PARCEL-BASED SCENE CLASSIFICATION: DATASET, MODEL EVALUATION, AND IMPACT OF CHANGES

Suzanna Cuypers*, WuFan Zhao, and Maarten Vergauwen
Department of Civil Engineering, Geomatics Section, KU Leuven
9000 Ghent, Belgium
*Email: suzanna.cuypers@kuleuven.be

KEY WORDS: Scene classification, parcel-based land use, urban sprawl, deep learning, remote sensing

ABSTRACT: Land use in rural areas has a significant impact on its immediate surroundings, particularly in terms of land cover and sealed surfaces. This study proposes a novel approach that focuses on identifying non-conforming land use and quantifying its impacts on the relevant parcel and its direct environment. In this study, we built a parcel-based land use scene classification dataset focusing on rural areas in Flanders, Belgium, a region notable for having the highest degree of urban sprawl in Europe. To classify a parcel's land use, we trained and tested deep learning networks on our dataset. The trained CNNs exhibit overall accuracies ranging from 55.66% to 61.43%, outperforming a ViT model. We observed that pretraining significantly influences deep learning network performance, but pretraining on a remote sensing dataset does not necessarily yield a better model. We quantified the environmental impact of land use alterations by leveraging existing resources such as land use, imperviousness, and naturalness maps. Furthermore, we proposed a Human Land Cover Index (HLCI), a metric designed to gauge human influence via the land cover map.

1. INTRODUCTION

Increasing settlement area is a growing concern in Europe, particularly in Flanders, the northern region of Belgium. Currently, settlement areas account for approximately 33.4 % of the total land area in this region (Pisman et al. 2021). The expansion of settlement areas in Flanders grows continuously and is haphazardly distributed across the entire region (Buitelaar and Leinfelder 2020; Pisman et al. 2021). This growth is predominantly attributed to residential and garden development (Pisman et al. 2021). "Settlement area" refers to the land designated for human activities, which includes housing, industrial and commercial zones, infrastructure, as well as areas earmarked for recreational purposes such as parks and gardens. One significant driver behind this expansion is urban sprawl. According to a report from the European Environment Agency (EEA) that assesses urban sprawl, Belgium ranks second in the category of weighted urban proliferation—a measure of urban sprawl (European Environment Agency. and Swiss Federal Office for the Environment (FOEN). 2016). The European Commission has set out the goal of "no net land take" by 2050 (EU Soil Strategy), to prevent the increase in human settlement area. One of Flanders's efforts to meet the goals of the EU Soil Strategy and Europe's Green Deal is to focus on smart urban planning and limiting land take, which refers to the conversion of previously agricultural or natural areas into urban land with buildings and infrastructure. Thus, it is necessary to monitor land use changes in rural areas.

To further bolster endeavors to curb the expansion of settlement areas, it is imperative to undertake the measurement and mapping of human settlements and their direct environmental impact (Vogler and Vukomanovic 2021). By tracking detected land use changes, various corresponding environmental changes, such as alterations in land cover, sealed surfaces, and naturalness, can be monitored. Consequently, this study proposes a novel approach that centers on land use. We hypothesise that non-conforming land use (not following the zoning plan) in rural areas exerts a more substantial influence on their immediate surroundings. Hence, the primary goal of this study is to identify non-conforming land use in rural regions and quantify its implications on the relevant parcel and its direct environment, particularly in terms of land cover and sealed surfaces.

Measuring the impact of land use changes in agricultural zones is challenging since knowledge of actual land use is limited. Dataset of industrial and commercial companies in Flanders exists, but are not reliable in terms of time accuracy, economic activity, and location. Consequently, an accurate understanding of historical land use is lacking. This lack of knowledge hinders the precise quantification of the influence of land use changes on the surrounding environment. While time-series maps showcasing land cover, imperviousness, and naturalness are available, a time-series map illustrating the potential drivers behind these changes remains absent. To fill this gap, we propose a parcel-based land use classifier. In pursuit of this objective, we present a parcel-based land use dataset. This dataset is distinct in its emphasis on land use within agricultural and rural domains—a focus that sets it apart from conventional scene classification datasets. Our aim is to illustrate the intricate relationship between land use and land cover dynamics. Through this endeavor, we seek to highlight the causal chain linking land use decisions to subsequent changes in land cover.

In contrast to much of the existing research on land use and land cover analysis, our approach centers on scene classification rather than pixel-wise classification. In the context of low-resolution images, the size of objects tends to be similar to the pixel resolution. However, when dealing with Very High Resolution (VHR) images, attempting to discern land use at the pixel level becomes less practical(Cheng et al. 2020; Xia et al. 2017), often referred to as semantic image segmentation. Instead, groups of pixels collectively represent objects of interest. This leads to Geographic Object-Based Image Analysis (GEOBIA) as the logical progression. GEOBIA involves initially identifying objects by clustering pixels based on attributes such as color and texture. Subsequently, each object's class is determined by analyzing the average pixel value within that object.

Our previous work (Cuypers, Nascetti, and Vergauwen 2023) explores Land Use/Land Cover (LULC) mapping using GEOBIA and a combination of satellite and VHR aerial imagery. The objects were generated through an unsupervised clustering algorithm, which can be enhanced with deep learning methods. Notably, the VHR bands derived from aerial imagery were found to significantly influence class prediction accuracy. When it comes to land use management applications, a parcel-based analysis proves to be more fitting and relevant than a pixel or object-based approach (Bin et al. 2014). As a result, our present research focuses on harnessing the power of deep learning in conjunction with VHR imagery to enhance parcel interpretation and classification accuracy.

The primary contributions of this work are:

- We developed a novel framework for parcel-based land use classification in rural areas. This workflow incorporates automated dataset preparation and applying deep learning networks for classification.
- We built a parcel-based land use scene classification dataset focusing on rural areas in Flanders, a region notable for having the highest degree of urban sprawl in Europe. Our dataset is the first focusing on rural areas and comprises ten distinct classes which will help in human settlement monitoring.
- We trained state-of-the-art Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs) and investigated various training optimization strategies such as employing pretrained weights..
- Furthermore, we propose the Human Land Cover Index (HLCI), a metric designed to gauge human influence via the land cover map. We quantified the environmental ramifications of land use alterations by leveraging existing resources such as land use, imperviousness, and naturalness maps to pinpoint the land use changes that yield the most significant human impact.

The rest of this article is organized as follows. Section 2 describes the current state of the art on scene classification, the relevant datasets, as well as some works on measuring the impact of land use on the environment. Our dataset is extensively described in Section 3. The methods to establish scene classification and measure land use impact are explained in Section 4. The experiments, results, and discussion are listed in Section 5. We conclude with Section 6.

