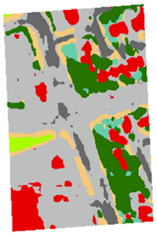
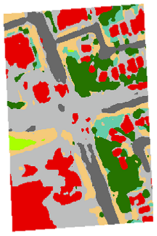
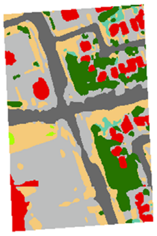
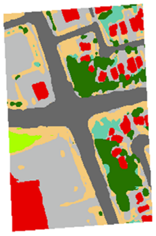
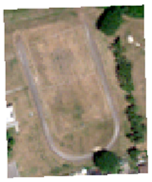
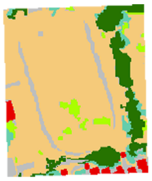
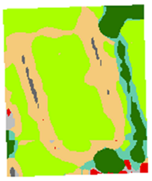
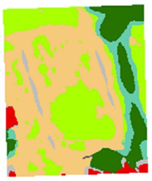
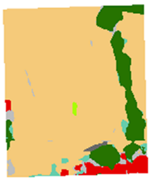
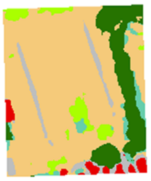
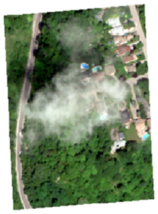
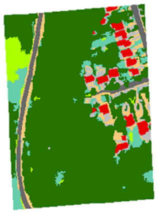
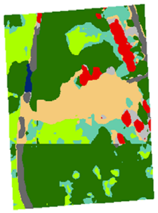
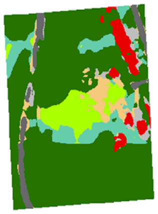
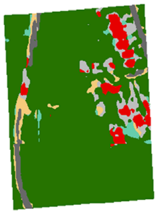
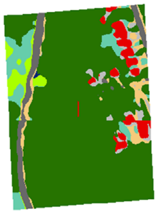
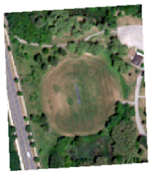
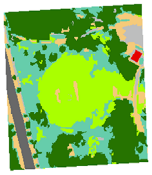
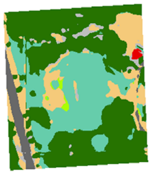
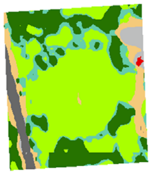
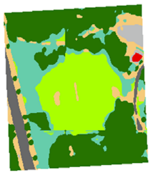
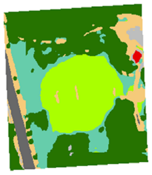
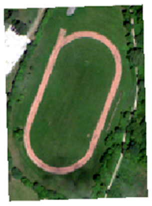
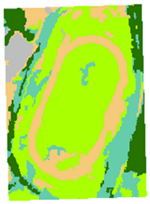
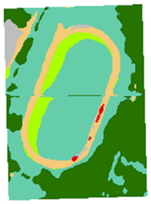
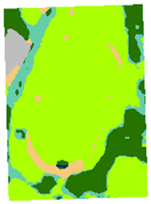
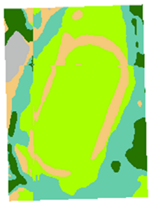
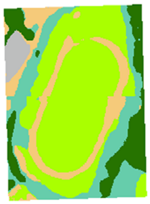
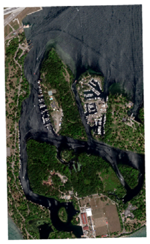
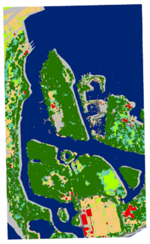
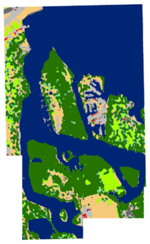
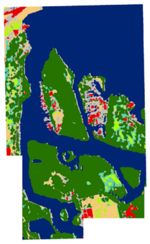
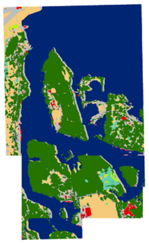
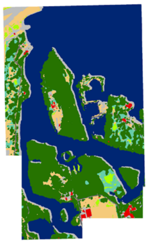
ULC	Color	Input	Ground Reference	Unet	Unet + SSL	Resnet50 +AttUnet	Resnet50 +AttUnet + SSL
Other	■						
Road	■						
Bare	■						
Trees	■						
Shrubs	■						
Grass	■						
Water	■						

Table 6. Qualitative comparison of model outputs.

