École Polytechnique Fédérale de Lausanne

Environmental Computational Science and
Earth Observation Laboratory (ECEO)

Master Thesis

Predicting Land Usage from Aerial Images with Deep Learning: A Case Study in the Valaisan Alps focusing on Class Imbalance

---

Zermatten Valérie
Prof. Devis Tuia
Dr. Benjamin Kellenberger

February 5, 2021

Key-words:
Deep Learning, Class Imbalanced Distribution, Land Use/Land Cover,
Remote Sensing Scene Classification, Swiss Areal Statistics

Abstract

The Swiss Areal Statistics consists of a land use and land cover classification used in many fields such as urban planning, hydrology or regional administration. These periodic surveys give a deep insight into the territory changes and allow to develop new policies to ensure favourable social, economic and environmental evolution. Up to now, the land use and land cover classification was performed by manual photo-interpretation at the Federal Office for Statistics (OFS). However, recent advances in artificial intelligence offer new perspectives to replace the time consuming classification task by an automatic method.

Deep convolutional neural networks (CNNs) produce promising results in remote sensing scene classification. This method is based on the analysis of the spatial and structural features of an image and attributes a semantic label to each photograph of the territory. Practical applications of such a method on a real life dataset introduce some challenges. On the one hand, the prediction of the categories constitutes itself an important difficulty: the delimitation of each class can be unclear since several categories may exhibit similar objects, or reversely, a large range of distinct features may appear within the same class, leading to poor recognition from the classifier. On the other hand, real world datasets often present unequal representation of the different classes. Most of the samples illustrate a few classes and all other categories are represented only by a small number of examples. This problem is known as the imbalanced distribution and has a detrimental effect on the classification performances.

Motivated by the fact that addressing the class imbalance could improve classification results, we employ resampling techniques and specific loss functions which aim to enhance the prediction accuracy of rare classes. During our study, we employ techniques such as undersampling, two-phases training, class balanced loss, focal loss and equalisation loss. These methods are tested on aerial photographs from the Swiss Alps with 28 different land use labels. However, no significant improvement is perceived for the rare classes accuracy, as the model performances are peculiarly damaged by the presence of ambiguous classes.

Several classes exhibit high visual similarity whereas some other posses impracticable samples. We decided to proceed to a dataset cleaning by clustering classes with high similarity and removing the classes with uncertain label. We apply again the methods targeting the imbalanced distribution problem. After the dataset cleaning, the focal class balanced model leads to a significant improvement in accuracy for the rare classes.

The effectiveness of the proposed techniques are compared to the predictions performed by ADELE, the classification model developed by the OFS. Even though the predictions results on the rare categories are poorer due to our restrained dataset size and the absence of auxiliary data, we reached a similar level of accuracy to ADELE for the more frequent classes.

In sum, this report shows that the use of methods addressing the class imbalance problem improves the prediction accuracy for the rare classes and could be further used for the production of the Swiss Area Statistics.

Code has been made available at: https://github.com/vzermatt/ClassImbalance

Résumé

La Statistique suisse de la Superficie est une classification de l'utilisation et de la couverture du sol utilisée dans de nombreux domaines tels que l'urbanisme, l'hydrologie ou l'administration. Ces relevés périodiques donnent un aperçu approfondi de l'évolution du territoire et permettent d'élaborer de nouvelles politiques pour assurer une évolution sociale, économique et environnementale favorable. Jusqu'à présent, elle était réalisée par des experts en photo-interprétation à l'Office Fédéral de la Statistique (OFS), mais l'intelligence artificielle offre de nouvelles perspectives pour automatiser la tâche de classification.

Les réseaux neuronaux convolutifs (*en anglais, convolutional neural networks, CNNs*) offrent des résultats prometteurs dans la classification des images issues de la télédétection. Cette méthode basée sur l'analyse des caractéristiques spatiales et structurelles d'une image attribue un label sémantique à chaque photographie du territoire. La mise en place d'une telle méthode sur des données du monde réel présentent plusieurs défis. D'une part, la prédiction des catégories constitue en soi une difficulté importante: la délimitation des classes peut être peu claire puisque plusieurs catégories peuvent présenter des caractéristiques similaires ou, à l'inverse, un large éventail d'objets distincts apparaît parfois au sein d'une même classe, ce qui entraine une faible reconnaissance à travers notre méthode de classification. D'autre part, les données du monde réel présentent souvent une représentation inégale des différentes classes, avec une grande majorité des échantillons illustrant quelques classes et les autres catégories ne contenant qu'un petit nombre d'exemples. Ce problème de distribution déséquilibrée des classes a un effet néfaste sur les résultats de la classification.

Motivés par le fait que résoudre le problème de déséquilibre des classes pourrait améliorer les résultats de la classification, nous employons des techniques de rééchantillonnage et des fonctions de pertes spécifiques qui visent à améliorer la précision des prédictions des classes les moins fréquentes. Au cours de notre étude, le sous-échantillonnage, l'entraînement en deux phases, et différentes fonctions de pertes incluant les méthodes appelées *equalisation loss, focal loss* et *class balanced loss* sont testées sur des photographies aériennes dans les Alpes suisses avec 28 labels d'utilisation du sol différents. Ces méthodes ne permettent aucune amélioration de la précision des classes rares, car les performances de notre modèle sont particulièrement dégradées par la présence de classes ambiguës.

Nous avons décidé de regrouper les classes présentant une grande similarité et d'en supprimer d'autres dont les labels les plus douteux. Nous appliquons à nouveau les méthodes ciblant le déséquilibre des classes. Ce processus conduit à une amélioration significative de la précision pour les classes rares.

L'efficacité des techniques proposées est comparée aux prédictions effectuées par ADELE, le modèle de classification développé par l'OFS. Nos résultats concernant les classes rares sont inférieurs à ceux d'ADELE. Cela est dû à notre zone d'étude restreinte et à l'absence de données auxiliaires, notre niveau de précision est comparable à ADELE pour les classes les plus fréquentes. La méthodologie spécifique ciblant le problème du déséquilibre des classes et de la confusion entre les catégories est capable d'atteindre un niveau de précision égal pour les classes fréquentes mais pas pour les autres catégories.

En somme, ce rapport montre que l'utilisation de méthodes pour résoudre le problèmes du déséquilibre des classes mène à une amélioration des prédictions pour les classes rares et pourrait être utilisées dans le cadres de la Statistique de la Superficie de la Suisse.

Contents

List of Figures

List of Tables

1 Introduction

Land cover and land use (LCLU) monitoring plays an important role for long term territory management. This spatial information allows the monitoring of the landscape evolution due to anthropogenic or natural transformations and serves in many fields of applications such as urban planning, regional administration and sustainable development (Liu et al., 2017). Knowing the recent changes in land use as well as their probable modifications makes it possible to verify whether the territorial transformations are in line with land use policies in order to ensure a favourable social, economic and environmental development (Patino & Duque, 2013).

In Switzerland, the Federal Statistical Office (OFS) is in charge of monitoring these changes with regular surveys. Assessing LCLU has traditionally been carried out manually through visual photo-interpretation by experts (OFS, 2017). The classification labels comprehend 46 land use categories such as agricultural land, housing, industrial areas, etc. and 27 land cover classes such as buildings, vegetation or water bodies (OFS, 2016). With one sampling site for each hectare of the Swiss territory, this procedure takes several years to complete and consumes an important part of the OFS resources allocated for the Areal Statistics (Facchinetti, 2019a).

Owing to the current rapid economic and urban development, OFS desires to gradually increase the release frequency for the Areal Statistics from 12 to 6 years and to allocate more budget for data analysis and communications (Facchinetti, 2019b). As traditional methods are particularly limiting regarding time constraints and monetary costs, the future of the Areal Statistics is heading towards automatic classification methods.

A promising direction for this task is to adopt deep learning, a machine learning method characterised by its multiple layers that automatically learn features from data (LeCun et al., 2015). These algorithms constitute the basic building blocks of most computer vision processes. Famous models such as AlexNet (Krizhevsky et al., 2012) or ResNet (He et al., 2016) are able to recognise pictures from a thousand types of objects with excellent precision. CNNs are more and more widely employed on remote sensing data sets (Xia et al., 2017; Yang & Newsam, 2010; Zhao et al., 2016; Zou et al., 2015).

The task of attributing LCLU labels consists in assigning a category to each aerial photograph of the territory, but the complexity of LCLU categories makes it a challenging task. Similarly to object recognition, scene classification includes identifying entities on the images, but it additionally requires to analyse the spatial arrangement of the features and their context (King et al., 2016). For instance, different land use categories can present similar objects but belongs to different LCLU classes (Zhu et al., 2017). Typically, residential and commercial areas both potentially exhibit roads, buildings and trees but are associated with two distinct categories. Only the difference in spatial organisation, such as building density and the construction sizes can determine the category. The problem of low between-class difference becomes even more important in the case of very fine and precise categories (Castelluccio et al., 2015). Conversely images within the same category can present high variability between samples due to different shapes, orientations or spatial structure (Hu et al., 2018). Large intra-class differences also occur in the case of wide class definitions grouping several disparate objects under one label.

Another issue related to the dataset itself implicates the important unequal distribution of samples between categories (Liu et al., 2019). In contrast to carefully filtered datasets where samples are uniformly allocated between the categories, realistic datasets often present a few classes that have a significantly higher number of samples than other classes. This unequal number of in-

stances per category has been shown to have a detrimental effect on the classifier performances (Buda et al., 2018). Traditional deep learning models give equal importance to each sample regardless of the frequency of the label (Krawczyk, 2016). Their predictions tend to favour categories with numerous images and therefore they have difficulties to recognise rare classes. A specific model design must be adopted to give the rare categories a more equal representation. This well-studied challenge is known as the imbalanced distribution or the long-tail distribution and it has been widely discussed over the last decade.

An intuitive solution to overcome the class imbalance problem consists in the use of sampling methods that aim at balancing the data distribution by increasing or reducing the number of samples of a few classes (He & Garcia, 2009). Another approach is to use specific loss functions that adapt the misclassification costs for target samples (Cui et al., 2019; Lin et al., 2018; Tan et al., 2020).

This work explores a number of deep learning techniques that enable the automated classification of land use in the context of the Swiss Areal Statistics. This study focusses on the class imbalance problem and tries to reduce its negative impact on the classification performances. Basic data cleaning methods are applied such as class grouping and class removal to observe the effects of the ambiguous delimitation of categories.

To address this problem, we design a deep CNN and we predict the land use labels on aerial photographs in the Swiss Alps. We compare our results with the performances of ADELE, a deep learning model designed by the OFS to predict the LCLU labels.

In this report, we first introduce our study area and our dataset. Then we present background information related to the Areal Statistics and previous works performed to automate its production. Next, we explain on the working principles of deep neural networks for image classification and the theoretical background on the methods used to tackle class imbalance. We describe our methodology and the experimental results and we discuss them critically.

2 Data

This section describes the methodologies used by the OFS to produce LCLU labels and the main challenges that are expected from this dataset for the task of automatic land use classification.

2.1 Production of Swiss Areal Statistics

The Swiss Areal Statistics is a raw surface statistics performed on the entire surface of Switzerland. It intents to measure the qualitative and quantitative changes in soil coverage and utilisation. As Switzerland is a country with limited space, land is a crucial resource and these statistics help to make decisions in the field of spatial planning and development. These statistics constitute a long-term time-series as they were completed on the exact same locations in 1979/85, 1992/97, 2004/09, 2013/18 and a new survey started in 2020 (Beyeler, 2018).

The semantic labelling work is a demanding task that requires a high degree of reproducibility and reliability, to ensure the compatibility of labels with the LCLU time-series (Facchinetti, 2019a). This exercise is intricate due the large number of categories and their internal complexity with sometimes wide definitions. As a result, this manual work requires expertise and must be carried out with transparent and precise criteria. This extremely time-consuming process performed at the OFS is described below.

Sampling sites and surface of reference
The Swiss Areal Statistics is created by photo-interpretation based on aerial photographs produced by the Federal Office of Topography (Swisstopo). The labels are produced in a point-wise manner on a regular grid formed by hectometric coordinates from the Swiss national maps (OFS, 2016). A reference surface of 50mx50m centred on the sampling site is interpreted. All the sampling sites are located at a distance of 100 meters from each other, leading to more than 4 million points across Switzerland.

The label determining the LCLU category depends on the features located at the exact sampling point location, together with the reference surface that provides contextual information. This manner of labelling can be difficult to predict for deep learning models. Contrarily to the usual classification labels, the land use category of a site does not always correspond to the class that covers the largest area on the photograph. For example in Figure 1, the white square defines two reference surfaces and the yellow crosses are located at the exact position of the sampling sites. Both images show similar features, but they receive different land use labels depending on the central element of the image.