2. RELATED WORK

2.1 Scene Classification Datasets

Various scene classification dataset exists and are used in benchmarking scene classification methods, but land use, land cover, and object classes are mixed together in the available datasets. The UC Merced dataset (Y. Yang and Newsam 2010) contains 21 scene categories with 100 samples for each category, which include land cover classes as well as object classes. The images are 256 x 256 pixels and have a resolution of one foot (30.48 cm). NWPU-RESISC45 (Cheng, Han, and Lu 2017) has 45 scene classes with 700 samples per class. The images are 256 x 256 pixels with varying resolutions ranging from 30 to 0.2 m. AID (Xia et al. 2017) contains 30 classes with a total amount of 10,000 multi-source images and a varying amount of samples per category. The images are 600 x 600 pixels and have varying resolutions ranging from 8 to 0.5 m. The more recent MillionAID (Long et al. 2021) has similar classes to our dataset, namely 51 classes that can be grouped in 28 parents nodes which belong to eight base classes. These base classes are agriculture land, commercial land, industrial land, public service land, residential land, transportation land, unutilized land, and water area. Each of the 51 classes has 2000 to 45,000 sample images. However, the images have mixed resolutions ranging from 0.5 to 153 m. The image sizes also vary, ranging from 110 × 110 to 31,672 × 31,672 pixels. A parcel-based, very high resolution dataset of aerial imagery focusing solely on land use is missing. Therefore, we propose a new dataset to bridge this gap.

2.2 Scene Classification Methods

Recent developments on remote sensing scene classification methods focus on deep learning with Convolutional Neural Networks (CNNs). Attention-based models such as vision transformer have also been employed for scene classification.

CNN modules can be through neural network ensembles. Zhang et al. introduce a combination of CNNs for scene classification and test it on the UC Merced dataset and a manually annotated satellite image of Sydney with 1 m resolution.

They call their method Gradient Boosting Random Convolution Network (GBRCN). It incorporates two elements: first, a Gradient Boosting Machine, and, second, reusing the learned weights in each CNN for the base learner called the Random Convolutional Network (F. Zhang, Du, and Zhang 2016). Wang et al. incorporate a recurrent attention mechanism in a CNN module in order to reduce computing power and redundant data. They validate their new network on their developed dataset, OPTIMAL-31, and prove their model achieves state-of-the-art performance for remote sensing scene classification (Q. Wang et al. 2019). The dataset OPTIMAL-31 contains 1860 images with varying resolutions and 31 classes including object classes such as airplane and bridge. Contrary to Wang et al., Tong et al. focus on channel attention and create a much lighter network, named channel-attention-based DenseNet (CAD). They validate their network on the UC Merced, AID, and NWPU-RESISC45 datasets. Their proposed attention outperforms the traditional CNN models: ResNet50 and VGG16 (Tong et al. 2020).

Scene classification can also be an initial step in building detection. The authors of (Sun, Tang, and Zhang 2017) develop a two-step method using CNNs to first detect villages in a rural environment and then, for the detected villages, detect buildings by classifying smaller images with a resolution of 0.27 m. The authors conclude that CNNs offer an efficient method and satisfactory results for building detection in rural areas with complex backgrounds. However, the usage of the building or building type is not distinguished.

Other methods to enhance CNN training is through data augmentation as proposed by Yang et al. (N. Yang et al. 2018). The authors propose a channel-dropping training data augmentation strategy in which one of the image's input bands is removed during training. The benefit of this data augmentation technique is demonstrated on datasets with four input channels (RGB and NIR) and the UC Merced dataset with 3 input channels (RGB). Though performance with the 4-channel datasets increased, the performance of the method on the 3-channel dataset severely decreases. However, it still performed better than above-mentioned GBRCN (F. Zhang, Du, and Zhang 2016). Moreover, they find that a fine-tuning method following (Nogueira, Penatti, and Dos Santos 2017) leads to much better results.

Vision transformers have been applied to the task of remote sensing scene classification. Lv et al. (Lv et al. 2022) evaluate ViT models on the UC Merced and the NWPU-RESISC45 dataset and compare to CNN models. They found that the channel-attention-based CNN (CAD) by Tong et al. (Tong et al. 2020) showed better performance than the classical CNNs: VGG16 and ResNet50. ViT-based methods performed even better, especially when pretraining is performed. CTNet, a combination of a pretrained CNN and a pretrained ViT module, proposed by Deng et al. (Deng, Xu, and Huang 2022), outperforms all above mentioned methods on the AID and NWPU-RESISC45 datasets. It has the ability to both extract semantic features through the ViT module and structural information through the CNN branch. The authors apply their framework using ResNet34 and MobileNet v2. In this work, we validate our dataset on classical CNNs and compare the performance against a ViT.

2.3 Parcel-Based Land Use Classification

A combination of data sources can help in classifying land use. The authors of (W. Zhang et al. 2017) utilize VHR aerial imagery, Lidar, Google Street View (GSV) images, and GIS data to obtain thirteen parcel features and train a random forest classifier. The features include parcel size, number of buildings, maximum building levels, Normalized Difference Vegetation Index (NDVI), and length of detected text in the GSV images. All the used features are derived from the above-mentioned input data sources: building detection and story height estimation is approximated from the Lidar data, the NDVI is obtained through the near infra-red and red bands of the VHR aerial image, and the length of text in the GSV images requires a text detection and recognition module. The method achieves an overall accuracy of 77.5% within in the study area of New York. However, its applicability to different regions is limited due to the requirement of specific input data, which is not available universally. Moreover, the time-sensitive nature of the data further restricts its application, as certain data sources are generated intermittently.

2.4 Land Use Impact

The significance of land use classification is shown in the literature through metrics of human footprint and naturalness. Vogler and Vukomanovic examine the effects of urbanization and population growth by quantifying the human footprint. The authors generate metrics for and quantify changes in housing density, imperviousness, per capita land consumption, and land use efficiency (Vogler and Vukomanovic 2021). Recent work by Ekim et al. proposes a Naturalness Index (NI) to quantify and visualize human influence on the environment. To create this index, the authors consider four influencing factors: population density, land cover, accessibility, and electrical power infrastructure (Ekim et al. 2021). In our work, we also use land cover and assume the same weights as Ekim et al. to create our Human Land Cover Index (HLCI).

3. DATASET CREATION

For the task of parcel-based land use classification in non-urban zones in Flanders, Belgium, we create a large dataset with VHR aerial images and its associated land use labels. The dataset includes ten classes: daytime recreation, retail, community, hospitality (food & beverage), industry, office, agricultural land, agricultural infrastructure, leisure accommodation, and residential. The aerial images are gathered from 23 municipalities between 2018 and 2022. An overview of the image samples per class can be found in Table 1. The images are automatically labeled using available GIS datasets described below. We refer to this dataset in the rest of the text as *D*.

The proposed dataset is suitable for the task of parcel-based scene classification because

- it focuses purely on land use
- it is parcel based, and can thus be utilized to monitor individual activities
- it is a real dataset, i.e. not manually created. A model that performs well on this dataset, will perform well in real applications.

Table 1 The classes present in dataset D and their sample counts

Class	# samples
0: Daytime recreation	428
1: Retail	399
2: Community	369
3: Hospitality (food & beverage)	187
4: Industry	1,089
5: Office	1,267
6: Agricultural land	30,052
7: Agricultural infrastructure	6,959
8: Leisure accommodation	109
9: Residential	6,266

The sizes of the images exhibit variation, with a mean width of 505 pixels. Specifically, the tenth percentile corresponds to 119 pixels, while the ninetieth percentile reaches 941 pixels. The values of image heights are comparable to image width.