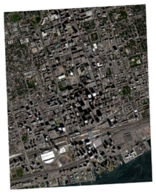
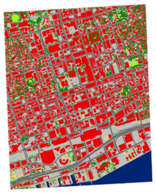
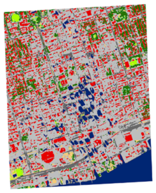
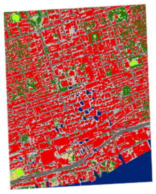
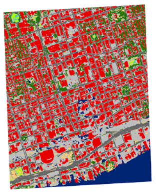
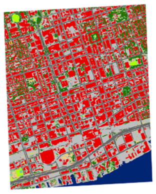
ULC	Color	Input	Ground Reference	Unet	Unet + SSL	Resnet50 +AttUnet	Resnet50 +AttUnet + SSL
Building	🟥						

Table 7. Validation metrics.

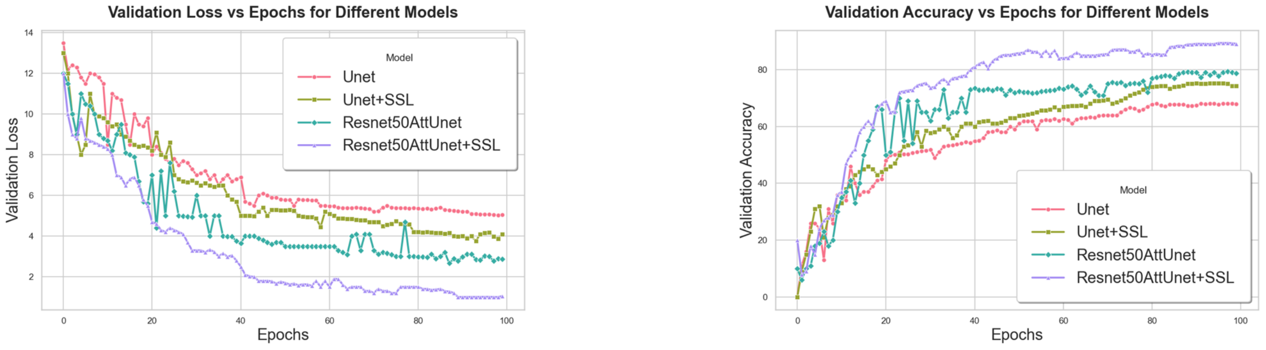


Table 8. Quantitative comparison of model outputs.

Model Name	IoU	F1 Score	OA
Unet	0.56	0.61	0.69
Unet + SSL	0.65	0.69	0.74
Resnet50AttUnet	0.71	0.74	0.80
Resnet50AttUnet + SSL	0.76	0.83	0.88

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 rather than a 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 is helpful in this case, and Resnet50AttUnet + SSL can accurately 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 incorrectly classifies it as 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 some spurious water pixels remain 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