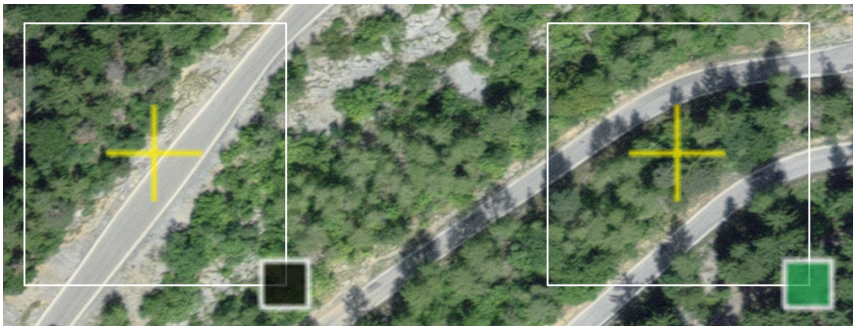


Figure 1: Examples of reference surfaces (white square) with the centre location indicated as a yellow cross. The coloured squares indicate the land use category: black for road, green for forest. ©OFS

Photo-interpretation
The interpreter performs the classification by means of a stereoscopic visualisation system and polarised glasses. With a three-dimensional view on the scenes, he or she is able to distinguish slopes and to give estimates of heights. A second screen providing numerous additional information facilitates the interpretation for the challenging sampling points (OFS, 2017):

- The two most recent Swiss national maps at 1:25'000
- Buildings information such as the constructions usage from the RegBL (Federal Register of Buildings and Housings) or the REE (Register of companies and establishments) and number of inhabitants according to the population census.
- The allocation of building zones from the Federal Office for Land development (ARE) and the cadastral survey for the extend of the habitable plots and industrial surfaces produced by the service for the coordination of geoservices (KOGIS at Swisstopo)
- The inventory of swampy sites, high- and low-marshes, protected by the Federal Office for Environment (OFEV)

- The forest perimeters, information related to forest damages and the canopy height model produced by the Swiss Federal Institute for Forest, Snow and Landscape Research (WSL)
- SWISSIMAGE orthophotos in RGB and FCIR colour, taken three years before or after the survey
- Satellite images providing infrared channel and improving the classification in areas of poor contrast or between agricultural lands and pastures
- The access to the internet if needed

In order to reduce mistakes, a second interpreter will check the labels given by the first operator, except for very vast classes such as forests. Field verification is possible if the category remains unclear after both interpreters observations. The operator gives to each sampling points two labels: one category of land use and one for land cover. The unique combination of these two labels are used to produce the standard label.

As a result, the Areal Statistic is a remarkable labelled dataset for several reasons. First, it comprehends more than 4 million labelled sampling points (OFS, 2017) and its size is relatively large compared to major LCLU benchmarking datasets that possess from a few thousands images for UC-Merced (Yang & Newsam, 2010) up to tens of thousands images with 31'500 samples for NWPU-RESISC45 (Cheng et al., 2017). Secondly, it is a high quality dataset with uniform and reliable labels since it is produced by experts through photo-interpretation and explicit definitions. Moreover, the large number of scene categories makes the Areal Statistics labels very fine grained. Last, it covers a long time frame with permanent locations for the sampling sites, which makes it a reliable and long term database.

2.2 Study area and raw data

Our experiment is based in the greater region of Sion (VS) in the south-west of Switzerland. The study area illustrated in Figure 2 spreads over approximately 600 km^2, with the following coordinates: 2'585'000, 1'117'300 to 2'620'600, 1'134'000 (from south-west to north-east, CH1903+/MN95).

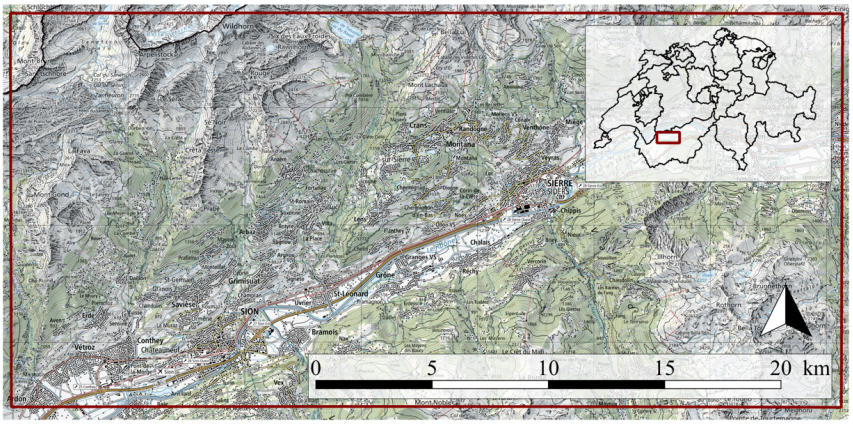


Figure 2: Location of the study area

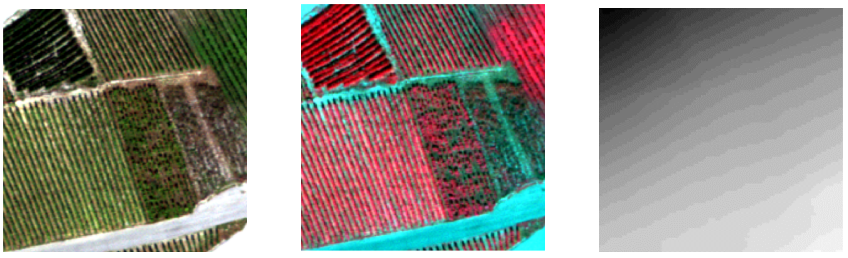
The studied surface exhibits a large variety of different land use types. The low lands are predominantly covered by urban areas and crop fields, the hillsides present mostly forests, vineyards and small villages whereas pastures and rocky areas occupy the higher altitudes.

Data description
Aerial images in the Sion area were collected by Swisstopo between the 26 May and 13 September 2020. Images were captured in five flights campaigns with a plane equipped by a Leica ADS100 digital camera. We use raw digital image strips without post-processing or ortho-rectification. The entire area of interest is covered by images with 25 cm resolution. Additional images with 10 cm resolution are available for the plain areas (Swisstopo, 2020b). The near infrared band (NIR) is available in addition to the three visible colour bands: red, green and blue (RGB). In total, it resulted in 585 images with 16 bit per channels.

The digital elevation model (DEM) originates from the Swisstopo product Swiss Alti3D obtained by digital photogrammetry. It presents the Swiss relief without vegetation and constructions with a spatial resolution of 50 cm. The last update for the area of interest dates from 2016 (Swisstopo, 2018, 2020a). It consists in 962 tiles of size 1 by 1 km.

The labels for the area of interest are based on aerial photography from 2013, they are part of the 2013/18 Areal Statistics survey (OFS, 2017). Labels for 2020 image collection corresponding to the images used on this study are not available yet. Some issues related to the time difference between the production of the labels and the aerial image collection are discussed later in the report.

2.3 Image dataset preparation



(a) (b) (c)

Figure 3: The five bands of a training sample illustrated with (a) the true colour image (red, green and blue bands), (b) the false colour image (near-infrared, red and green bands) and (c) the DEM band

The raw images strips are merged together with *gdal merge*[1] to form regular tiles of 2 by 5 km over the area of interest. Image strips with 10 cm spatial resolution are discarded, as they only cover plain areas and the 25 cm tiles entirely overlap them. Longitudinal (east-west) overlapping areas is automatically processed by *gdal*, as they follow the flight trajectory, but lateral (north-south) overlap is manually treated by selecting the corresponding image strips. Raster tiles for the DEM are produced with oversampling from 50 cm to 25 cm resolution. The reference surfaces are produced by stacking the DEM and the colour images and by extracting the 50x50m reference surfaces corresponding to each sampling site.

[1] https://gdal.org/

Each reference surface is given a tile ID number composed by 8 digits, named RELI formed as

$$RELI = X * 100 + Y/100$$

by the X,Y coordinates of the sampling point in the MN03 coordinate system. This ID number is used to bind the tiles and the labels produced by OFS. As a result, the total dataset obtained 59'976 tiles with size 200x200 pixels with five channels: near-infrared, red, green, blue and the oversampled DEM.

2.4 Dataset characteristics

For our study, we decide to use the land use labels, because they are less sensitive than land cover to the object located at the exact centre of the image. We also discard the standard classification as it derives from a logical conversion table from both LCLU, and the number of classes is much higher (72 instead of 46), leading to a reduced number of samples per class.

The land use categories contain 46 basic classes from four main domains: settlement and urban areas, agricultural lands, forests and unproductive areas. The OFS produces a documentation with a detailed description for each land use category (OFS, 2016, 2017). The labels are freely accessible online through the Swiss geodata catalogue[2].

The imbalanced distribution

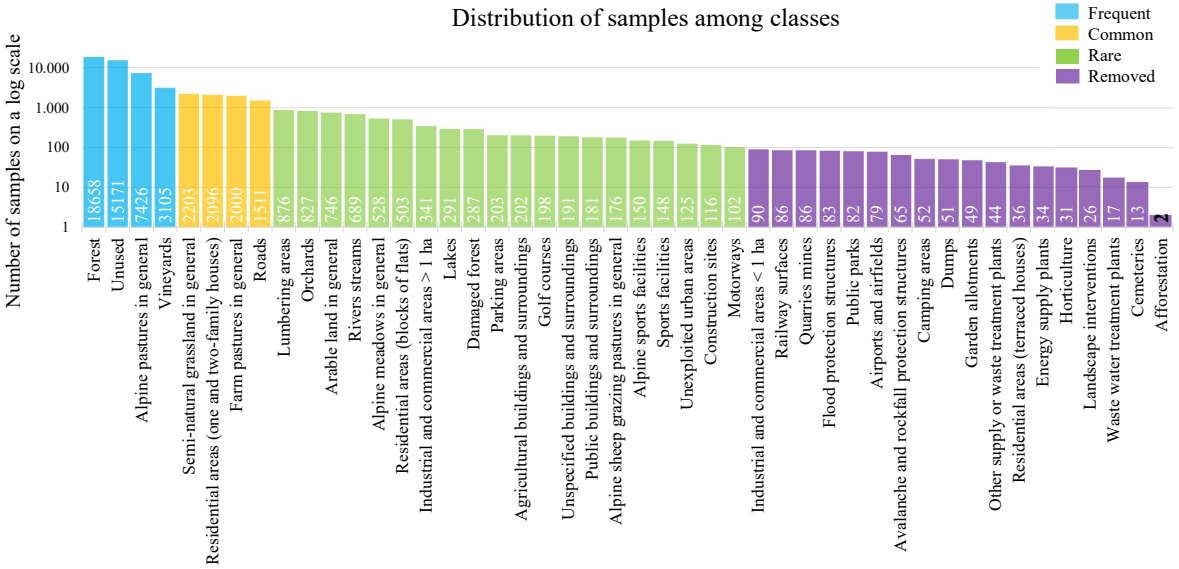


Figure 4: Number of samples per class for the land use categories in the 2013/18 survey over the study area, on a log scale.

Our dataset faces a typical problem of the real world dataset: The distribution of unprepared data shows important inequalities between categories (Liu et al., 2019). As illustrated in Figure 4, our study area presents all 46 classes but with a strongly skewed distribution towards some very

[2]Available at: https://www.bfs.admin.ch/hub/api/dam/assets/14607225/masterl

frequent categories: 10 land use categories make up 90% of the data, whereas the 10% remaining samples represent the 36 tail classes. Typically, forests cover about one third of Switzerland (OFS, 2015) and a similar share in our samples illustrates this class, whereas less frequent land use types such as waste water treatment plants or cemeteries are much rarer with only a few occurrences.

The class imbalance ratio computed as the minimum number of samples over maximum number of samples across all classes is the range 1:10'000 (2:18'658). This ratio is very large in comparison to other study targeting the imbalanced distribution, where they usually target imbalance in the range of 1:10 to 1:500 (Cui et al., 2019; Tan et al., 2020), but very few studies exist on extremely imbalanced data in ranges from 1:1000 to 1:5000 (Krawczyk, 2016). Due to their small number of samples and the extreme imbalance ratio, we removed categories with less than 100 samples, making a total of 18 suppressed classes, and established a more usual imbalance ratio of 1:200 (100:18'658). As a result, the dataset used in the following experiments is comprised of 59'047 tiles representing 28 categories. Images from the 28 categories are exposed in the Appendix (Figures 27,28,29).

Figure 4 splits the samples into four bins: the frequent categories in blue (>3'000 samples per class), the common ones in yellow (between 1'000-3'000 samples per class), the rare ones in green (<1'000 samples per class) and the removed ones in purple (<100 samples).

(a) Blocks of flats (b) One and two-family houses (c) Terraced houses

Figure 5: Three categories of residential areas

The absence of auxiliary data

We anticipate another drawback from this dataset regarding its fine grained classes. Usually, a large number of categories is perceived as an advantage for a land use dataset, since it implies smaller differences between samples from the same category (Xia et al., 2017). Consequently, a high uniformity within all class members would typically improve their identification by a deep learning framework and it could potentially lead to better performances. However, this principle does not apply in our dataset. Since labels are attributed by photo-interpretation with auxiliary information, the aerial photograph itself occasionally does not contain all the features required to differentiate between related classes. Accurate predictions for categories with similar aspects may become very difficult.

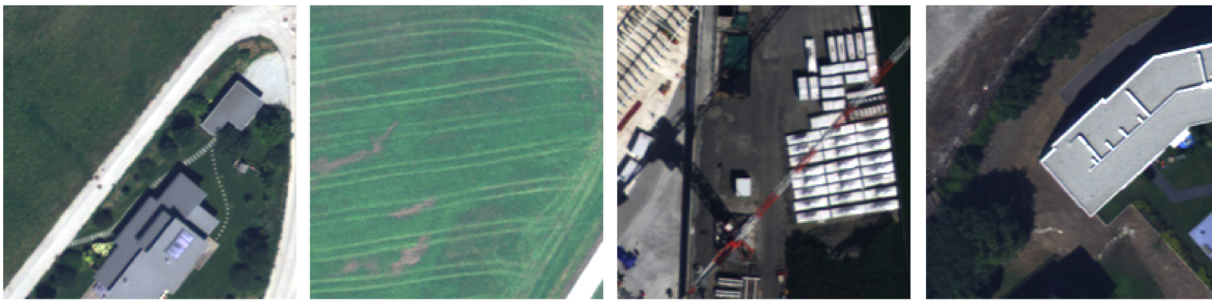
The OFS provides through several documents (OFS, 2016, 2017) a detailed definition of all classes and the auxiliary information employed to recognise them. Below we present a few examples of classes whose identification would be arduous without external data.

- Industrial and commercial areas are spilt in two separate categories depending on their total extent occupying more or less than one hectare. The size of interpretation surface spreads on a much larger area than the reference surface used in our study and the aerial pictures alone are not sufficient to distinguish these two categories.
- Categories of residential areas shown in Figure 5 consist of terraced houses, one or two-family houses and blocks of flats and require cadastral information such as the count of inhabitants or the number of floors to be distinctly identified from an aerial point of view.
- A list of camping surfaces are provided to the OFS interpreter. Even if some samples might be relatively straightforward to identify visually with the regular presence of small construction sites (see 6a), access roads and trees, other parts of the camping sites may look very similar to forests, parking lots or residential buildings (see 6b) and become difficult to attach to their usage without auxiliary indications.