Although the dataset is parcel-based, we choose not to mask the areas around the parcels for several reasons. First, the shape of the parcels can give away the function, especially for agricultural land which often has an irregular shape. However, since we specifically want to detect illegal land use that resulted from a change from agricultural use to a non-conforming land use, we do not want shape to be a deciding factor to land use class.

3.1 Source Data Collection

To generate our dataset *D*, we wrote various pyQGIS scripts that process a list of input data sources, shown in Table 2. In this section, we expand on the input data sources. In the following section, we show an overview of the GIS Python scripts used to automatically match each orthomosaic to a label and the output labels each of them generate.

Imagery - Aerial orthomosaics of Flanders are typically created twice a year: in the winter and in the summer. Available orthomosaics for Flanders include low (1 m) resolution, medium (25 cm and 15 cm since 2022) resolution, and high (10 cm) resolution. Only one high resolution image is available, and it spans the years 2013-2015. In order to facilitate model training, we opt not to use high resolution imagery but use only medium resolution imagery with a resolution of 25 cm and 15 cm in 2022. Furthermore, we decide to use winter images exclusively to reduce foliage obstructing visibility in the imagery. We limit the dataset to imagery no older than 2018. Although imagery from before then is available, there is limited label source data available and the reliability of that data deteriorates.

Dual-sensor models are explored in research for scene classification, such as RGB-D scene classification and segmentation (Jin et al. 2022). However, for the study area, only two Lidar-based digital surface models (DSM) (including building heights) were captured over the time spans 2001-2004 and 2013-2015. To ensure the usability of the trained model, we will only train on RGB imagery since this data is available for each year.

Labels - Acquiring information about the functional use of a parcel is not evident, which is the main motivation for creating a deep learning model that can identify the functional use from an aerial image. The labels are extracted from five sources: Business directory by Leiedal, Sport infrastructure, Agricultural parcels, Points of Interest (POI), and Registered companies (Dutch *: VKBO*). Evident from this list, a public dataset with companies already exists (*VKBO*). However, it is highly unreliable for two reasons. First, the location is not the location of the business' activity, but the

headquarters. For small business, this address is often the owner's residence. Second, the actual business' category cannot be extracted from this dataset. Businesses in Flanders list all their economic activities when they register, but since any change to this list costs money and there is no limit to the number of activities that can be listed, companies list numerous activities making it impossible to put them in only one of the categories required for this work. Thus, extracting location and economic activity from this dataset is highly unreliable. Therefore, we integrate other data sources and limit the usage of the *VKBO* to build our dataset. One of these is the sports infrastructure dataset which we directly obtained from the Flemish government. It localizes sport infrastructures by coordinates and classifies them by type (such as tennis fields, gym halls, fishing ponds, etc.). This dataset was created in 2018 and frequently gets updated, so that it is reliable in the time dimension as well as location.

The data source Business directory by Leiedal is a private, manually annotated dataset in possession of inter-communal Leiedal. It localizes companies in 13 municipalities and lists the three most dominant business functions. Further, we enrich our dataset with economic activities from a public POI (Points Of Interest) Web Feature Service (WFS) and publicly available shapefiles of Agricultural parcels. The latter is available since 2008 with yearly new versions.

Supporting data - Scene classification requires to capture the surrounding area around the parcels as well. We use the administrative parcels to generate a rectangular area that encompasses the entire parcel and parts of the surrounding area. We decided not to mask the area outside the parcel for two reasons. First, this would reveal the shape of the parcel which can be influenced by the land use of the parcel. For example, agricultural land is more likely to have irregular shapes. Since the goal is to detect non-conforming land use changes, the shape of the parcel should not be used by the trained model as a factor. Second, we hypothesise that the surroundings of a parcel are influenced by the economic activity on a parcel. For example, a growing industry will require adapted road infrastructure. Besides parcels, we also utilize address points, municipality borders, and a zoning plan. The zoning plan enables us to only extract imagery from parcels located in agricultural and natural areas. Table 2 lists all used data sources used in this work and their relevant information.

Table 2 Data sources used for the generation of training samples and for the effect measurement of land use changes.

Dataset name (*Dutch*)	Resolution	Available years	Format
Winter aerial orthomosaic	25 cm & 15 cm (2022)	2000-2003, 2005-2007, 2008-2011, 2012, 2013,..., 2022	.tif
Business directory by Leiedal (*Bedrijvengids*)	13 municipalities	2018 - 2023	.csv
Sports infrastructure (*Sportinfrastructuur*)		2018 - 2023	.csv
Agricultural parcels (*Landbouwgebruikspercelen*)		2008 - 2023	.shp
Points Of Interest (POI)		2023	WFS
Registered companies (*VKBO*)		2023	.shp
Adresses (*CRAB*)		2023	.shp
Adminstrative parcels (*Administratieve percelen GRB*)		2014 – 2023	.shp
Zoning plan (*Gewesplan*)		2000	.shp
Municipalities (*Gemeenten*)		2023	.shp
Greenmap (*Groenkaart*)	1 m	2009, 2012, 2015, 2018, 2021	.tif
Soil sealing map (*BAK - Bodem Afdekkingskaart*)	5 m	2012, 2015, 2018	.tif

3.2 Scene and Label Extraction

With the given input data listed in Table 2, we set up rule-based pyQGIS processing scripts to extract image patches with its associated land use function. The scripts and the class outputs they generate are visualized in Figure 1Figure 1 The data extraction pyQGIS processing scripts with their data input and class outputs. Of each class three samples are shown.. We employ High Throughput Computing with Condor (HTC) to divide the "GDAL:cliprasterbyextent" processes over multiple cores. The GDAL clipping script only run on CPU and takes up most of the running time. This way, we can divide the jobs over multiple cores and generating the dataset is sped up significantly.