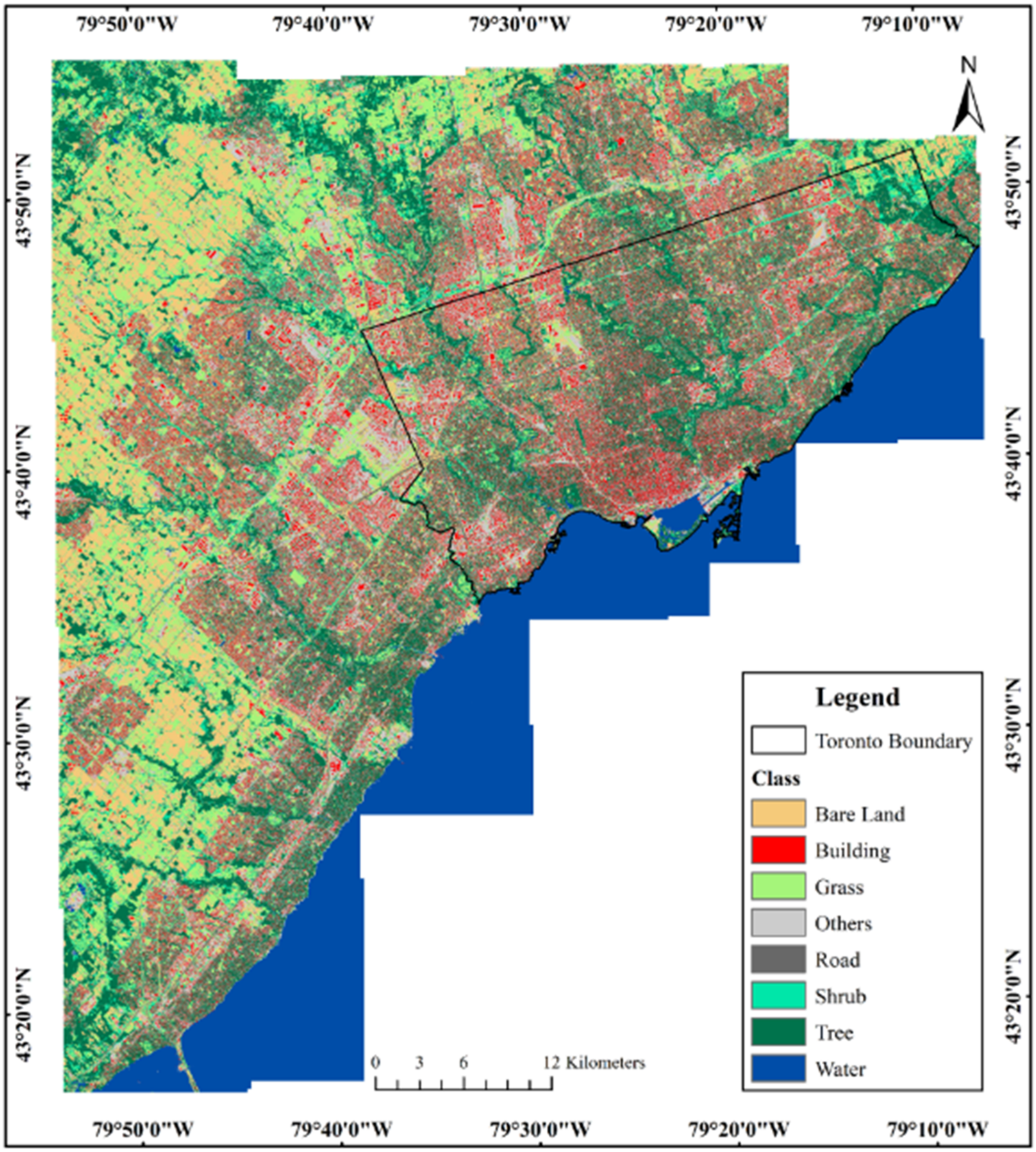


Figure 9. Final land cover output.

shadow pixels and properly classify them as either buildings or roads (Figure 9).