(a) (b)

Figure 6: Camping sites with typical features in 6a and more difficult samples in 6b

The time gap between the labelling and the data collection
An important time gap of 7 years exists between the production of the land use labels in 2013 and the aerial photography survey in 2020. We observe that some sampling sites have changed their usage through this period, as a result of the land use evolution. This randomly affects a number of samples from most of the classes as illustrated on Figure 7. According to the report from the company Picterra (Picterra, 2017), 9.5% of the land use labels changed between the 2013/18 Areal Statistics survey and the one from 2004/09. Thus it is likely that a similar amount of labels in our study has changed during the time interval of 7 years.



Construction site →Residential area | Landscape interventions → Arable land | Semi-natural grassland → Construction site | Vineyards → Residential area

Figure 7: The visible land use differs from the one of the label due to time difference. The legend indicates first the true category present on the label from the 2013 survey followed by a more realistic suggestion of label for the picture in 2020.

3 Deep Learning for Land Use Classification

This section gives some background explanations on automatic land use classification with deep learning models. We first describe the concepts and the key features of deep learning algorithms and the training process. Next we introduce previous studies performed for the Areal Statistics. The section 4 focusses on the methods to address imbalanced datasets with deep learning.

3.1 Deep Learning for Image Classification

Deep learning is a method of machine learning that learns data representation through its multiples layers. Through their adaptative architecture, the layers memorise patterns in data such as pictures or one dimensional data. This allows them build a complex representation of concepts by combining simpler features (Goodfellow et al., 2016).

Convolutional neural networks (CNNs) are one of the most commonly used deep learning algorithm and they are designed to process and analyse images (LeCun et al., 2015). CNNs layers are composed of artificial neurons that are interconnected. The outputs from one layer are given as inputs to the next layers. The initial layers extract simple features such as edges from the raw input, whereas deeper layers identify more complex pattern such as digits or objects. For instance, a simple neural network was already used by LeCun et al. (1989) to recognise hand written digits for the U.S. Postal Service.

The development of performant computer architectures and the increasing amount of available training data have lead deep learning algorithms to solve increasingly complicated questions over time. Deep learning is now used in a wide variety of disciplines such as speech and audio processing, natural language processing, bioinformatics, medicine, video games, search engines, online advertising and finances (Havaei et al., 2017; LeCun et al., 2015; Zhang et al., 2020; Zimmermann, 2018).

3.1.1 Layers in Convolutional Neural Networks

A CNN is composed of a sequence of interconnected blocks, each of them performs a series of non-linear operation that transforms the data representation.

- In a typical deep CNN, data are fed to the network via the **input layer** and are stacked into groups called batches. Each batch contains a number of images as a multi-dimensional array.
- The **convolutional layers** perform the basic operations required by the CNN on the input batches. The convolution operation consists of a filter containing a pattern of interest sliding over the entire image. Formally, it computes the dot product of two matrices, the filter and a patch of the image for each pixel on the picture.

  Sliding the filter on the entire image produces activation maps illustrated in Figure 8. They show high values where the images exhibit patterns similar to those of the filter. The size of the convolutional filter determines the size of the neighbourhood taken into account during the convolution operation. Convolutional layers are very powerful tools for image analysis. They are translation invariant, meaning that they are able to recognise patterns independently of their location within the image (Zhang et al., 2020).

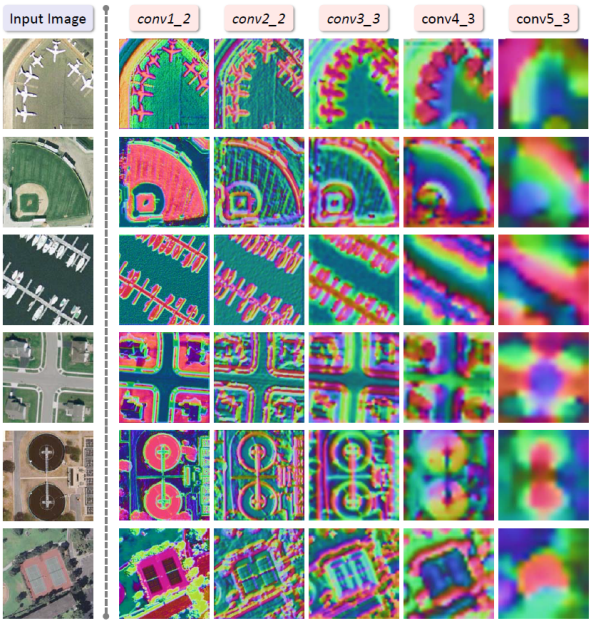


Figure 8: Visualisations of activation maps produced by 5 successive convolutional layers on some land use images (Hu et al., 2016)

- **Normalization layers** such as batch normalization (Ioffe & Szegedy, 2015) commonly follow a convolutional layer in order to recentre and rescale the outputs of the convolution. They allow stable and fast learning which in turn reduces the required number of iterations during the classifier training. Several alternative normalization methods exist, arguing that the batch normalisation applied on the entire batch might be biased for small batch size. On the one hand, group normalization (Wu & He, 2018) divides the channels into groups and computes within each group the mean and variance for normalization. On the other hand, instance normalization (Ulyanov et al., 2017) goes one step further and computes the means and variance for each channel in each training example.
- CNNs contain one or several **pooling layers**. Like a convolutional layer, the pooling operator consists of a window that slides over the inputs, but it computes a single output

for each patch (Zhang et al., 2020). This process leads to a gradual reduction of the spatial resolution and enables features grouping to extract patterns. This downsampling operation also reduces the network sensitivity to the exact location of an object.

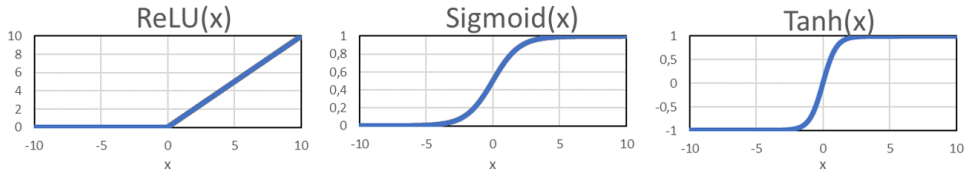


Figure 9: Non-linear activation functions

- To go from one layer to the next, the inputs passes through a **non-linear function** that acts as a selective threshold and allows the network to model non-linear data. By stacking several of these non-linear functions, the network can recognise very complex details from the input. A popular non-linear function is the rectified linear units or $ReLU = max(x, 0)$ that was first introduced by Krizhevsky et al. (2012). Other non-linearities exist such as hyperbolic tangent or sigmoid (see Figure 9) but ReLU is usually preferred since it is several times faster to compute and does not saturate contrarily to the sigmoid and the hyperbolic tangent.
- He et al. (2016) developed a network called ResNet that possesses **shortcut connections** by-passing convolutional layers. This short-cut connection can be seen as a block encompassing several convolutional layers with batch normalization and ReLu layers. This connection directly feeds its input by identity mapping i.e. without modification to the outputs of the stacked layers. This new architecture allows a better training of very deep networks while keeping a low complexity, leading to very performant networks.
- CNNs used for classification usually terminate with one or several **fully connected layers** that combine the network output features together and passes them to an **activation function** such as softmax which finally attributes to each input a predicted category.

3.1.2 Loss functions

During the training of a CNN, the layer weights are periodically updated, in order to reduce the error rate on the training set. To do so, a measure of fitness is required. The loss function, also called criterion or objective function, quantifies the distance between the ground truth and prediction through a numerical value (Zhang et al., 2020).

The loss is usually a non-negative number, where perfect predictions receive values close to zero and larger losses indicate less accurate predictions. For tasks such as classifications with CNNs, we need a loss function able to compare categorical values. The **cross entropy loss** is commonly used in this case. It allows to compare two probability distributions. The ground truth labels can be seen as binary distribution and the outcome of the network gives an estimate of the probability for each sample regarding all categories.

The cross-entropy loss is often combined with a **softmax activation function** $\sigma(\hat{y})$. This function takes as input the output of the network $\hat{y}$ and transforms it into a probability distribution consisting of positive numbers summing to 1. The softmax activation function regards each class as mutually exclusive (Tan et al., 2020). If the likelihood of one class increases, the other have to

decrease by an equal amount. It is widely adopted in image classification. The predicted output probabilities for all class is $\hat{y} = [\hat{y}_1, \hat{y}_2, ..., \hat{y}_C]^T$ with C the total number of categories. For a ground truth label y_c the softmax function $\sigma(\hat{y}, y_c)$ can be written as:

$$\sigma(\hat{y}, y_c) = \frac{exp(\hat{y}_c)}{\sum_{k=1}^{C} exp(\hat{y}_k)} \tag{1}$$

The softmax cross entropy loss L_{SCE} is computed between the ground truth label y_c and the output probabilities $\hat{y}$ for all classes. The cross entropy loss L_{SCE} can be formulated as:

$$L_{SCE}(\hat{y}, y_c) = -\log\left(\frac{\exp(\hat{y}_c)}{\sum_{k=1}^{C} \exp(\hat{y}_k)}\right) \tag{2}$$

In practice the ground truth labels are one hot encoded:

$$y_k = \begin{cases} 1 & if \quad k = c, \\ 0 & otherwise \end{cases} \qquad and \qquad \sum_{k=1}^{C} y_k = 1 \tag{3}$$

3.1.3 Optimiser

The optimisation of a CNN consists of iteratively updating the model weights by a small amount in order to minimize the loss function. The **learning rate** determines the size of the changes in the model parameters. Large learning rates may cause instability in the learning process whereas too low learning rates may result in a model unable to converge (Goodfellow et al., 2016). The learning rate is usually reduced during the training, a large variety of schedules exists. A common practice is to start with an elevated value to cover a rather broad range of parameters, before continuing with a smaller value in order to finely estimate the best weights for the network.

A simple optimisation method is to use a gradient descent that basically follows the steepest descending gradient of the loss function. However, the computational cost of this optimizer increases linearly with the dataset size, leading to very long computations for large datasets (Zhang et al., 2020). The **stochastic gradient descent** (SGD) reduces its computational cost by calculating the gradient on a random set of samples at each iterations.

Adam or "adaptative momentum" (Kingma & Ba, 2015) has become one of the most robust and effective optimisation algorithms to use in deep learning (Zhang et al., 2020). One of Adam's advantages over SGD is that it automatically adapts the learning rate based on the gradient momentums. Furthermore, the weight updates of Adam are independent of the magnitude of the gradient, which allows faster convergence when going through areas with small gradients, where SGD may sometimes get blocked.

3.1.4 Addressing overfitting

CNNs are high capacity models and therefore are prone to overfitting or over-training. This means that is the model tends to memorise properties of the training data that do not help them to generalise on unseen data. As a result, it performs poorly on unseen data (Zhang et al., 2020).

Models overfit for several reasons: First, a small number of training examples easily lead to overfitting, as the small number of images is not representative of the full data distribution and the model would not generalise well on unseen data. Second, models also tend to be more

susceptible to overfitting when the number of parameters is high, as they basically have more neurons to memorize the inputs. Further, it is also important to adapt their weights by small increments at the time to lower the effect of peculiar samples. Overfitting often appears when weights can take a too wide range of values (Goodfellow et al., 2016). Last, more training iterations give the model more opportunities to learn small details which are not useful for generalisation (Zhang et al., 2020). Several **regularization techniques** were developed against overfitting, a few important ones are presented here below.

- **Weight decay** is used to keep the model parameters small by adding a penalty term to the loss function (Zhang et al., 2020). The penalty term increases when the neurons weights and biases grow, in order to avoid extreme values.
- **Early stopping** is a strategy for saving training time and reducing the risk of overfitting. As the model trains, the error on training set steadily decreases over time but sometimes the validation error starts to increase again. The network stops improving on unseen data and become worse, a sign of overfitting. The training can be terminated after a defined number of epochs without improvements and the weights with the best validation accuracy can be selected (Goodfellow et al., 2016).
- In 2014, Srivastava et al. (2014) noticed that injecting noise to the network layers enforces the model generalisation ability. His method called **dropout** randomly disable outputs from neurons in the network in order to push the model to learn features that are useful on their own and avoid co-adaptation between neurons. Similarly, each hidden unit in a neural network trained with dropout has to learn to work with a randomly chosen sample of other units. During the testing phase, dropout is not used. This computationally inexpensive yet powerful method is now a standard technique for CNN training (Goodfellow et al., 2016).
- **Dataset augmentation** is another way to inject noise in the learning process. A fundamental rule of deep learning is that more data lead to better models, but data are limited and expensive to collect. However, creating fake data has shown to be particularly effective for image recognition. An enormous range of factors of variation can be easily simulated such as scale augmentation, horizontal and vertical flips, random rotations and colour modifications in hue, brightness, contrast or luminosity (He et al., 2016). In any case, special care must be given not to apply transformations that would change the correct label or would be unrealistic given the data.

3.1.5 Transfer learning

Fully training a CNN requires a considerable amount of data. Even though large remote sensing datasets are publicly available, large labelled image datasets are far less common. Training a deep learning model from scratch on a limited dataset would quickly lead to a model with poor generalisation ability even with regularization techniques (Castelluccio et al., 2015).

Transfer learning attempts to adapt the knowledge learned in one computer vision field to another. The idea is that patterns learned by a CNN such as edges, geometric shapes or colours as illustrated in Figure 10 are identifiable across datasets illustrating different subjects, and could be transferred to another classification task as well (Castelluccio et al., 2015).