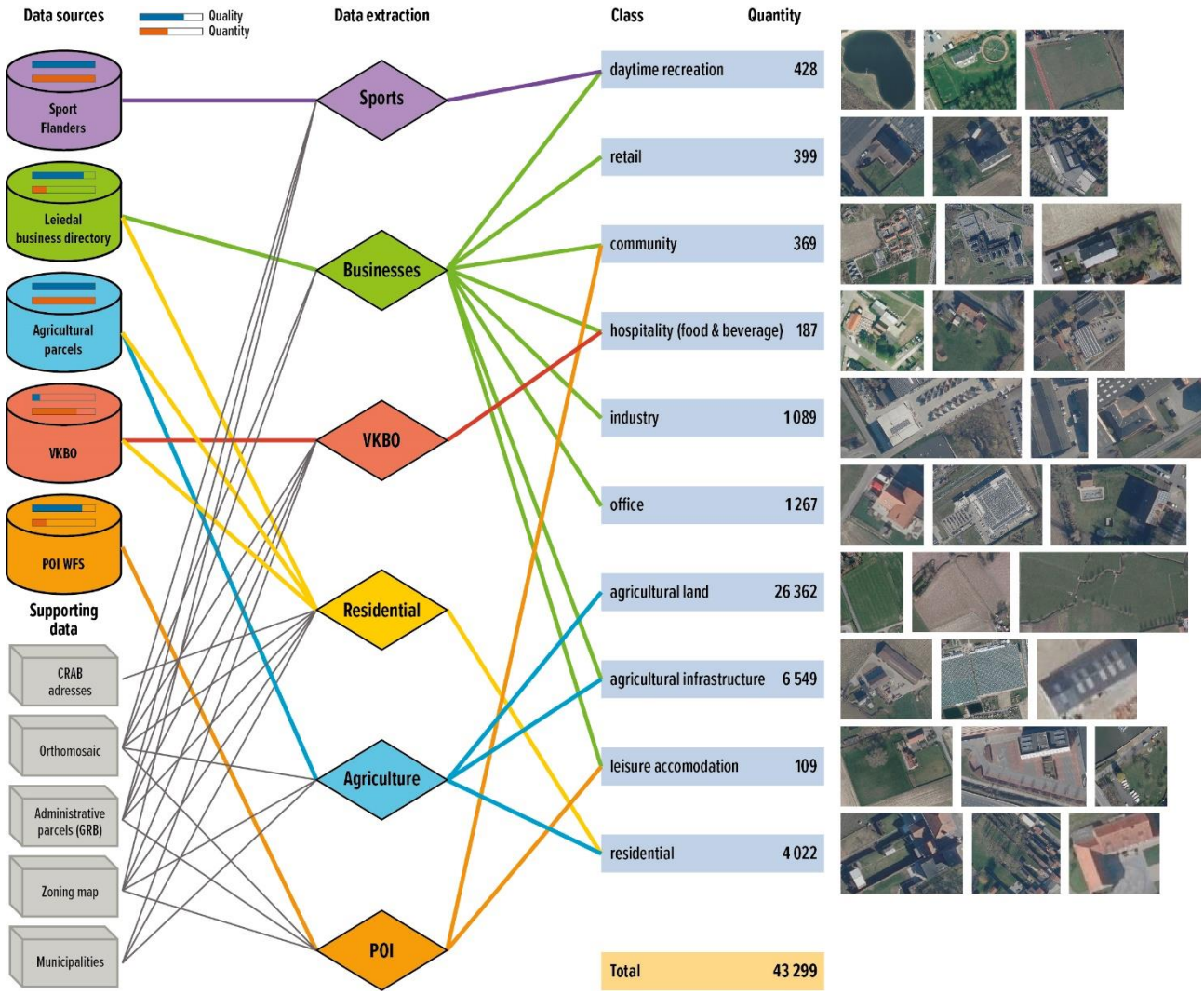

Figure 1 The data extraction pyQGIS processing scripts with their data input and class outputs. Of each class three samples are shown.

3.3 Data Splits

To reduce class imbalance, we create a subdataset, called D_{2000}, where the samples are limited to 2000 per class. The class distribution of this dataset is shown in Figure 2. Due to the large class imbalance, we carefully split the data in train, validation, and test set so that each class is optimally represented in each split. Furthermore, we split based on the municipality where the image is captured, so that a single municipality is only present in one split. As a result, images of the same parcel available over multiple years are in the same data split. Secondly, we make sure that the most underrepresented class (leisure accommodation) is split in such a way that the largest split is the train set. Overall we aim for the largest split to be the train set followed by the test set.

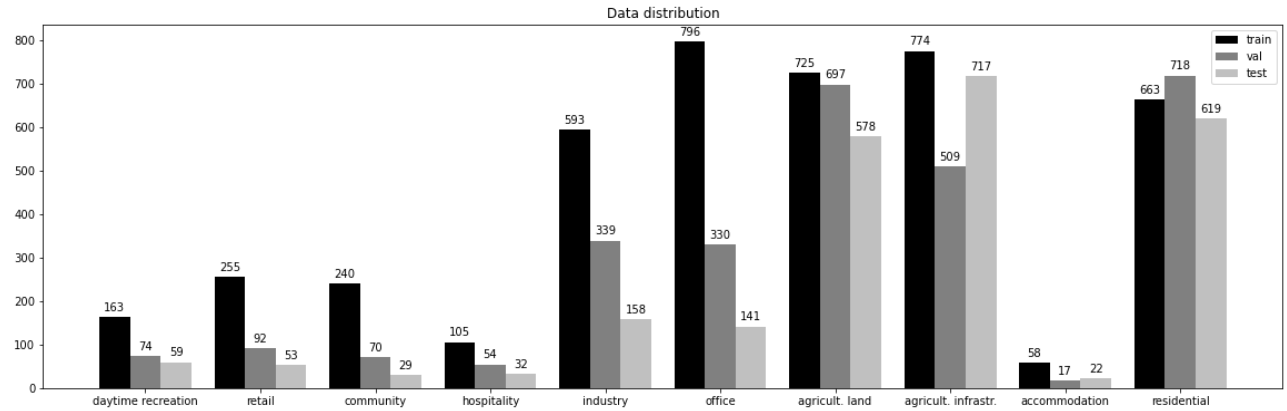

Figure 2 Class distribution in the balanced D2000 dataset.

4. METHODS

4.1 CNNs and ViTs

Large CNN models pretrained on natural image datasets such as ImageNet perform well on benchmarking datasets because they are pretrained on very large datasets. For remote sensing scene classification, imagery is widely available but labels are much harder to obtain. Therefore, transfer learning is the most common approach in remote sensing scene classification. The CNN is first trained on a large natural image dataset and then fine-tuned on the target dataset. In this work, we employ pretrained CNNs and freeze all layers except the last layers to then fine-tune the model on our own dataset D_{2000}.

The drawback of fine-tuning is that training the final layers alone does not suffice to teach the model to understand inter-class similarity and intra-class diversity that is typical of remote sensing datasets (D. Wang, Zhang, Du, et al. 2023). Since labels are mostly absent but a plethora of remote sensing imagery exists in the field of remote sensing, researchers are investigating other learning approaches such as unsupervised and self-supervised learning. Momentum Contrast (MoCo) (He et al. 2020) is a self-supervised learning technique where a network is pushed to improve image representations with a contrastive loss. The loss function pushes positive pairs (an image and its augmented version) closer together and pushes negative pairs (the input image and images from a different classes) further apart. Ayush et al. build further upon the MoCo training technique and consider remote sensing images of the same location taken at different times as positive pairs (Ayush et al. 2022). Besides using these temporal positive pairs, the authors combine the contrastive loss learning with geo-location classification, where the classes are the location where the images is taken. The network can deduce the location from the context of the image (for example, an image of an Indian elephant is most likely taken in India). Therefore, their framework includes contrastive learning, where the loss is computed based on the features from the input image and images of the same location taken at a different time (temporal positive pairs) - which can be seen as the augmented images in the MoCo framework - and geography aware classification. The resulting trained network is intended as a pretrained encoder and can further be fine-tuned on a remote sensing task, such as object detection and semantic segmentation. There is however, recently, a new large-scale remote sensing dataset that is sufficiently large for training a network from scratch in a supervised manner, namely MillionAID (Long et al. 2021).