The table above compares the performance of four models on land cover segmentation tasks, quantified by Intersection over Union (IoU), F1 Score, and Overall Accuracy (OA). The baseline U-Net model demonstrates relatively low performance across all metrics, with an IoU of 0.56, an F1 score of 0.61, and an overall accuracy (OA) of 0.69. These results indicate the challenges posed by limited labeled data and the simplicity of U-Net's encoding architecture. The encoder in the basic U-Net is not powerful enough to capture the rich spatial patterns required for high-quality land cover segmentation, especially when the availability of labeled samples is constrained. Introducing self-supervised pre-training using Barlow Twins significantly boosts the performance of the U-Net, as evidenced by the metrics of U-Net + SSL (IoU = 0.65, F1 = 0.69, OA = 0.74). The substantial improvements across all validation metrics underline the efficacy of self-supervised learning (SSL) in addressing data scarcity. By leveraging SSL pre-training, the encoder learns more robust and generalized features from the data, thereby enhancing the downstream segmentation task without requiring extensive labeled samples. Moreover, the ResNet50AttU-Net model, which incorporates a ResNet50 backbone and attention mechanisms, outperforms the U-Net and U-Net + SSL models. The IoU improves to 0.71, the F1 score to 0.74, and the OA to 0.80. This superior performance can be attributed to the more powerful and deeper ResNet50 encoder, which facilitates better feature extraction. Additionally, the attention gates integrated into this architecture allow the model to focus on the most relevant spatial information, further refining segmentation accuracy. Finally, the ResNet50AttU-Net + SSL model achieves the highest performance, with an IoU of 0.76, an F1 score of 0.83, and an overall accuracy (OA) of 0.88. This indicates the positive synergy between the attention-gated ResNet50 architecture and SSL pre-training. The sophisticated encoder, attention mechanisms, and SSL enable this model to extract and utilize feature representations more effectively, particularly in scenarios with limited labeled data, culminating in the best overall segmentation results.

These results highlight the significant advantages of integrating SSL pre-training into encoder-decoder segmentation models and the utility of attention mechanisms in improving the precision and accuracy of land cover segmentation tasks.

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 to learn 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 pan-sharpening 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 in terms of 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 with constrained resources. Efficient usage of unlabeled data opens the possibility of applying land cover classification to new regions and periods with minimal manual labeling. However, we find that the most frequent sources of misclassification in our LULC map are: (1) shadow artifacts, where areas shaded by tall buildings or dense tree canopies exhibit spectral signatures similar to water or impervious surfaces; (2) mixed-pixel effects, which occur along boundaries between built and natural covers (e.g., roof–vegetation edges) and lead to ambiguous class assignments; and (3) spectral confusion between spectrally similar materials—most notably bare soil and concrete or asphalt; and (4) cloud contamination. Addressing these issues will require improved shadow-compensation algorithms, sub-pixel classification techniques, and the incorporation of additional spectral or contextual features in future model iterations.

In conclusion, Barlow Twins-based SSL presents a promising approach for 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.

Acknowledgements

The authors gratefully acknowledge Natural Resources Canada for providing the very high-resolution satellite imagery, procured from Maxar Technologies, and the various agencies and field teams that collected the ground reference data used in this study. We also thank the developers of the open-source software libraries PyTorch Lightning and TorchGeo. We further appreciate the support of NVIDIA's Academic Grant Program, whose donation of two NVIDIA Quadro RTX A6000 GPUs to the INRS Environmental and Northern Remote Sensing Laboratory made the deep learning experiments possible.

Disclosure statement

No potential conflict of interest was reported by the author(s).

References

Ayush, K., Uzkent, B., Meng, C., Tanmay, K., Burke, M., Lobell, D., and Ermon, S. 2020. Geography-aware self-supervised learning. https://arxiv.org/abs/2011.09980. Version Number: 7.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. 2020. A simple framework for contrastive learning of visual representations. https://arxiv.org/abs/2002.05709. Version Number: 3.

Chen, X., and He, K. 2020. Exploring simple Siamese representation learning. https://arxiv.org/abs/2011.10566. Version Number: 1.

Cheng, G., Yang, C., Yao, X., Guo, L., and Han, J. 2018. "When deep learning meets metric learning: remote sensing image scene classification via learning discriminative CNNs." *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 56 (No. 5): pp. 2811–2821. May. ISSN 0196-2892, 1558-0644. http://ieeexplore.ieee.org/document/8252784/. doi:10.1109/TGRS.2017.2783902.