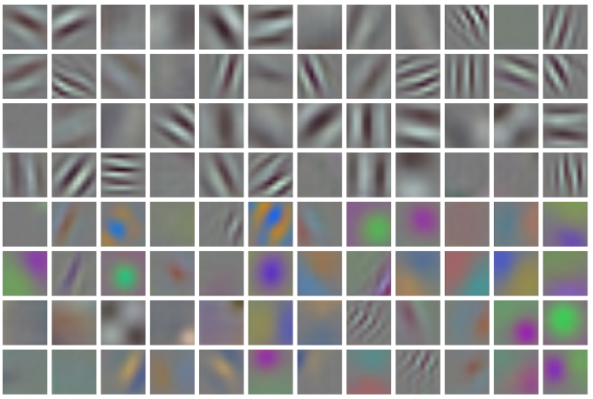


Figure 10: Illustration of the filters learned by the first layer of a deep learning model trained for object classification. It is sensitive to patterns with various frequencies and orientations and different colours transitions (Castelluccio et al., 2015).

The simplest form of transfer learning consist in using features extracted from a model pre-trained on a reference dataset and give them to a classifier. Popular dataset for pre-training include ImageNet (Russakovsky et al., 2015) for object classification, MS-COCO (Lin et al., 2014) or Places (Zhou et al., 2018) for images of every day scenes and AID (Xia et al., 2017) or UC Merced (Yang & Newsam, 2010) for remote sensing scenes recognition.

Adapting the network parameters makes better usage of the CNN capacities. The weights of the pre-trained model are only used to initialise the model parameters. The **fine tuning** of the model parameters allows a deeper adaptation of the model to the data of interest as all the convolutional filters within the CNN can learn patterns specific to the new dataset. It is now a standard method for obtaining baseline results on a new target dataset (Van Horn & Perona, 2017). Pre-trained CNNs on the ImageNet dataset have been successfully used on optical remote sensing images (Castelluccio et al., 2015; Nogueira et al., 2017), since they show similar basic features.

3.2 Existing work on the Swiss Areal Statistics

The OFS proposed a new classification method involving partial automation with several objectives (Beyeler, 2018; OFS, 2016, 2017). The most important point is that the time interval between two surveys must be reduced to 6 years, and the new process must ensure the continuity with the labels from the chronological time series. Additionally, the visual interpretation workload should be reduced in order to develop the analysis and the diffusion of the Areal Statistics results. OFS evaluated the suitability of complementary data sources for spatial information such as satellite images or crowd sourced geospatial data such as OpenStreetMap (OSM). The introduction of this new methodology is the source of several studies and reports aiming at analysing and developing a suitable classification model (Jordan, Lack, et al., 2019; Jordan, Meyer, et al., 2019; Lutz, 2019; Picterra, 2017; Schar et al., 2017). Two of them are presented in the following sections.

3.2.1 Picterra feasibility study

The start up Picterra[3] presented a comprehensive feasibility study matching the requirements of the methodology revision and tested several possible models (Picterra, 2017). They demonstrated that the Areal Statistics with the abundant existing labels is an typical task for the application of deep learning (Beyeler, 2018). Their model is based on a ResNet-50 model that was pre-trained on ImageNet and acts as a feature extractor, and a random forest as classifier. The random forest classifiers are based on a collection of random trees, that are a set of rules organized in a hierarchical manner to predict data values. Their experiment involved raw aerial image classification for separated LC and LU predictions. Their model performed well on LU categories with several thousands images per class for training, but the predictions are drastically degraded when classes possess less samples.

As the aim of their work was not to produce a fully functional model but to test several approaches, their results for LCLU classification show an important potential for improvement. Nevertheless, their report mentioned significant limitations and remarks that were useful for our study. After comparing several methods for the Swiss Areal Statistics, such as change detection or LCLU classification, their results concluded that LCLU categorisation with a deep CNN would be the most appropriate method for the OFS task. The usage of digital surface model and near-infrared band in addition to the RGB colours would probably improve the model predictions. Auxiliary information such as those used by OFS would also ameliorate results. Fine-tuning and data augmentation are not applied, but would probably aid the classifier.

They targeted the issue of imbalanced distribution by simple frequency re-weighting for LC and LU, but no significant performance improvement was observed. In addition to that, their work indicated that classes with abundant samples and rather simple patterns are more likely of being supported by automation. By being able to accurately predict the majority classes, the visual interpretation work load could be reduced by 50% (Beyeler, 2018).

3.2.2 ADELE

ADELE or the *Arealstatistik DEep LEarning* project is ordered by the OFS and performed by the Fachhochschule Nordwestschweiz (FHNW) in collaboration with the company ExoLabs. It aims at developing deep learning technologies for automatic or semi-automatic image classification of the Swiss Areal Statistics (Jordan, Meyer, et al., 2019). The ADELE model is based on a random forest classifier with several types of inputs: Predictions of land use classes from a CNN based on 50m by 50m aerial photographs with RGB and FCIR imagery with 25 cm spatial resolution, a time series of 12 LANDSAT images, the cadastral surface category from the building register and the forest perimeter and the canopy height model produced by the Swiss Federal Institute for Forest, Snow and Landscape Research (WSL).

Their best land use classification model trained on 2013/2018 data uses 665‘401 samples for the training and the same number for the testing phase. They achieve an 84.2% overall accuracy with a precision over 90% for 10 classes over 46 LU classes (Jordan, Meyer, et al., 2019). These classes represent 44% of the sampling points and include categories with abundant samples and homogeneous features such as arable lands, alpine pastures, forests or lakes. Some rare classes with a frequency below to 0.05% (less than 1 sample in over 2’000) for instance cemeteries, parking lots or golf fields are recognised with a precision of over 70%.

[3]More details at: picterra.ch

On the one hand, particularly low scores are recorded in rare classes with a small number of samples that also present a heterogeneous aspect (i.e. alpine sport facilities, landscape interventions). On the other hand, rare classes that are visually similar to very large classes tend to be misclassified into the larger class. For example, afforestation, damaged forests and lumbering areas tend to be misclassified as forests.

Interestingly, they also give an insight on the weight received by each data type in the random forest classifier. The features extracted by the CNN on the RGB images and satellite time-series are the principal parameters and represent close to 50% (20% and 25% respectively) of the weights. The building register (16%), the false colour images (13%) and the altitude (11%) also bring valuable information.

Some potential sources of improvement are mentioned such as grouping into one classes the categories that present a high similarity. The fine classification could be performed by a second classifier that learn to specifically distinguish these classes. They also consider that alternative sampling methods such as oversampling the minority classes and undersampling the majority classes. Moreover, they suggest a better time consistency between the labels and the other information sources such as satellite and the elevation model, could yield better accuracies.

4 The Class Imbalance Problem

This section introduces the effects of the imbalanced distribution on the learning process. Next we present several samplings methods and loss functions that we will use in our experiments.

4.1 Effects of the imbalanced distribution

CNNs require a significant amount of data for training. The best results are achieved on reference datasets where careful collections and filtering processes are applied (Dong et al., 2019). However, real world datasets often exhibit more challenging features with significant imbalance and some class overlap (He & Garcia, 2009). As a result, the model parameters will be largely driven by the classes with abundant samples, while results are degraded for the under-represented classes with less representatives.

The impact of class imbalance has been shown to increase with the scale of the imbalance, i.e. the more imbalanced a dataset, the more degraded its performances. Since the tail usually contains most of the categories, it dominates the average classification results (Van Horn & Perona, 2017). The reason is that with increasing imbalance, the availability of training data per class become scarce. Thus the recognition and the generalisation ability of the model on unseen data are reduced.

The number of training images per class is a critical factor in the context of class imbalance: classification error more than doubles every time the number of training images is cut by a factor of 10 (Van Horn & Perona, 2017). This is due to the fact that popular learning algorithms expect balanced distribution and adapt their weights with equal misclassification costs for all samples. Hence the model ability to properly learn the minority classes is compromised, since the rare samples are observed much less often by the model (He & Garcia, 2009). While majority classes receive appropriate labels, rare categories tend to be misclassified into majority categories (Wang et al., 2016).

As a result, special attention must be given when training a CNN on an imbalanced dataset. Minimizing the skewed distribution by collecting more tail samples is a difficult and expensive task when constructing datasets (Wang et al., 2020). Thus the class imbalance problem needs to be addressed by other means. The following section reviews different techniques that have been typically used to solve this issue including the sampling methods, the algorithmic methods and the mixing of these two types of methods.

4.2 Sampling methods

Sampling deals with the data imbalance problem from a data perspective by artificially balancing the distribution of the input data (Wang et al., 2016). Various sampling techniques have been proposed, the most commonly used are presented below.

4.2.1 Oversampling

Oversampling is an intuitive solution to overcome the class imbalance problem by simply replicating the samples from the minority classes. Randomly chosen samples are duplicated and passed to the training set, increasing the total number of samples up to the level that completely eliminates the imbalance (He & Garcia, 2009). This method was shown to be effective and is frequently used thanks to its simplicity (Buda et al., 2018; Wang et al., 2016). However, increasing the size of the training set augments the convergence time for the model. Moreover, the models tend to show overfitting, as replicated data are memorised by the network leading to poor performance on unseen data (Buda et al., 2018; He & Garcia, 2009; Wang et al., 2016).

4.2.2 Undersampling

Undersampling is classical method of sample selection. They are typically used with traditional machine learning methods, such as decision trees as they require less samples for training. It is now also successfully applied on deep learning models (Buda et al., 2018). Random majority oversampling simply removes a portion of the frequent classes in order to reduce the imbalance, sometimes until all classes have the same number of samples. A significant disadvantage of this method is that it discards a proportion of available data and it may cause the classifier to drastically reduce its performances on the most abundant class.

For extreme ratios of imbalance and a large portion of classes belonging to the minority, undersampling performs similarly to oversampling. If training time is an issue, undersampling is a better choice than oversampling since it reduces the size of the training set (Buda et al., 2018). Therefore, undersampling is often preferable to oversampling (Cui et al., 2019; Huang et al., 2016).

4.3 Loss functions

In addition to sampling techniques, another way to deal with the class imbalanced distribution is the use of specific loss functions that solve the data imbalance problem by taking into consideration the penalty associated with misclassifying samples (He & Garcia, 2009). The false classification of a sample is more or less penalised depending on its difficulty or its rarity. This chapter describes several methods that tackle the class imbalance problems with specific loss functions.

4.3.1 Inverse class frequency weighting

A common method for addressing class imbalance is to introduce a weighting factor for each class that matches the data distribution. Re-weighting by inverse class frequency is frequently adopted (Huang et al., 2016; Wang et al., 2017). In practice, a smoother version of this loss, the square root inverse class frequency has shown better results (Mahajan et al., 2018). For a sample of true class c, its frequency λ_y corresponds to the ratio of samples with label c over the total number of samples in the dataset. The inverse class frequency weight for the class c equals $w_c = 1/\lambda_c$, the squared inverse frequency weight is $w_c = 1/\lambda_c^2$. The re-weighted cross entropy loss L_{IF} is defined as:

$$L_{IF}(\hat{y}, y_c) = -w_c \; L_{SCE}(\hat{y}, y_c) = -w_c \; \log\left(\frac{\exp(\hat{y}_c)}{\sum_{k=1}^{C}\exp(\hat{y}_k)}\right) \tag{4}$$

The effects of this method are similar to those of oversampling, as the importance of all rare samples are multiplied to approach the one of majority classes, without the disadvantages of increasing training time. Consequently, oversampling is not be implemented in our experiments.

4.3.2 Class balanced loss

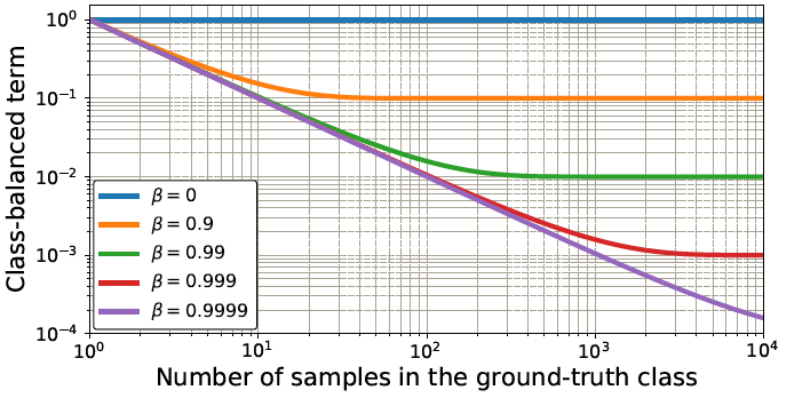


Figure 11: Visualization of the class-balanced term for various number of samples in the ground truth and different values of β (Cui et al., 2019). Both axis are in log scale.

Cui et al. (2019) developed a more advanced re-weighting strategy, the class balanced loss L_{CB}. They argue that when the number of sample increases within a class, the benefit of these new samples decreases, as the information they carry tend to overlap more and more.

The L_{CB} defines an effective number of samples for each category that represents the theoretical number of independant samples representing a class. Similarly to the re-weighting method, each class receives a weighting term depending on the inverse number of effective samples. This technique has shown significant improvement in performances compared to commonly utilized loss functions on long-tailed datasets. The L_{CB} is easy to implement, as it only requires to add a class balanced weight term to the loss functions. It can be combined with different loss functions such as the softmax cross entropy or the focal loss (see Section 4.3.4).

The class balanced loss can be written as:

$$L_{CB}(\hat{y}, y_c) = \frac{1}{E_n} L_{SCE}(\hat{y}, y_c) = -\frac{1}{E_n} \log\left(\frac{\exp(\hat{y}_c)}{\sum_{k=1}^{C}\exp(\hat{y}_k)}\right) \quad with \quad E_n = \frac{1-\beta^{n_c}}{1-\beta} \tag{5}$$

E_n denotes the effective number of samples, with n_c the number of samples for the ground truth class c. The hyper-parameter $\beta \in [0, 1]$ enables to adjust between no re-weighting with β values close to zero and re-weighting similarly to inverse class frequency with β values close to one. Figure 11 illustrates the effect of number of samples in the ground truth class on the class-balanced term for different β values. As our dataset contains classes with size comprised between roughly 100 to 20'000, interesting β values for our study range from 0.99 to 0.9999. A smaller β will give a more equal weight term for all classes, whereas higher β values will reduce the weights of samples from the majority categories.