In this work, we initialize a ResNet-50 model with weights obtained from pretraining in a supervised manner on ImageNet-1k and MillionAID and from pretraining in a self-supervised manner on imageNet (MoCo) and on a remote sensing dataset considering temporal pairs and location (MoCo-v2-Geo).

Besides CNNs, vision transformers are also becoming more popular in the field of scene classification. The Vision Transformer (ViT) was proposed by (Dosovitskiy et al. 2021) in 2021, and has since inspired researchers in the field of remote sensing (D. Wang, Zhang, Xu, et al. 2023; Deng, Xu, and Huang 2022; D. Wang, Zhang, Du, et al. 2023), In a ViT, an image is divided into 16 patches, and each is fed as an input to the model. This limits the model's ability to learn global features (J. Zhang, Zhao, and Li 2021). Although the ViT does not outperform CNNs of similar size trained on mid-sized datasets, it becomes interesting when trained on large-scale datasets such as ImageNet-21k. The authors of (Dosovitskiy et al. 2021) show that fine-tuning the pretrained ViT model on domain specific datasets reaches state-of-the-art performance. In this work, we fine-tune a ViT-base (ViT-B) model on our dataset.

4.2 Fine tuning

We train the model on the train set, and save the best model based on its performance on the validation set. To test the final model's performance, we evaluate on the test set.

We choose to fine-tune our models instead of training from scratch as it has been proven that the performance is better in the field of scene classification (Nogueira, Penatti, and Dos Santos 2017; Pires De Lima and Marfurt 2019). Models pre-trained on large natural image datasets and fine-tuned on small remote sensing datasets outperform models directly trained on the smaller remote sensing datasets. Even for ViTs, transfer learning outperforms training from scratch (Steiner et al. 2022). This can be attributed to the generality of the initial layers of the network across all datasets; these layers are capable of detecting fundamental elements like lines, shapes, and blobs. The domain-specific datasets are generally not large enough to train a network to detect these visuals.

Fine-tuning a pre-trained network starts by reusing the weights of the net trained on a very large dataset from a different domain (such as ImageNet) and then continuing training on the target dataset with a lower learning rate. There are two fine-tuning options: continuing training all layers or freezing the first layers and only adjusting the last layers. We use three different initialization weights from pretrained ResNet50 on Imagenet-1k, MoCo-v2, and MillionAID. For the ViT base model (Vit-B), we use the weights from pretraining on ImageNet-21k. The exact implementation details used in this work are described in Section 5.

4.3 Human Land Cover Index

Land use has an impact on the land cover of its direct environment. A change in land use will influence the land cover of the parcel, but also the area directly around the parcel. To quantify the influence of land use changes on the direct environment, we create a Human Land Cover Index (HLCI) derived from the Flemish Greenmap (Dutch: *Groenkaart*) by weighting land cover class's counts over a parcel and its direct surroundings. The classes in the Greenmap are: high green, low green, agriculture, and non-green. The weights for the respective classes are: 0, 2, 8, and 10. Consequently, the HLCI ranges from 0 to 10 with 0 being the most natural and 10 having the most human influence. Our HLCI is adapted from (Ekim et al. 2021) where the same land cover weights are used.

Per parcel, the number of occurrences of each class are added resulting in a pixel count. Then, the class counts are multiplied by a weighting factor. The final weighted counts are added and divided by the total number of class counts in the parcel. The result is a value between 0 (completely covered by "high green") and 10 (no green present). The HLCI gives an indication of the intensity of land usage by humans.

Let v be the list of land cover occurrences of each class and w be the list of corresponding land cover weights. Then the weighted land cover values v' can be obtained through element-wise multiplication:

$$v' = v_i * w_i \; for \; i = 1,2,3,4 \tag{1}$$

The HLCI can be calculated as the sum of the weighted values v' divided by the sum of all values v:

$$HLCI = \frac{\sum_{i=1}^{4} {v'}_i}{\sum_{i=1}^{4} v_i} \tag{2}$$

5. EXPERIMENTS

5.1 Metrics

We evaluate the trained models based on their performance on the test set. We consider the Overall Accuracy (OA) which is the sum of all correct classification divided by the total number of classifications. Since OA is heavily influenced by class size, we include the Average Accuracy (AA), which is the average of the individual class accuracies so that the model is evaluated over all the classes equally. Additionally, Cohen's Kappa Coefficient considers agreement between predicted and true class while correcting for agreement that occurs by chance. It is thus a good metric to evaluate the model performance on a specific dataset considering the number of classes.

5.2 Implementation Details

Both Convolutional Neural Networks (CNNs) and Vision Transformer (ViT) models are set up using the PyTorch framework. To fine-tune ResNet50, we freeze all layers except for layer 4 and the fully connected (fc) layer. We use a weighted cross entropy loss function, batch size of 32 and Stochastic Gradient Descent (SGD). We achieve stable models at about 10 epochs. From our experiments, we found that increasing the number of epochs to 20 did not increase the performance. The learning rate is adjusted with a cosine function; the base learning rate is 0.03 and the final rate is 1e-5 with a weight decay of 1e-4. The resolution of the input images is 224 x 224. The ViT model is loaded from Hugging Face (backbone: google/vit-base-patch16-384), pretrained on ImageNet-21k and fine-tuned on ImageNet 2012 at a resolution of 384 x 384. We further fine-tune it on our dataset with a batch size of 8 for 10 epochs. We set the learning rate at 1e-5.

We trained each model five times and list the averages of the overall accuracy (OA), average accuracy (AA), and Cohen's Kappa Coefficient. We found that each model training was very consistent and showed little variation.

5.3 Classification Results

In Table 3, we report the results of the train classification networks on our balanced dataset D_{2000}. The best AA is achieved with ResNet50 pretrained on ImageNet-1k (40.86 %), followed closely by ResNet50 pretrained on MillionAID (40.67 %). Although a higher performance was expected on the model pretrained on the remote sensing dataset MillionAID, it could not outperform the model pretrained on the natural images. This can be explained by the mixed and

widely ranging resolutions in the images of MillionAID, whereas the resolution of the images in our dataset are 25 or 15 cm, which is much higher and a small range compared to MillionAID. Based on the OA, ResNet18 achieves the best performance (61.43 %). This metric is much higher than the AA because it is heavily influenced by the classification results on the largest data classes, i.e. agricultural land and agricultural infrastructure. A Kappa value between 0.41 and 0.60 indicates moderate agreement of the models.

The MoCo pretrained models perform lower than the fully supervised pretrained models, which is expected since these models were pretrained in an unsupervised fashion. However, the pretrained models on geo-specific data do not achieve significantly better OA. MoCo-TP is pretrained on temporal positive pairs of remote sensing images. Thus, we expect it to understand remote sensing imagery better. When considering the OA, the geo MoCos - geo, geo+TP, and TP - outperform regular MoCo, with MoCo-geo achieving the highest OA of 61.02 %. The authors of (Ayush et al. 2022) find that for the task of classification on the training dataset, functional Map of the World, MoCo-TP achieves the highest OA (68.32 %) outperforming MoCo-geo by four percentage points. However, when fine-tuning the whole model on the dataset, the model's performance increases to 71.55 %, which might indicate that the model was not pretrained enough.