Ciotola, M., Poggi, G., and Scarpa, G. 2023. "Unsupervised deep learning-based pansharpening with jointly enhanced spectral and spatial fidelity." *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 61 pp. 1–17. ISSN 0196-2892, 1558-0644. https://ieeexplore.ieee.org/document/10198408/. doi:10.1109/TGRS.2023.3299356.

City of Toronto. 2022. High resolution land cover dataset for Toronto with eight land cover classes: (1) tree (2) grass (3) bare (4) water (5) building (6) road (7) other paved surfaces and (8) shrub. August. https://open.toronto.ca/dataset/forest-and-land-cover/.

Conway, T.M., Khan, A., and Esak, N. 2020. "An analysis of green infrastructure in municipal policy: Divergent meaning and terminology in the Greater Toronto Area." *Land Use Policy*, Vol. 99: pp. 104864. December. ISSN 02648377. https://linkinghub.elsevier.com/retrieve/pii/S0264837720302064. doi:10.1016/j.landusepol.2020.104864.

Deng, L-J., Vivone, G., Paoletti, M.E., Scarpa, G., He, J., Zhang, Y., Chanussot, J., and Plaza, A. 2022. "Machine Learning in Pansharpening: A benchmark, from shallow to deep networks." *IEEE Geoscience and Remote Sensing Magazine*, Vol. 10 (No. 3): pp. 279–315. September. ISSN 2168-6831, 2473-2397, 2373-7468. https://ieeexplore.ieee.org/document/9844267/. doi:10.1109/MGRS.2022.3187652.

Dong, C., Loy, C.C., He, K., and Tang, X. 2015. Image super-resolution using deep convolutional networks. https://arxiv.org/abs/1501.00092. Version Number: 3.

Grill, J.-B., Strub, F., Altché, F., Tallec, C., Richemond, P.H., Buchatskaya, E., Doersch, C., *et al.* 2020. Bootstrap your own latent: A new approach to self-supervised learning. https://arxiv.org/abs/2006.07733. Version Number: 3.

He, K., Fan, H., Wu, Y., Xie, S., and Girshick, R. 2019. Momentum contrast for unsupervised visual representation learning. https://arxiv.org/abs/1911.05722. Version Number: 3.

Jean, N., Wang, S., Samar, A., Azzari, G., Lobell, D., and Ermon, S. 2018. Tile2Vec: Unsupervised representation learning for spatially distributed data. https://arxiv.org/abs/1805.02855. Version Number: 2.

Jung, H., Oh, Y., Jeong, S., Lee, C., and Jeon, T. 2022. "Contrastive self-supervised learning with smoothed representation for remote sensing." *IEEE Geoscience and Remote Sensing Letters*, Vol. 19 pp. 1–5. ISSN 1545-598X, 1558-0571. https://ieeexplore.ieee.org/document/9397864/. doi:10.1109/LGRS.2021.3069799.

Lin, D., Fu, K., Wang, Y., Xu, G., and Sun, X. 2017. "MARTA GANs: Unsupervised representation learning for remote sensing image classification." *IEEE Geoscience and Remote Sensing Letters*, Vol. 14 (No. 11): pp. 2092–2096. November. ISSN 1545-598X, 1558-0571. http://ieeexplore.ieee.org/document/8059820/. doi:10.1109/LGRS.2017.2752750.

Masi, G., Cozzolino, D., Verdoliva, L., and Scarpa, G. 2016. "Pansharpening by Convolutional Neural Networks." *Remote Sensing*, Vol. 8 (No. 7): pp. 594–2072. July. ISSN 4292. https://www.mdpi.com/2072-4292/8/7/594. doi:10.3390/rs8070594.

Maurer, T. 2013. "How to pan-sharpen images using the gram-schmidt pan-sharpen method – a recipe." *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, Vol. XL-1/W1: pp. 239–244. doi:10.5194/isprsarchives-XL-1-W1-239-2013.