4.3.3 Equalization loss

Tan et al. (2020) analysed the problem of extreme imbalance with a new perspective. For one category, samples of the other categories are seen as negative samples. When a sample of a certain class is utilized for training, the commonly used losses such as softmax cross entropy give to other classes a small but non-zero discouraging gradient. Since objects of rare categories occurs much less often than those from the majority classes, the predictor for these categories receive mostly discouraging gradients. The minority categories weights might be overwhelmed by negative samples, leading to poor predictions performances. Finally, even samples from the rare categories receive a low probability of predictions from the model.

The idea behind equalization loss L_{EQL} is to bring all the classes to a more equal status by introducing an ignoring strategy. This method uses a weight term for each sample from the rare categories, which reduces the influence of the negative samples. Experiments on several reference datasets for image segmentation and image classification demonstrated the effectiveness of the equalization loss.

The softmax equalization loss is a variant of the softmax cross entropy loss with a specific weighting term $\tilde{w_k}$ introduced in the softmax function. For a sample with ground truth label c and a predicted probability $\hat{y}$, the L_{EQL} can be formulated as:

$$L_{EQL}(\hat{y}, y_c) = -\log\left(\frac{\exp(\hat{y}_c)}{\sum_{k=1}^{C}\tilde{w_k}\exp(\hat{y}_k)}\right) \tag{6}$$

$$with \qquad \tilde{w_k} = 1 - \theta T_\lambda(f_k)(1 - y_k) \tag{7}$$

The weighting term $\tilde{w_k}$ is composed with three binary terms. The random variable θ is used to randomly maintain the gradient of negative samples with a probability of ρ to be 1 and $1-\rho$ to be 0 and takes values within the range $[0.5, 1]$. The ground truth distribution y is one-hot encoded. For a sample with ground truth category c, we have:

$$y_k = \begin{cases} 1 & \text{if } k = c \\ 0 & \text{otherwise} \end{cases} \qquad \theta = \begin{cases} 1 & \text{with probability } \rho \\ 0 & \text{with probability } 1-\rho \end{cases} \tag{8}$$

Finally, the threshold function T_λ is used to distinguish the rare classes from the majority classes. The frequency f_j of the category j equals the number of samples for the j class divided by the number of samples in the entire dataset. The frequency utilised to distinguish minority categories from other is called λ. The tail ratio TR is used to find the value of λ, it equals to the number of images in rare classes divided by the total image number. Experimentally, values of TR between 2-10% have shown to work best.

$$T_\lambda = \begin{cases} 1 & \text{if } x < \lambda \\ 0 & \text{otherwise} \end{cases} \qquad with \qquad TR(\lambda) = \frac{\sum_{j=1}^{C} T_\lambda(f_j) N_j}{\sum_{j=1}^{C} N_j} \tag{9}$$

In other words, the rare categories are defined by a threshold (λ) and the negative gradient from negative samples for the minority categories are ignored. However, if all the negative samples were ignored, the rare categories would be too easy to guess and the model would predict a large number of false positives. Thus, the rare categories still receive some negative gradients from a number of randomly chosen negative samples due to the θ variable.

4.3.4 Focal loss

The focal loss has been designed by Lin et al. (2018) to specially tackle the class imbalance problem in object detection, but it works well for other tasks such as multi-classes image recognition. The idea behind focal loss is to down-weight the losses assigned to well classified samples and focus on more difficult samples. As mentioned for the equalization loss, the commonly used cross entropy losses can be problematic with imbalanced datasets and produce a discouraging gradient for the rare classes, leading to poor prediction performances for minority categories. The focal loss tries to avoid this issue with a scaling factor that decays to zero for well classified samples. When a sample is misclassified and its confidence is small, the scaling factor is close to 1, and the loss is unaffected. When classification confidence is high, the scaling factor and its loss are close to zero.

The focal loss L_F can be used with the softmax cross entropy loss multiplied by a modulating factor α_t for each sample. In addition to that, a weighting term can be used for each class. For a sample of ground truth category y_c, the focal loss with a softmax activation function is defined as:

$$L_F(\hat{y}, y_c) = -\alpha_t \ (1-\hat{y})^\gamma L_{SCE} = -\alpha_t \ (1-\hat{y})^\gamma \log\left(\frac{\exp(\hat{y}_c)}{\sum_{k=1}^{C}\exp(\hat{y}_k)}\right) \tag{10}$$

In the equation above, $\hat{y_c} \in [0,1]$ is the model estimated probability for the class with label c. The tunable hyper-parameter $\gamma \in [0,5]$ determines the rate at which easy samples are down-weighted. The value $\gamma = 0$ cancels the focal loss and leads to a softmax cross entropy loss. The categorical weighting factor is present as $\alpha_t \in [0,1]$ for each class. For instance, it shows good performances for inverse class frequency weighting. The $(1-\hat{y})$ term differentiates between easy and difficult samples.

4.4 Other methods

Two-phases training is a method that trains the network in two steps (Havaei et al., 2017). First, the model is trained on a balanced dataset with equi-probable class frequency. Next the output layers are fine-tuned on the entire dataset with a distribution of the labels corresponding to the real world distribution. The basic idea is to have the benefits from undersampling where the network learns a balanced class representation, without the disadvantage of discarding a portion of the dataset.

5 Experimental setup

This section describes the processes utilised and the choices made for the experimental part of our study. We start by describing the model and training parameters and we explain the three phases of our investigations. First, we report the methods used on the full dataset where we try several approaches addressing the class imbalance problem. Next, we characterise the classes that have been removed or grouped together to reduce confusion risks for the investigations on the clean dataset. In a third part, we indicate the methods used to compare the performances obtained by ADELE. The different training sets used for our experiments are summarised in Table 1. Last we present the evaluation metrics commonly employed to evaluate classification performances on imbalanced datasets.

The code has been made available online[4].

5.1 CNN architecture and training specifications

This section reviews the practical aspects of our experimentations. We describe the model architecture and the general training parameters.

5.1.1 Network architecture

We choose ResNet-50 pre-trained on the dataset ImageNet as implemented in the *Pytorch*[5] framework (Paszke et al., 2017) for all experiments. ResNet is commonly used in multi-class classification problems (Liu et al., 2019), it reaches high performances on benchmarking datasets (Picterra, 2017) and it offers a good balance between top accuracy and model complexity.

The ResNet-50 model architecture is adopted for all our experiments with several modifications. Since we opted for a model pre-trained on ImageNet, its classes are not specific for LCLU classification. We update the weights of all parameters at each epoch to enforce the model to be specific to our problem.

The input layer is modified in order to accept our data with five channels instead of the three RGB bands. The pre-trained weights for the input layer for the RGB bands are given to the corresponding RGB channels in our data. The weights for the near-infrared and DEM channels are initialised with the values from the red channel. We experimentally observed that a 50% dropout layer added between the average pooling layer and the fully connected layer reduced overfitting and improved classification performances. The number of outputs in the fully connected layer is modified from 1'000 to 28 or 21 categories to fit the number of land use classes present in the dataset.

[4]https://github.com/vzermatt/ClassImbalance
[5]https://pytorch.org/

5.1.2 Training parameters

Several configurations are tested to establish optimal training parameters for the baseline model and for the subsequent experiments.

The model is trained on a single GPU (GeForce GTX TITAN X 12GB). For all experiments, we use a batch size of 128, that is the largest batch size possible for our machine. The best initial learning rate is experimentally determined to $1e^{-5}$, with a decay by a factor of 0.1 every 40 epochs. The model is usually trained for 100 epochs, the implementation of early stopping allowed sometimes but not always an improvement in performances. When in use, early stopping is mentioned with the number of epochs without improvement.

Empirically, we find out that Adam worked better than SGD as an optimizer. We trained the baseline model with the softmax cross entropy loss and Adam with the recommended default parameters from Kingma and Ba (2015) (coefficients that control the exponential decay of the running averages of gradient and the squared gradient: $\beta_1 = 0.9, \beta_2 = 0.999$, term added to the denominator to improve numerical stability $\epsilon = 1e^{-8}$) and a weight decay of 0.1. We opted for batch normalization as it performed slightly better in trials than instance normalisation.

5.1.3 Data augmentation

Data augmentation is performed during the training and validation phases. Since aerial data have a top-down perspective, rotations or random flips do not affect the ground truth and both methods are used, each with 50% probability. However, lateral movements or random crops of the image could lead to a relative displacement of the image centre, to which the label is strongly dependant and thus they are not employed. Finally, a random colour jittering on the RGB channels is applied, modifying the brightness, contrast, hue and saturation by a random amount of up to 30% each and to 5% for the hue setting. For all training, validation and test phases, data are rescaled channel-wise to 2-98% quantiles, except for the DEM that is resized between 475 and 3242, being the minimum and the maximum altitude present in our study area.

5.2 Experiments on all classes ("full dataset")

Our first experiment aims to determine if the unbalanced distribution of samples among classes is damaging the classification results. To this end, we evaluate several techniques targeting the class imbalanced distribution against a baseline model where no special measures are applied.

The baseline model is trained on the full dataset composed by images from classes with more than 100 representatives, leading to a selection of the 28 among 46 categories. For the model training process, the data are split into 60% training, 10% validation and 30% test sets. Stratified random sampling is used to ensure similar representation of classes between the three sets. During the experiments, the softmax cross entropy loss is employed if not otherwise mentioned.

5.2.1 Sampling methods

For the sampling methods, we implement two versions of undersampling. We start by a total correction of the imbalance as recommended by Buda et al. (2018) and we remove samples from the full dataset until all classes have 100 examples. The images from the undersampling dataset (`und-100`) are selected in the full training and validation set with exactly 60 samples per class for training and 30 for validation for all the classes. The test phase is performed on the full test set.

As this first selection drastically reduced the number of training samples, we also performed a second undersampling experiment (`und-1000`) with a partial reduction of the number of samples per class where a maximum of 1'000 samples per categories are retained. For `und-1000`, the training and validation samples originate from training and validation sets from the full dataset, and a maximum of 600 samples per class for training and 100 for the validation phase are selected. The test is performed on the full test set.

As mentioned in section 4.3.1, we do not implement the oversampling method since its effects are similar to inverse frequency re-weighting, but as samples are replicated, the training is much longer due to the increased size of the dataset.

For the two-phases training experiment, we start the training with the fully trained baseline model and we adapt its parameters for 100 more epochs on the partially balanced (`und-1000`) dataset.

5.2.2 Methods involving specific loss functions

We start with the re-weighting methods with inverse frequency weighting (`inv_freq`) and inverse square root frequency weighting (`sq_inv_freq`). The weights are computed on the number of samples in the train set and are normalised so that the mean of all weights equals to 1.

The class balanced loss is tested with β values in the range $\beta \in [0.99, 0.999, 0.9999]$ with the softmax cross entropy loss (`sCBL`). The CBL is then evaluated with the focal loss (`fCBL`) with the best β values from the `sCBL` and varying values for $\gamma \in [0, 0.5, 1, 2, 3]$.

The equalization loss (`EQL`) is tested with several combination of hyper-parameters with a θ (ignore probability) ranging from 0.5 to 0.95, and the threshold for the minority class $T_\lambda \in [300, 600, 900, 1500]$. For all methods involving loss functions, the full training, testing and validation sets are used.

5.3 Experiment on a reduced set of classes ("clean dataset")

He and Garcia (2009) reported that dataset complexity is a primary factor of classification deterioration. It includes issues such as the overlap between classes or the high variability within class with the presence of sub-categories that makes these classes more difficult to learn for the CNN. The main idea of the second part of our investigations is to apply manual data decontamination methods to observe if we can increase accuracy over the rare classes through methods addressing the class imbalance.

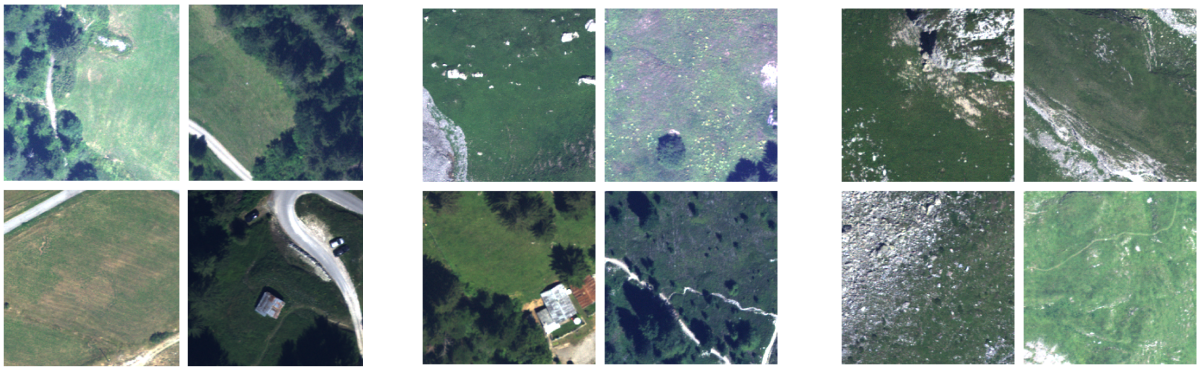
Dataset cleaning

The dataset cleaning or decontamination involves relabelling or removal of some selected examples (Buda et al., 2018). Its goal is to identify inappropriate labels to produce a more robust dataset. Through an in-depth analysis of the type of samples present in each class and their relation to other classes, the cleaning intents to detect class overlapping, class label noise, mistakes and unclear borders between classes (Krawczyk, 2016).

Even though label noise can occur even in the case of expert annotators (Algan & Ulusoy, 2021), we estimate it very low in the case of the OFS procedure for the label production. However, the addition of auxiliary data for labelling and the time gap between the label production and the survey may introduce some incoherent or unpractical labels as mentioned in section 2.4.