Transferring the pretrained geo MoCos, the authors of (Ayush et al. 2022) found that the MoCo-geo weights gave the best fine-tuned results on a land cover dataset, outperforming the net initialized with ImageNet weights. This is also what we see in the OA of our results.

Although unsupervised pretrained networks do not outperform fully supervised pretrained networks yet for the task of down-stream remote sensing scene classification, domain-specific unsupervised pretraining tasks do bridge the gap between unsupervised pretraining and fully supervised pretraining. Our results back up the findings of (Ayush et al. 2022).

Table 3 Fine-tuning results on balanced dataset D_{2000}.

Model	Weights	OA	AA	Kappa
ResNet18	ImageNet-1k	**61.43**	39.75	0.5259
ResNet50	ImageNet-1k	61.32	**40.86**	**0.5262**
ResNet50	MillionAID (Long et al. 2021)	61.22	40.75	0.5256
ResNet50	MoCo (He et al. 2020)	55.66	37.96	0.4662
ResNet50	MoCo-geo (Ayush et al. 2022)	61.02	39.85	0.5222
ResNet50	MoCo-geo + TP (Ayush et al. 2022)	59.89	38.44	0.5084
ResNet50	MoCo-TP (Ayush et al. 2022)	58.24	38.24	0.4932
ViT-B	ImageNet-21k	49.37	34.14	0.4361

Class accuracies on D_{2000} - Since our dataset suffers class imbalance, we list the classification accuracies per class of the models with the highest AA in Table 4. The individual accuracies in this table are thus not the result of averages of several runs but the result of one single model. As expected, the classes with the highest presence in the dataset achieve the highest accuracies, such as agricultural land and agricultural infrastructure. The class residential is the third largest class but only achieves a maximum AA of 37.56 %. This can be explained by the fact that many residential lots also serve another function such as community and office. Therefore, it is more difficult to distinguish these classes from one another. The classes that are most difficult for the models to predict right are retail, community, industry, and leisure accommodation. Agricultural land is the most easily identified since it is the only class that seldom contains buildings.

Table 4 Class accuracies and Average Accuracy

Class	ResNet18 ImageNet-1k	ResNet50 ImageNet-1k	ResNet50 MillionAID	ResNet50 MoCo	ResNet50 MoCo_geo	ResNet50 MoCo_geo+TP	ResNet50 MoCo_TP
0: Daytime recreation	72.88	67.80	71.19	71.19	72.88	66.10	67.8
1: Retail	15.09	28.30	20.75	5.66	18.87	15.09	3.77
2: Community	13.79	6.90	3.45	10.34	17.24	3.45	13.79
3: Hospitality (food & beverage)	28.12	25.00	40.62	40.62	25.00	31.25	31.25
4: Industry	13.92	13.92	11.39	27.22	7.59	8.86	5.70
5: Office	59.57	51.06	42.55	48.94	48.23	46.10	55.32
6: Agricultural land	93.50	93.50	92.68	90.24	92.20	92.68	92.85
7: Agricultural infrastructure	81.54	83.33	78.24	66.53	80.17	80.72	77.55
8: Leisure accommodation	0	18.18	13.64	9.09	18.18	27.27	13.64
9: Residential	32.17	32.01	37.56	26.15	31.85	31.38	29.48
Average Accuracy	41.06	42.00	41.21	39.60	41.22	40.29	39.11

Figure 1**Error! Reference source not found.** shows the confusion matrix of a trained ResNet50 model. The results are normalized over the true classes. From the confusion matrix, it is clear that class 6 (agricultural land) is the easiest for the model to classify, followed by class 0 (daytime recreation) and 7 (agricultural infrastructure).

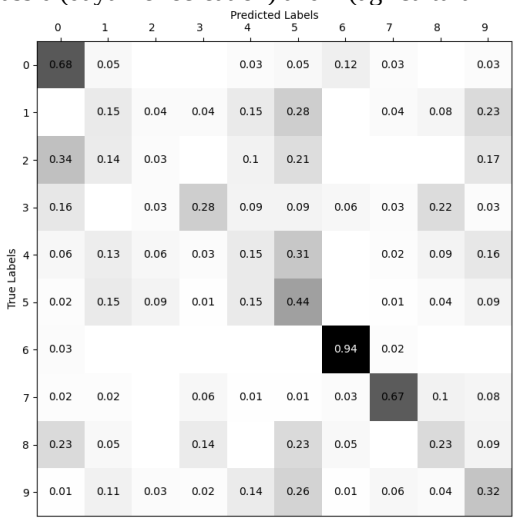

Figure 3 The normalized confusion matrix of a ResNet50 model pretrained on ImageNet-1k.

6. DISCUSSION

The impact of land use changes in rural areas is significant. The importance of investigating land use in rural areas has been shown by other researchers. Vogler and Vukomanovic found in their study on the US that mean per-capita land consumption was 8.5 times higher than in exurban, suburban, and urban density lands (Vogler and Vukomanovic 2021). Zeng and Ramaswami also came to a similar conclusion: the direct impact of land use in low-density rural lands is ten times greater than in urban areas due to larger homes and parcels, and about five times more road surface per person is required in rural areas (Zeng and Ramaswami 2020).

In Flanders, this issue is also pressing. The Spatial Report in Flanders (Pisman et al. 2021) indicates that the increase in sealed surface is higher than the population growth in Flanders. In fact, every new resident requires four times more built-up surface compared to 100 years ago. It is reported that 10-20 % of land use changed in Flanders between 2013-2019.

In future work, we will investigate the HLCI changes over time and connect them with the changes in land use per parcel. Specifically, we will look at parcels that were intended for agricultural usage and have recently undergone functional change that is not conforming with the zonal plan. Investigating these changes, we hope to pinpoint which economic activities lead to the largest impact on its direct environment in terms of human land cover and imperviousness. An example of a case study is shown in Figure 4.