McFeeters, S. K. 1996. "The use of the normalized difference water index (ndwi) in the delineation of open water features." *International Journal of Remote Sensing*, Vol. 17 (No. 7): pp. 1425–1432. doi:10.1080/01431169608948714.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. 2013. Efficient estimation of word representations in vector space. https://arxiv.org/abs/1301.3781. Version Number: 3.

Moore, C., and Hester, D. 2023. "A self-supervised approach to land cover segmentation." https://arxiv.org/abs/2310.18251. Version Number: 1.

Oktay O., and Schlemper. 2018. "Attention u-net: Learning where to look for the pancreas." *arXiv Preprint arXiv:1804.03999*, MIDL Conference. doi: 10.48550/arXiv.1804.03999.

Qin, R., and Liu, T. 2022. "A review of landcover classification with very-high resolution remotely sensed optical images—Analysis unit, model scalability and transferability." *Remote Sensing*, Vol. 14 (No. 3): pp. 646. January. ISSN 2072-4292. https://www.mdpi.com/2072-4292/14/3/646. doi:10.3390/rs14030646.

Rouse, J. W., Haas, R. H., Schell, J. A., and Deering, D. W. 1974. Monitoring vegetation systems in the great plains with erts. In *Third Earth Resources Technology Satellite-1 Symposium*, Vol. 1, pp. 309–317. Washington, DC: NASA SP-351.

Saboori, M., Homayouni, S., Shah-Hosseini, R., and Zhang, Y. 2022. "Optimum feature and classifier selection for accurate urban land use/cover mapping from very high resolution satellite imagery." *Remote Sensing*, Vol. 14 (No. 9): pp. 2097. April. ISSN 2072-4292. https://www.mdpi.com/2072-4292/14/9/2097. doi:10.3390/rs14092097.

Statistics Canada. 2021. Census Profile, 2021 Census of Population: Toronto, City (Census subdivision), Ontario. Available at: https://www12.statcan.gc.ca/census-recensement/2021/dp-pd/prof/details/page.cfm?Lang=E&SearchText=toronto&DGUIDlist=2021A000011124,2021A00053520005&GENDERlist=1,2,3&STATISTIClist=1,4&HEADERlist=0

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

van den Oord, A., Li, Y., and Vinyals, O. 2018. Representation learning with contrastive predictive coding. https://arxiv.org/abs/1807.03748. Version Number: 2.

Wald, L., Ranchin, T., and Mangolini, M. 1997. “Fusion of satellite images of different spatial resolutions: Assessing the quality of resulting images.” *Photogrammetric Engineering and Remote Sensing*, Vol. 63 pp. 691–699. November.

Walter, K., Gibson, M. J., and Sowmya, A. 2020. Self-supervised remote sensing image retrieval. In *IGARSS 2020 - 2020 IEEE International Geoscience and Remote Sensing Symposium*, 1683–1686, Waikoloa, HI, USA: IEEE. September. ISBN 978-1-72816-374-1. doi:10.1109/IGARSS39084.2020.9323294.https://ieeexplore.ieee.org/document/9323294/.

Wang, Y., Albrecht, C.M., Braham, N.A.A., Mou, L., and Zhu, X.X. December 2022. “Self-supervised learning in remote sensing: A review.” *IEEE Geoscience and Remote Sensing Magazine*, Vol. 10 (No. 4): pp. 213–247. ISSN 2168-6831, 2473-2397, 2373-7468. https://ieeexplore.ieee.org/document/9875399/. doi:10.1109/MGRS.2022.3198244.

Yaloveha, V., Podorozhniak, A., Kuchuk, H., and Garashchuk, N. 2023. “Performance comparison of cnns on high-resolution multispectral dataset applied to land cover classification problem.” *Radioelectronic and Computer Systems*, (Volume 0 No. 2): pp. 107–118. (doi:10.32620/reks.2023.2.09.

Zbontar, J., Jing, L., Misra, I., LeCun, Y., and Deny, S. 2021. Barlow twins: Self-supervised learning via redundancy reduction. https://arxiv.org/abs/2103.03230. Version Number: 3.