Dedicated methods have been developed for large scale dataset cleaning (Algan & Ulusoy, 2021), but this work focusses on manual category grouping or removal. To identify classes damaging the classification results, we proceed to a visual dataset examination and to a misclassification analysis from the results of the first experiment.

As illustrated in section 2.4, some classes present a very low visual inter-class differences that makes it almost impossible to separate them without auxiliary data. We decide to merge some of them into a more meaningful category. The formation of the two new super-classes formed from seven classes exhibiting ambiguous features are explained below.



(a) Alpine meadows in general (b) Alpine pastures in general (c) Alpine sheep grazing pastures

Figure 12: Three types of alpine pastures

- Alpine meadows, alpine sheep grazing pastures and alpine pastures in general comprehend grassland at high altitude that are located away from permanent residential areas (OFS, 2016, 2017). These categories show a high visual similarity as shown in Figure 12, but some differences in their usage appear with their respective description: Alpine meadows are used for haymaking, occasional roads are present, which is not the case for the other two classes. Both alpine sheep grazing pastures and alpine pastures in general are used for seasonal cattle grazing. The sheep grazing areas tend to be more rocky pastures with complicated access due to their remote location. The alpine pastures in general shows more favourable conditions. During the photo-interpretation, the expert refers to the label given during the previous survey, if no land use change is visible.

(a) Farm pastures in general

(b) Semi-natural grassland in general

Figure 13: Semi-natural grasslands and farm pastures

- Farm pastures shown in Figure 13a and semi-natural grasslands shown in Figure 13b consist in grasslands located in low density areas with permanent settlements (OFS, 2016, 2017). Their main difference lays in the fact that grasslands are mowed at least once a year, whereas the farm pastures are used for cattle grazing. The delimitation of these local pastures requires the consultation of the cadaster from the Federal Office for Agriculture (FOAG) for agricultural production. Depending of the seasonal vegetation changes, the characteristics of these pastures are more or less visible. For instance, variations in vegetation density and colours appear between the photographs from the early and late summer and are visible in Figure 13a.

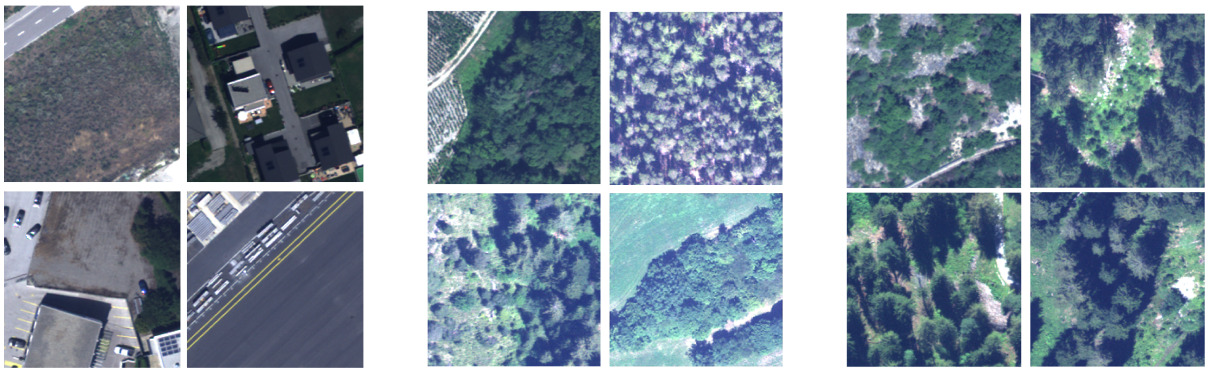
In addition to these two new super-classes, four categories are entirely removed from the training set.

- The category **construction sites** is severely damaged by the time difference between the label production in 2013 and the photography survey in 2020.
  This class comprehends surface where temporary construction work is in progress and it may include deposit of excavation material, machinery and equipments, cleared surfaces, etc (OFS, 2016). In our dataset, the majority of samples with this label do not any more exhibit signs of construction work as shown in Figure 14b. The 7 years period allowed the termination and the disappearance of most of the constructions sites. We also find some evidences of new construction sites on some samples with other land use labels. Due to the confusing samples present in this class, the construction sites labels are no longer valid for our dataset and are removed.

(a) (b)

Figure 14: (a) Typical construction sites (b) Samples labelled as construction sites but with probable new land use

- The **unexploited urban areas** shown in Figure 15a include brownfields and unused buildings located within the residential and commercial perimeters which have not yet found a new use (OFS, 2016). A wide variety of land cover may be present: uncultivated agricultural land, abandoned industrial buildings, disaffected roads, old touristic sites and buildings, ruins of houses or public buildings. The results on the full dataset experiment showed that samples from this class tend to be misclassified into many different categories due to the high intra-class variations. Additionnaly, due to the 7 year gap between labels and our data collection, some of the unexploited areas seemed to have a new affectation for examples the construction of new habitations on unused agricultural land (see Figure 15a). We decide to withdraw this class for the data cleaning experiment.
- Similarly, the category **unspecified buildings and surroundings** encompasses many different building types within and out of the urban areas such as hotels, restaurants or local shops (OFS, 2016). The identification of the correct class requires information from both the building register (RegBL) and the agricultural exploitations register (NOGA), since some samples tend to be misclassified as residential area or agricultural buildings during our full dataset experiment. As a result, the samples of this class are removed for the clean dataset experiment.
- The **lumbering areas** present a strong similarity with the forests class (see Figure 15b and 15c). The former includes forest stands that have been cut for silvicultural activities reducing the tree coverage by at least 60% (OFS, 2016). The forests category requires by definition a tree coverage superior to 60%. Some lumbering sites present a clear cut, but other places are sparsely logged. Without auxiliary data it is difficult to differentiate lumbering areas from true forest stands. Consequently this class is removed.



(a) Unexploited urban areas (b) Forests (c) Lumbering area

Figure 15: Illustration of classes removed from the clean dataset. (a) Illustration on the left of typical unexploited urban areas, and on the right unexploited urban area with probable new land use. (b) and (c): Illustration of the similarity between the categories for forests and lumbering areas

Experiments on the clean dataset

The clean datasets are based on the same distribution of samples as in the full, the `und-100` and `und-1000` datasets, except for the four classes that are completely removed and the five classes that have been fused into two super-classes. The new total number of categories equals 21 classes instead of 28. The number of samples for each class in each dataset is mentioned in the Appendix (Figure 5).

The methods addressing the class imbalance applied during the experiments on the full dataset are replicated on the clean dataset. The best hyper-parameters from the first experiment are employed for the `sCBL,fCBL` and the `EQL`. When early stopping is used, both the shortened models and the model trained on 100 epochs are tested on the clean test set and the best of them is selected.

5.4 Comparisons with ADELE predictions

The last part of our investigations consists in comparing the predictions of our best performing models with the results from the ADELE methods on the same data. Adrian F. Meyer is member of the ADELE project at the FHNW and provided us predictions data produced by ADELE model as described in section 3.2.2 for comparison with our data. The predictions were produced in 2018/early 2019 from the data of the 2013/18 survey and they cover approximately one quarter of our study area as shown in Figure 16.

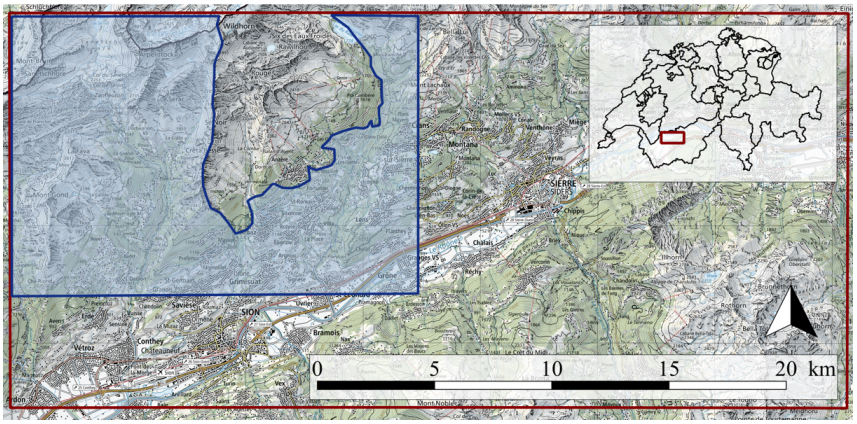


Figure 16: Location of the test area with ADELE predictions (in blue) within our study area (in red)

The data contain the predictions for a total of 6634 unique samples, the corresponding ground truth labels, the RELI identifiers and the values of random forest confidence for each prediction. On the test area, they reach 86.3% overall accuracy, slightly higher than the average in their report. More details are provided in the results section.

The predicted samples come from 36 ground truth classes. For a fair evaluation of ADELE predictions, we only selected samples where both the ground truth and the predicted labels corresponded to one of the 28 or 21 classes class present in our model training sets and we removed samples issues from categories not classified by our models. 14 samples in total belonging to 8 of these classes are removed from the test. As a result, the FHNW test set (`FHNW_test`) used for comparison with our model contains 6620 samples from the same 28 classes as our full dataset.

Our experimentation compares ADELE predictions with the model that performed best on the full dataset, and then with the model that performed the best on the clean dataset. For a fair comparison of our model performances to those of ADELE, we retrained our best model on a new training set and validation set is required where there are no overlaps between our training points and the FHNW test set.

We start by excluding the samples present in the FHWN test set from the full dataset. For each class, 10% of the remaining data are used for the validation and 90% for training. A clean training and validation sets are produced similarly to the methods presented for the clean dataset. As this procedure is different from the one for the full dataset presented earlier, some variations in performances may be introduced. Table 1 summarises the produced datasets and the Table 6 in the Appendix lists all the experiments.

	Full dataset experiment			**Clean dataset experiment**			**Comparison with ADELE**	
Datasets name	full	und_100	und_1000	clean	clean_100	clean_1000	fhnw	clean fhnw
Number of classes	28	28	28	21	21	21	28	21
Number of samples per class	>100	100	100-1000	>100	100	100-1000	-	-
Test set	full test set			clean test set			fhnw test set	clean fhnw test

Table 1: Summary of the dataset used for the different experiments

5.5 Evaluation metrics for multi-class classification problem with imbalance

The **confusion matrix** is used to analyse classification errors between different classes by placing the true label and the prediction for each sample in a table, as illustrated in Figure 17. The green elements on the main diagonal are correctly predicted with true positives (TP) and true negatives (TN), whereas all other cells present the wrong predictions with false positives (FP) and false negatives (FN). Several metrics based on it are presented in the equations 12 and 11 below. These measures all lie in the range $[0, 1]$ with 1 being perfect predictions.

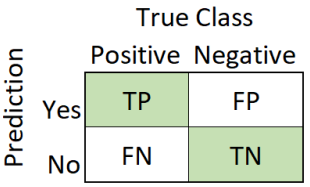


Figure 17: Confusion matrix

The **precision** is calculated by dividing the correctly classified samples by the total number of samples obtaining that predicted class. It is also called error of commission as it indicates whether the samples classified into a class actually belong to this class. It tells if a classifier is over-predicting a given class producing to many false positives. In our experiment we compute the average precision per group of rare, common and frequent classes and the average precision of all classes with all categories receiving the same weight.

The **recall** or producer accuracy is computed by dividing the correctly classified samples by the total number of samples predicted with this ground truth class. It measures the completeness of the classifier predictions, assessing whether the classifier is under-predicting a given class with many undetected samples. It is related to the error of omission, since it informs whether the classifier manages to find all samples belonging to a class.

$$Precision = \frac{TP}{TP + FP} \qquad Recall = \frac{TP}{TP + FN} \tag{11}$$

$$F-1\ score = 2 * \frac{Recall * Precision}{Recall + Precision} \qquad Cohen's\ Kappa(\kappa) = \frac{P_{obs} - P_{est}}{1 - P_{est}} \tag{12}$$

Precision and recall are best used together as the classifier is balancing between both: Increasing recall usually leads to reducing the precision and vice-versa. Specifically, the **F1-score** combines these two metrics through a geometric mean and give a measure of effectiveness of the classifier. We compute the F1-score for each class. For the entire classifier we calculate the F1-score with the means of the individual F1-score for all samples and we do not take label imbalance into account.

A frequently used metric to asses a classifier performance is the **overall accuracy**. It is formed by dividing the sum of the correctly classified sample points by the total number of sample points. However, this measure is well known to provide inadequate indications in the context of class imbalance (Buda et al., 2018; Cheng et al., 2017; He et al., 2016). When the test set is imbalanced, the overall accuracy favours classes with a large number of samples. The result can be misleading, as the performances for the categories with few samples are barely taken into consideration. The overall accuracy must be interpreted with care. Similar reasoning and limitations apply for the overall error rate. An alternative to the overall accuracy is to adopt the **average precision** as it is independent of the number of samples per class.

Cohen's Kappa compares the predicted distribution with a random distribution of samples among classes, respecting the number of instance per categories. It is a robust measure for imbalanced datasets as it takes into account the agreement occurring by chance, which frequently arrives when a few majority classes are present.

6 Results and analysis

6.1 Results for the experiment on the full dataset

In this section, we evaluate the impacts of the sampling methods and the specific loss functions targeting the imbalanced distribution on the full dataset with 28 classes. We first compare the general effects of the tested methods before focussing on the performance per class and studying the confusion between categories.

6.1.1 Effects of the methods targeting the class imbalance

	K	**OA**	**AP**	**AR**	**F1**	**Pr**	**Pc**	**Pf**
`Baseline`	**78,2%**	**82,5%**	**63,1%**	52,0%	**54,6%**	**58,3%**	62,0%	88,0%
`und_100`	54,9%	61,2%	40,9%	54,0%	41,7%	29,2%	54,0%	86,0%
`und_1000`	67,7%	72,9%	48,4%	55,7%	48,9%	38,6%	57,8%	87,8%
`two-phases`	70,8%	75,7%	53,3%	**56,1%**	52,6%	45,4%	58,2%	88,2%
`inv_freq`	73,4%	78,1%	54,9%	55,8%	53,6%	47,8%	56,5%	**88,3%**
`sqrt_inv_freq`	76,4%	81,0%	60,6%	50,4%	51,8%	55,2%	61,2%	87,0%
`EQL`	77,3%	81,8%	58,6%	48,0%	49,2%	52,4%	61,2%	86,8%
`sCBL`	77,3%	81,8%	61,1%	50,4%	53,1%	56,0%	60,5%	86,8%
`fCBL`	77,3%	81,8%	62,3%	50,7%	52,4%	57,3%	**62,8%**	86,2%

Table 2: General results for different methods targeting the imbalanced distribution on the full dataset.