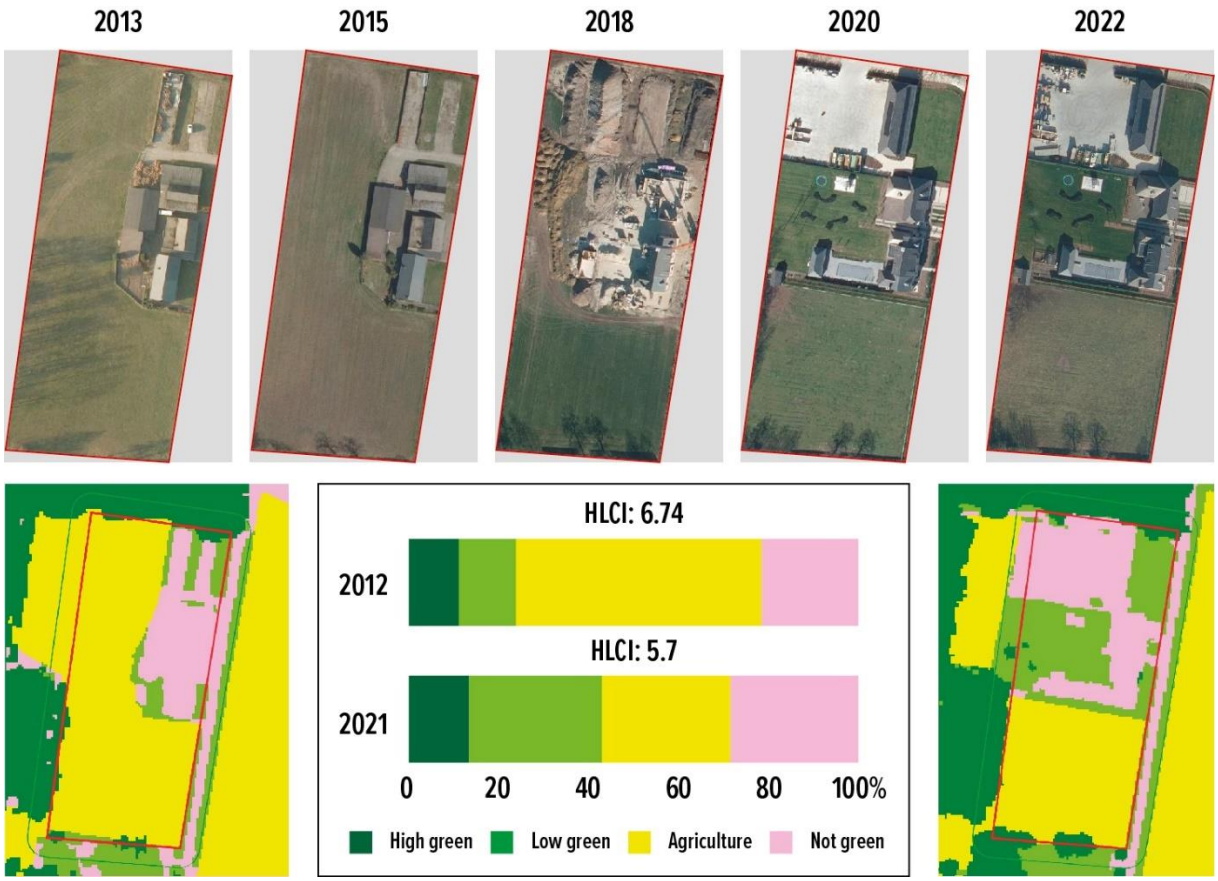

Figure 4 Case study of land cover for a parcel. The HLCI in 2012 is 6.74 and goes down to 5.7 in 2021 indicating that the land is less intensively used by humans. This is due to the transformation of agricultural land (yellow) to high and low vegetation (dark and light green)

7. CONCLUSION

We initiated this paper by highlighting the necessity of aerial-based land use classification. Especially in Flanders, Belgium, there is a strong need to monitor the expansion of settlement areas. We identified a significant lack of reliable data regarding land use. Consequently, we propose the training of deep learning classifiers, such as CNNs and ViTs, capable of classifying a parcels' land use based on a single aerial image.

To train these classifiers, we introduce a unique dataset. Given the undesirable the expansion of settlement area in rural regions, our dataset focuses on tracking land use changes in these areas. The dataset is developed using various data sources in a heuristic manner, so that training samples and the class labels are generated automatically. This approach yields several advantages, including rapid dataset creation, adaptability to any municipality with an available aerial orthomosaic, and scalability to incorporate new imagery. Our dataset is built on imagery from 13 Flemish municipalities, captured between 2018 and 2022. Each image corresponds to an administrative parcel, with its surrounding areas left unmasked and assigned one out of ten class labels. We emphasize the relevance of retaining the surrounding areas of the parcel.

The trained CNNs achieve overall accuracies ranging from 55.66 % to 61.43 % and average accuracies of 37.96 to 40.86 %, with the smaller ResNet-18 model outperforming the others. The ViT model performs notably lower, achieving an overall accuracy of 49.37 % and average accuracy of 34.14%. The average accuracies are lower than the overall accuracies as is expected with severe class imbalance. Some classes pose greater challenges for the models even though they are not among the least represented. For example, the residential class is difficult for the models, which can be attributed to the fact that many samples in other classes also serve the residential function. We conclude that pretraining is incredibly important for a deep learning network, but pretraining on a remote sensing dataset does not necessarily lead to a better performing model. Matching imagery resolution is crucial for knowledge transfer to the target dataset.

To quantify the human impact of the use of a parcel on the parcel and its surroundings, we proposed a Human Land Cover Index (HLCI). In future work, we will compare the HLCI before and after a change from agricultural land use to confirm our hypothesis that certain non-conforming land uses lead to more sealed surfaces and a higher human land cover index than other ‘less harmful’ usages.

8. REFERENCES

Ayush, Kumar, Burak Uzkent, Chenlin Meng, Kumar Tanmay, Marshall Burke, David Lobell, and Stefano Ermon. 2022. "Geography-Aware Self-Supervised Learning." arXiv. http://arxiv.org/abs/2011.09980.

Bin, Wu, Yang Jian, Zhao Zhongming, Meng Yu, Yue Anzhi, Chen Jingbo, He Dongxu, Liu Xingchun, and Liu Shunxi. 2014. "Parcel-Based Change Detection in Land-Use Maps by Adopting the Holistic Feature." *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 7 (8): 3482–90. https://doi.org/10.1109/JSTARS.2013.2291773.

Buitelaar, Edwin, and Hans Leinfelder. 2020. "Public Design of Urban Sprawl: Governments and the Extension of the Urban Fabric in Flanders and the Netherlands." *Urban Planning* 5 (1): 46–57. https://doi.org/10.17645/up.v5i1.2669.

Cheng, Gong, Junwei Han, and Xiaoqiang Lu. 2017. "Remote Sensing Image Scene Classification: Benchmark and State of the Art." *Proceedings of the IEEE* 105 (10): 1865–83. https://doi.org/10.1109/JPROC.2017.2675998.

Cheng, Gong, Xingxing Xie, Junwei Han, Lei Guo, and Gui-Song Xia. 2020. "Remote Sensing Image Scene Classification Meets Deep Learning: Challenges, Methods, Benchmarks, and Opportunities." *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 13: 3735–56. https://doi.org/10.1109/JSTARS.2020.3005403.

Cuypers, Suzanna, Andrea Nascetti, and Maarten Vergauwen. 2023. "Land Use and Land Cover Mapping with VHR and Multi-Temporal Sentinel-2 Imagery." *Remote Sensing* 15 (10): 2501. https://doi.org/10.3390/rs15102501.

Deng, Peifang, Kejie Xu, and Hong Huang. 2022. "When CNNs Meet Vision Transformer: A Joint Framework for Remote Sensing Scene Classification." *IEEE Geoscience and Remote Sensing Letters* 19: 1–5. https://doi.org/10.1109/LGRS.2021.3109061.

Dosovitskiy, Alexey, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, et al. 2021. "An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale." In *Proceedings of the ICLR 2021*. Virtual Event: arXiv. https://doi.org/10.48550/arXiv.2010.11929.

Ekim, Burak, Zeyu Dong, Dmitry Rashkovetsky, and Michael Schmitt. 2021. "The Naturalness Index for the Identification of Natural Areas on Regional Scale." *International Journal of Applied Earth Observation and Geoinformation* 105 (December): 102622. https://doi.org/10.1016/j.jag.2021.102622.

European Environment Agency. and Swiss Federal Office for the Environment (FOEN). 2016. *Urban Sprawl in Europe: Joint EEA FOEN Report.* LU: Publications Office. https://data.europa.eu/doi/10.2800/143470.

He, Kaiming, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. "Momentum Contrast for Unsupervised Visual Representation Learning." arXiv. http://arxiv.org/abs/1911.05722.

Jin, Jianhui, Wujie Zhou, Lv Ye, Jingsheng Lei, Lu Yu, Xiaohong Qian, and Ting Luo. 2022. "DASFNet: Dense-Attention–Similarity-Fusion Network for Scene Classification of Dual-Modal Remote-Sensing Images." *International Journal of Applied Earth Observation and Geoinformation* 115 (December): 103087. https://doi.org/10.1016/j.jag.2022.103087.

Long, Yang, Gui-Song Xia, Shengyang Li, Wen Yang, Michael Ying Yang, Xiao Xiang Zhu, Liangpei Zhang, and Deren Li. 2021. "On Creating Benchmark Dataset for Aerial Image Interpretation: Reviews, Guidances and Million-AID." arXiv. https://doi.org/10.1109/JSTARS.2021.3070368.

Lv, Pengyuan, Wenjun Wu, Yanfei Zhong, and Liangpei Zhang. 2022. "Review of Vision Transformer Models for Remote Sensing Image Scene Classification." In *IGARSS 2022 - 2022 IEEE International Geoscience and Remote Sensing Symposium*, 2231–34. Kuala Lumpur, Malaysia: IEEE. https://doi.org/10.1109/IGARSS46834.2022.9883054.

Nogueira, Keiller, Otávio A.B. Penatti, and Jefersson A. Dos Santos. 2017. "Towards Better Exploiting Convolutional Neural Networks for Remote Sensing Scene Classification." *Pattern Recognition* 61 (January): 539–56. https://doi.org/10.1016/j.patcog.2016.07.001.

Pires De Lima, Rafael, and Kurt Marfurt. 2019. "Convolutional Neural Network for Remote-Sensing Scene Classification: Transfer Learning Analysis." *Remote Sensing* 12 (1): 86. https://doi.org/10.3390/rs12010086.

Pisman, Ann, S. Vanacker, H. Bieseman, L. Vanongeval, M. Van Steertegem, L. Poelmans, and K. Van Dyck. 2021. "Ruimterapport 2021." Brussel: Departement Omgeving.

Steiner, Andreas, Alexander Kolesnikov, Xiaohua Zhai, Ross Wightman, Jakob Uszkoreit, and Lucas Beyer. 2022. "How to Train Your ViT? Data, Augmentation, and Regularization in Vision Transformers." arXiv. http://arxiv.org/abs/2106.10270.

Sun, Li, Yuqi Tang, and Liangpei Zhang. 2017. "Rural Building Detection in High-Resolution Imagery Based on a Two-Stage CNN Model." *IEEE Geoscience and Remote Sensing Letters* 14 (11): 1998–2002. https://doi.org/10.1109/LGRS.2017.2745900.

Tong, Wei, Weitao Chen, Wei Han, Xianju Li, and Lizhe Wang. 2020. "Channel-Attention-Based DenseNet Network for Remote Sensing Image Scene Classification." *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 13: 4121–32. https://doi.org/10.1109/JSTARS.2020.3009352.

Vogler, John B., and Jelena Vukomanovic. 2021. “Trends in United States Human Footprint Revealed by New Spatial Metrics of Urbanization and Per Capita Land Change.” *Sustainability* 13 (22): 12852. https://doi.org/10.3390/su132212852.

Wang, Di, Jing Zhang, Bo Du, Gui-Song Xia, and Dacheng Tao. 2023. “An Empirical Study of Remote Sensing Pretraining.” *IEEE Transactions on Geoscience and Remote Sensing* 61: 1–20. https://doi.org/10.1109/TGRS.2022.3176603.

Wang, Di, Qiming Zhang, Yufei Xu, Jing Zhang, Bo Du, Dacheng Tao, and Liangpei Zhang. 2023. “Advancing Plain Vision Transformer Toward Remote Sensing Foundation Model.” *IEEE Transactions on Geoscience and Remote Sensing* 61: 1–15. https://doi.org/10.1109/TGRS.2022.3222818.

Wang, Qi, Shaoteng Liu, Jocelyn Chanussot, and Xuelong Li. 2019. “Scene Classification With Recurrent Attention of VHR Remote Sensing Images.” *IEEE Transactions on Geoscience and Remote Sensing* 57 (2): 1155–67. https://doi.org/10.1109/TGRS.2018.2864987.

Xia, Gui-Song, Jingwen Hu, Fan Hu, Baoguang Shi, Xiang Bai, Yanfei Zhong, Liangpei Zhang, and Xiaoqiang Lu. 2017. “AID: A Benchmark Data Set for Performance Evaluation of Aerial Scene Classification.” *IEEE Transactions on Geoscience and Remote Sensing* 55 (7): 3965–81. https://doi.org/10.1109/TGRS.2017.2685945.

Yang, Naisen, Hong Tang, Hongquan Sun, and Xin Yang. 2018. “DropBand: A Simple and Effective Method for Promoting the Scene Classification Accuracy of Convolutional Neural Networks for VHR Remote Sensing Imagery.” *IEEE Geoscience and Remote Sensing Letters* 15 (2): 257–61. https://doi.org/10.1109/LGRS.2017.2785261.

Yang, Yi, and Shawn Newsam. 2010. “Bag-of-Visual-Words and Spatial Extensions for Land-Use Classification.” In *Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems*, 270–79. San Jose California: ACM. https://doi.org/10.1145/1869790.1869829.

Zeng, Lin, and Anu Ramaswami. 2020. “Impact of Locational Choices and Consumer Behaviors on Personal Land Footprints: An Exploration Across the Urban–Rural Continuum in the United States.” *Environmental Science & Technology* 54 (6): 3091–3102. https://doi.org/10.1021/acs.est.9b06024.

Zhang, Fan, Bo Du, and Liangpei Zhang. 2016. “Scene Classification via a Gradient Boosting Random Convolutional Network Framework.” *IEEE Transactions on Geoscience and Remote Sensing* 54 (3): 1793–1802. https://doi.org/10.1109/TGRS.2015.2488681.

Zhang, Jianrong, Hongwei Zhao, and Jiao Li. 2021. “TRS: Transformers for Remote Sensing Scene Classification.” *Remote Sensing* 13 (20): 4143. https://doi.org/10.3390/rs13204143.

Zhang, Weixing, Weidong Li, Chuanrong Zhang, Dean M. Hanink, Xiaojiang Li, and Wenjie Wang. 2017. “Parcel-Based Urban Land Use Classification in Megacity Using Airborne LiDAR, High Resolution Orthoimagery, and Google Street View.” *Computers, Environment and Urban Systems* 64 (July): 215–28. https://doi.org/10.1016/j.compenvurbsys.2017.03.001.