Table 2 illustrates the results obtained by the different training methods on the full test set. On average the frequent classes obtain the best precision with 88.3% precision whereas for the rare classes it is reduced to 58.3%. The common classes scores slightly better than the rare classes with a maximum of 62.8% of precision

The baseline model is globally the best classifier with the best κ, overall accuracy, average precision, F1-score and rare classes precision. Nonetheless, its average recall performed moderately well in comparison to other models, meaning that it tends to under-predict several classes. Slightly better predictions are obtained by the `fCBL` and the `inv_freq` for common and frequent categories. This result corroborates those obtained by the company Picterra in their study (Picterra, 2017), where no improvement is observed when applying re-weighting methods on the original dataset.

Both undersampling methods `und_100` and `und_1000` perform poorly. The reduction in the number of training data seems to alter results especially for the rare classes. The common classes experience a modest decrease in accuracy that is expected when decreasing the training dataset size. Nonetheless, the frequent classes precision is barely affected by the reduction in the training set size compared to the baseline model. These classes including forests or alpine pastures show a simple and uniform pattern repeated in all samples and they are well learned by the model even with a small number of representatives. Reversely, the significant increase in rare classes precision between the `und_100` and `und_1000` indicates that the model is highly sensitive to the number of training samples for these classes, as adding more data significantly ameliorates the predictions results. Surprisingly the `und_1000` model has a much poorer rare class prediction compared to the baseline model, even if they have the exact same samples for training on the rare classes.

The `two-phases` method obtains the best average recall but its average precision is among the lowest meaning that it trades a bit of precision for a higher recall, contrarily to other models. The F1-score performs within 2% from the baseline, but its rare classes precision is significantly lower than other models. As the `two-phases` model weights are initialised on the baseline parameters for the first phase of training, we can deduce that the second part of the training on the `und_1000` dataset damages the classifier instead of improving it.

The square root inverse frequency re-weighting (`sq_inv_freq`) method produces acceptable results on all metrics, even though they are always lower than the baseline model. As anticipated through the literature review (Mahajan et al., 2018), it performs better than its counter part with inverse frequency weights (`inv_freq`) regarding most of the metrics (K, OA, AA, Pr and Pc). The latter obtains the best precision on the frequent classes.

The results for the methods involving the equalisation loss (`EQL`) and the softmax (`sCBL`) and focal class balanced losses (`fCBL`) are produced by the combination of the best hyper-parameters. The results for the different tests are present in the Appendix (`EQL` see Table 7, `sCBL` and `fCBL` see Table 8).

For the `EQL` training, the frequent class precision remains relatively stable and seems unaffected by the different training parameters. The common classes precision is higher with larger threshold value for minority classes and high value for the probability to ignore the negative gradient. Its reactions to the change in number of training epochs is not constant.

For the rare classes precision, the pattern seems to be highly dependant on the number of epochs, as the the shorter training session produced the highest precision on rare classes. We selected an ignore probability θ of 75% and a threshold for minority class T_λ of 300 samples since it obtains the best K, OA, F1-score and it is the model performing the closest to the baseline. $T_\lambda = 300$ is the threshold for the minority class in the test set, which correspond to a total number of 1000 samples in the full dataset, which matches our definition for the rare categories. The chosen `EQL` reaches on average slightly less good results than the baseline but suffers from an important drop of the precision for the rare categories. The model with $\theta = 0.50$ is not selected since the frequent classes performances seems degraded compared to $\theta = 0.75$.

At the beginning of the training the equalisation loss concentrates on better learning the images from the minority categories defined by a frequency below the threshold T_λ. After a number of iterations, these samples are properly learned, and the EQL recognises them with a high precision. By continuing the training, the network will adapt to difficult samples from other classes and

it may damage the weights adapted for rare samples. This principle probably explains why the longer training alters the rare classes precision.

For the `sCBL` models, the experiment with $\beta = 0.99$ receives the best K, OA, AP, F1-score and rare class accuracy with early stopping after 30 epochs without improvement. Even though this model performs less accurately on the most frequent classes, the early stopping improves the precision for both the common and the rare classes compared to a model trained until 100 epochs.

The value $\beta = 0.99$ means that the model weights give more importance for very rare samples with less than about 200 occurrences (see Figure 11) and all other samples receive an equal weight independent from their frequency, meaning that the category importance is unbalanced for the others. By studying different γ values for the focal loss with $\beta = 0.99$, a slightly better rare class precision is obtained with a $\gamma = 1$. Both `sCBL` and `fCBL` show predictions within 2% from the baseline and even slightly better on the common classes precision.

To sum up, we observe that the methods targeting the class imbalance are ineffective on the full dataset. Even though the precision on the frequent categories for some methods are similar to the one of the baseline model, an important decrease in performances occurs for the rare classes. The next part aims to explain why these classes are difficult to learn.

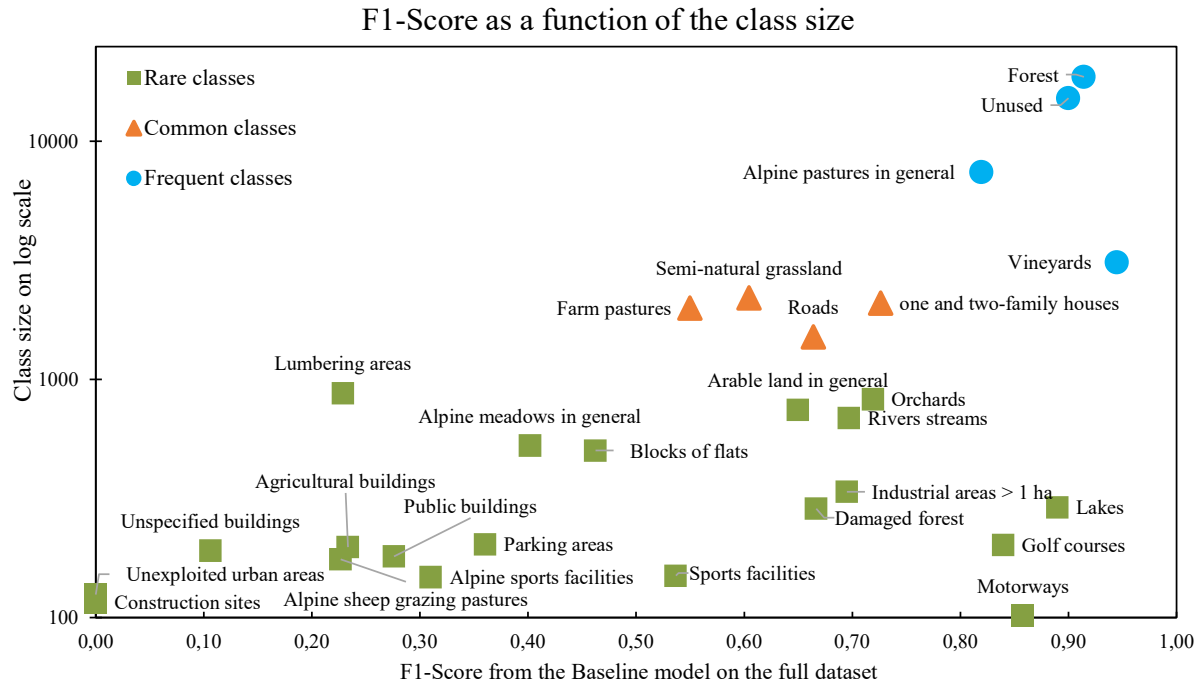


Figure 18: F1-score from the baseline model as a function of the number of samples per class in the full dataset (on log scale) with short labels

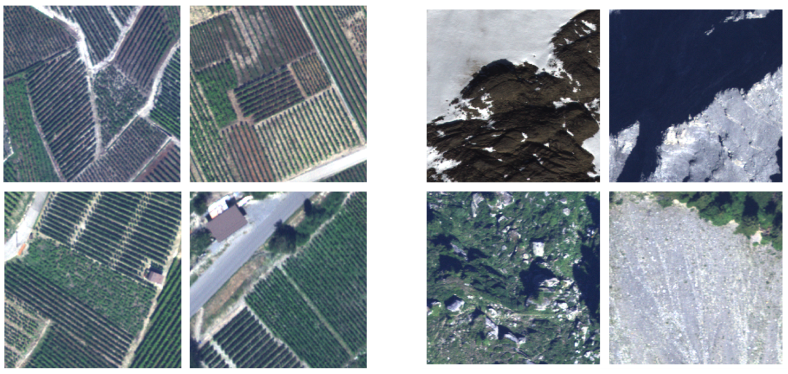
6.1.2 Results per class

This section analyses the class difficulty in relation to the number of samples. Table 30 in the Appendix gives the precision, recall and F1-score for class and method combination. Most

categories receive an evaluation in the same range of values with all models, with sometimes differences in the balance between precision and recall.

For a clearer overview on the class difficulty, the F1-score from the baseline is plotted as a function of the class size in Figure 18. Frequent classes obtain very reliable results with F1-scores around 0.8, common classes receive lower F1-scores between 0.55 - 0.75 and the scores for rare classes occupy the entire range of values.

The categories with abundant data including forests, unused areas, alpine pastures and vineyards are easily learned by the model. The vineyards illustrated in Figure 19a obtain the global best classification performances. Its repetitive lines of plantation are a very distinctive element for the convolutional layers.



(a) Vineyards (b) Unused areas

Figure 19: Examples of frequent categories

Unused alpine areas have a rather diversified appearance with the presence of landslides, bare rocks, glaciers and short vegetation as depicted in Figure 19b. Some samples present strong contrasts in lightning due to the irregular topography. However the unused lands are constantly located among the highest altitude and the DEM channel probably gives key indications of their category to the classifier. Forests, vineyards and alpine pastures exhibit a characteristic pattern of vegetation densities that are enhanced through the infrared channel and it is probably a significant source of information for their recognition.

The common categories are represented by less samples than the frequent ones. This reason alone is not sufficient to explain the drop in accuracy since the model achieves better classification accuracy on several rare classes. The common categories are more complex than the frequent classes, as they comprehend a larger variety of features on each image. For instance, residential areas with one or two-family houses exhibit buildings, roads and vegetations areas. Similarly, farm pastures and semi-natural grass land indicates the presence of grass fields, but roads or buildings frequently appear on their reference surface. Moreover these two classes show a high visual similarity as exposed in Figure 13 and tend to be confounded with each other.

The rare classes with typical and distinctive features such as golf courses, lakes and motorways shown in Figure 20 obtain a very good F1-score in the range of 0.8. As they present a recognisable pattern identifiable on most images, they reach similar level of precision as the frequent classes, despite their small number of representatives and the class imbalance factor.

(a) Motorways (b) Golf courses (c) Lakes

Figure 20: Examples of rare categories with high recognition accuracy from the network

Poor outcomes concern some rare categories that exhibit features similar to those of other classes. For instance several types of buildings (agricultural, public, unspecified buildings) undergo considerable misclassification rates due to their mutual confusion. Another drop in classification accuracy is related to land use groups with several specific objects. For example, the sport facilities label matches places as varied as tennis or football fields, horse riding fields or swimming pools. The high variability of the images in addition to the insufficient number of samples makes it very difficult for the model to recognise the sub-categories.

Concerning the two classes with the worst F1-score, unexploited urban areas and construction sites, the time gap between labels and images (7 years) renders the labels unusable for the classifier, as mentioned in section 5.3 and they are removed for the next experiment.

As a result, we observe that the classification does not only depend on the class frequency but also on the complexity of the class itself. The model can produce accurate labels for classes even with a small number of samples if this class shows distinctive visual characteristics that are repeated in many samples of the class and absent from other categories. Large error rates originate from classes who are visually too similar to another, or classes with high variability between samples.

6.1.3 Class interactions during the training

This part treats more in depth the problematic of class similarity through the analysis of the class interactions in the predictions. We base our analysis on the confusion matrix for the baseline model in Figure 23.

Classes similarity

As mentioned is section 5.3, an important similarity exists between the different types of grass fields and it appears clearly in the confusion matrix. For example, the alpine meadows are frequently misclassified as semi-natural grasslands, farm pastures and alpine pastures in general. These classes occur at various altitude with different degrees of ground slope. The presence of the DEM might partially reduce the confusion between the low altitude grass fields (arable lands in general, farm pastures and semi-natural grasslands) with the high ones (alpine meadows, alpine pastures in general, alpine sheep grazing pastures).

Numerous confusion appear with the sub-categories of the forested areas. The damaged forests samples usually present apparent dead trees laying on the ground (see Figure 21), but in some cases they have been withdrawn. Consequently this class obtains a high precision (75%) meaning that the model is able to properly recognise samples with these typical dead trees on the ground, but a lower recall (60%) since it forgets the samples without this attribute. Similarly lumbering areas present occasionally a cleared perimeter where trees have beencut but sometimes the area is sparsely logged. In both cases, the two forest sub-categories are difficult to observe on the pictures leading to an elevated rate of misclassification.

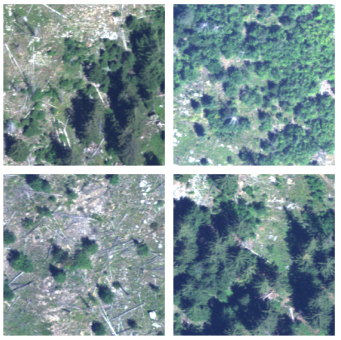


Figure 21: Damaged forest

Lack of auxiliary data

Buildings types as residential area, agricultural, public and unspecified buildings undergo considerable misclassification rates especially in low density areas. Their appearance from an aerial point of view tends to be similar with the presence of buildings surrounded by vegetation as illustrated in Figure 22. The absence of auxiliary data makes their distinctions very difficult. Interestingly the industrial and commercial buildings are less attained by this trend since the weak vegetation cover, the size of the buildings and the type of material render their appearance more distinguishable.

(a) Unspecified buildings and surroundings

(b) Residential area: One and two-families houses

(c) Agricultural buildings and surroundings

Figure 22: Examples of building categories in low density area

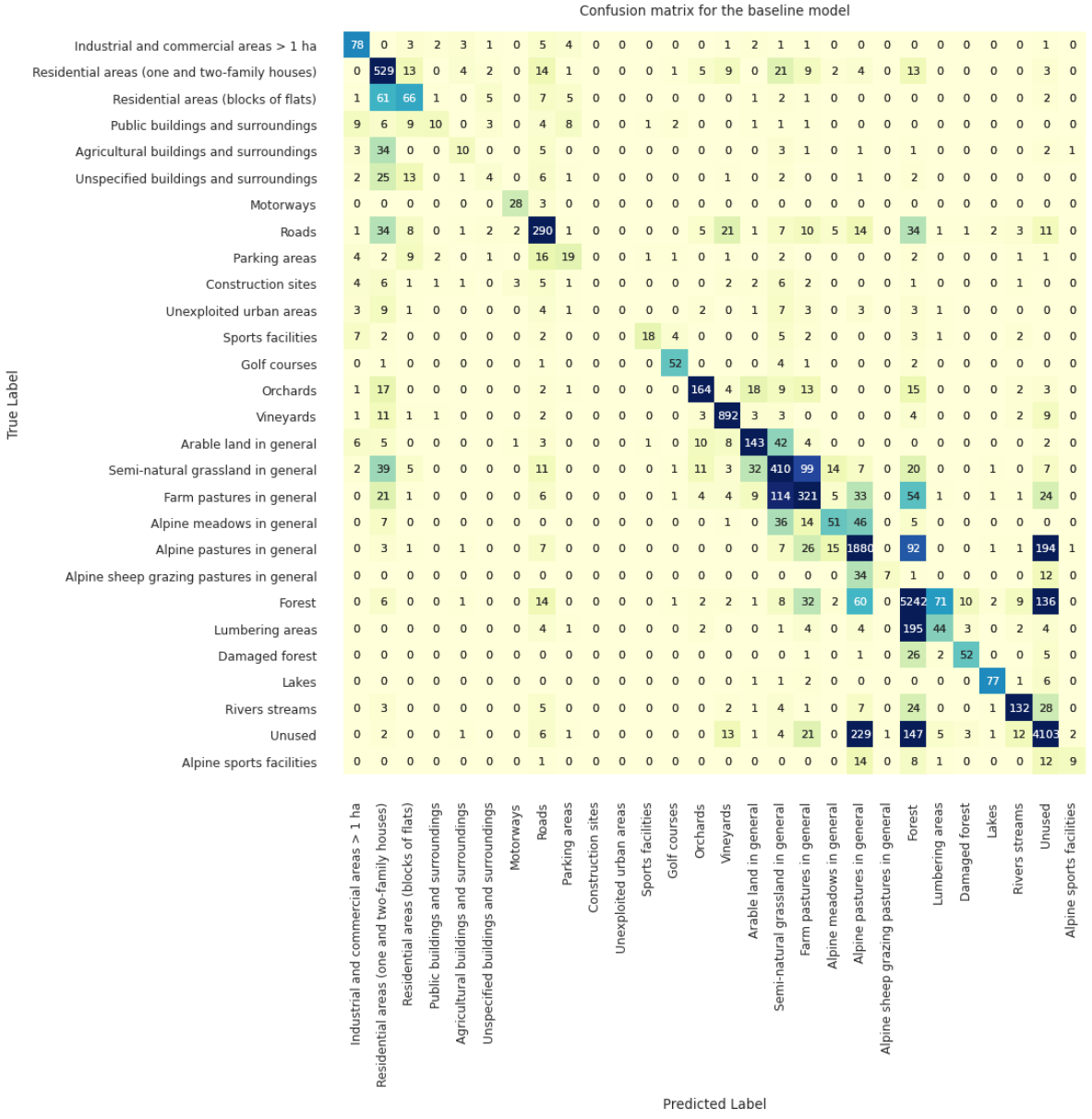
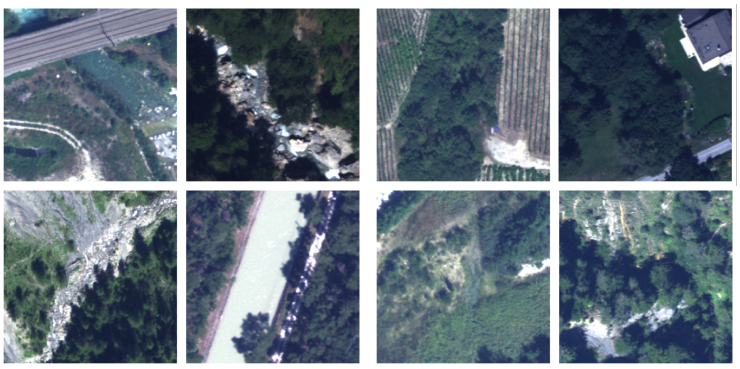


Figure 23: Confusion matrix from the baseline model on the full dataset

Some surprising errors are encountered. For example, residential areas are misclassified as semi-naturals grass land or forest areas. It can be explained by the fact that land use classes for buildings also include their direct surroundings (OFS, 2017), such as the grasslands and trees next to a residential areas or the parking lots deserving commercial and industrial areas. These are typical examples of overlap between classes and they might introduce some confounding effects in the network.



(a) (b)

Figure 24: Rivers and streams may have very heterogeneous appearance (a). The vegetation may entirely hide them, auxiliary information might be required to detect them (b).

Interpretation of the central features

As the object located at the centre of the image usually determines the land use label, this leads to some erroneous predictions. This particularly concerns sites located at the border between two separate land use classes. For instance, the roads category is highly affected by this fact, but also sites located at the limit between the forests and alpine pastures, or residential areas located next to vineyards etc.

Some rivers and streams have a high precision but a low recall. Rivers may have very different appearance (see Figure 24a). The classifier reaches a high precision and the model is able to identify samples correctly. The low recall indicates that it remains difficult to find all samples since we observe in Figure 24 that forgotten samples are hidden by the surroundings.

To summarise, the methods targeting the imbalanced distribution on the full dataset do not allow to improve the classification accuracy for the rare classes compared to the baseline model. However, this experiment highlights that the category complexity may be the driving source of miss classification for our dataset. The study of He and Garcia (2009) corroborates the idea that the dataset complexity is a determining factor of performances deterioration in the case of imbalanced datasets.

The class complexity includes issues such as the overlap between classes or the high variability within one class with the presence of sub-categories. Moreover since the labels were produced with the help of auxiliary data, the identification of several classes might be unpracticable for the model. Additionally, the time gap between the photography survey and the label production introduce some erroneous labels and the model unable to recognise the construction site class. We also noticed several mistakes when more than one land use class is present on the reference surface: the label depends on the land use of the central point which may be different from the land use covering the majority of the image.

For categories with numerous samples, the CNN manages to understand their concepts and diversity, but the complex classes with less representatives have difficulties to be recognised. We presume that the class complexity alters the general classification results. Consequently it may also damage the effect of the methods targeting the class imbalance. As a result we decided to reproduce the experiments on a dataset with a reduced class complexity, presented in the clean dataset experiment.

6.2 Results on the clean dataset

In this section, we evaluate the impact of the grouping and the removal of several ambiguous classes on the methods targeting the imbalanced distribution. We first compare the general effects of the tested methods on the clean dataset before focussing on the performances per class of the best method.

6.2.1 Effects of the methods targeting the class imbalance on the clean dataset

	K	OA	AP	AR	F1	Pr	Pc	Pf
Baseline	**82,1%**	**85,9%**	69,9%	60,6%	62,8%	64,1%	**72,3%**	88,5%
`und_100`	60,0%	66,3%	43,5%	62,3%	46,9%	29,3%	53,0%	86,2%
`und_1000`	75,3%	80,0%	62,7%	65,1%	60,1%	56,3%	59,7%	87,5%
`two-phases`	79,3%	83,3%	64,2%	**65,9%**	62,8%	57,1%	63,3%	89,5%
`inv_freq`	79,1%	83,2%	66,9%	60,9%	60,8%	51,2%	69,7%	88,0%
`sqrt_inv_freq`	80,8%	84,7%	66,9%	63,9%	63,7%	59,7%	72,0%	88,2%
`EQL`	80,5%	84,5%	60,5%	58,8%	57,3%	50,6%	67,7%	**89,8%**
`sCBL`	81,8%	85,6%	70,2%	63,4%	64,8%	64,8%	71,7%	88,0%
`fCBL`	82,0%	85,8%	**72,6%**	62,4%	**65,2%**	**68,1%**	71,7%	89,0%

K=Kappa, F1=F1-Score, OA=overall accuracy, AP= average precision, AR= average recall, Pr (resp. Pc, Pf) = average precision for the rare (resp. common, frequent) classes. The best value for each metric is in bold.

Table 3: General results for the different methods addressing class imbalance on the clean dataset (21 classes). All of them have been trained for 100 epochs.

Table 3 presents the results of the baseline model and the methods targeting the imbalanced distribution on the clean dataset. As expected, all models show an notable improvement in performances compared to the full dataset experiment. The frequent classes precision increases by 1.5%. The rare and categories obtain both an increase in average precision by about 10%, but removal or the fusion of several of their classes makes these metrics not properly comparable.

The best model is the one trained with the focal class balanced loss (`fCBL`), as it outperforms all other models regarding the rare categories precision, F1-score and the average precision. The baseline model produces the best performances regarding the κ value, the overall accuracy and the common categories precision. However the `fCBL` is always within 1 percentage point behind.

Concerning the other methods, the `sCBL` also beats the baseline on several metrics including the rare classes precision (AP, AR, F1-score, Pr) and corroborates the ameliorating effect exerted by the class balanced loss for the identification of the rare classes. All other models receive lower rare class precision.

Similarly to the first experiment, the use of undersampling in `und_100` and `und_1000` provides the least satisfying classification results. The equalisation loss and the inverse frequency re-weighting

overtake the focal class balanced loss for respectively the average recall and the frequent classes precision. Nevertheless they remain globally less performant. The two-phases learning model obtains again the best recall but balances it with the average precision.

Consequently, we observe that the focal class balanced loss can effectively address the class imbalance on our dataset after the removal of the most confusing classes. Reducing the class complexity makes the dataset more reliable and the model learns more coherent classes representation.

Label	Baseline			fCBL			Difference			Test set size
Rare classes	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	
Motorways	89,0%	88,0%	90,0%	82,0%	83,0%	81,0%	-7,0%	-5,0%	-9,0%	31
Alpine sports facilities	18,4%	56,0%	11,0%	24,2%	50,0%	16,0%	5,9%	-6,0%	5,0%	45
Sports facilities	34,4%	61,0%	24,0%	51,6%	85,0%	37,0%	17,1%	24,0%	13,0%	46
Public buildings and surroundings	22,4%	44,0%	15,0%	25,4%	43,0%	18,0%	3,0%	-1,0%	3,0%	55
Agricultural buildings and surroundings	0,0%	0,0%	0,0%	16,7%	50,0%	10,0%	16,7%	50,0%	10,0%	61
Golf courses	72,4%	63,0%	85,0%	83,0%	81,0%	85,0%	10,6%	18,0%	0,0%	61
Parking areas	28,6%	45,0%	21,0%	39,1%	46,0%	34,0%	10,5%	1,0%	13,0%	62
Damaged forest	66,9%	83,0%	56,0%	63,1%	74,0%	55,0%	-3,8%	-9,0%	-1,0%	87
Lakes	91,0%	93,0%	89,0%	88,8%	93,0%	85,0%	-2,1%	0,0%	-4,0%	88
Industrial and commercial areas > 1 ha	71,5%	66,0%	78,0%	68,6%	58,0%	84,0%	-2,9%	-8,0%	6,0%	101
Residential areas (blocks of flats)	52,5%	54,0%	51,0%	53,9%	52,0%	56,0%	1,5%	-2,0%	5,0%	152
Rivers streams	67,6%	83,0%	57,0%	68,6%	91,0%	55,0%	1,0%	8,0%	-2,0%	208
Orchards	72,9%	76,0%	70,0%	72,0%	73,0%	71,0%	-0,9%	-3,0%	1,0%	249
	52,9%	62,5%	49,8%	56,7%	67,6%	52,8%	3,8%	5,2%	3,1%	
Common classes										
Residential areas (one and two-family houses)	75,2%	71,0%	80,0%	74,0%	74,0%	74,0%	-1,2%	3,0%	-6,0%	630
Arable land in general	54,4%	85,0%	40,0%	56,7%	74,0%	46,0%	2,3%	-11,0%	6,0%	225
Roads	67,9%	71,0%	65,0%	66,5%	68,0%	65,0%	-1,4%	-3,0%	0,0%	454
Group of semi-natural grassland	75,5%	75,0%	76,0%	74,9%	73,0%	77,0%	-0,6%	-2,0%	1,0%	1262
	68%	76%	65%	68%	72%	66%	0%	-3%	0%	
Frequenct classes										
Vineyards	93,0%	91,0%	95,0%	93,5%	93,0%	94,0%	0,5%	2,0%	-1,0%	932
Group of alpine pastures	83,0%	82,0%	84,0%	82,5%	81,0%	84,0%	-0,5%	-1,0%	0,0%	2443
Unused	89,5%	91,0%	88,0%	89,5%	91,0%	88,0%	0,0%	0,0%	0,0%	4552
Forest	93,4%	90,0%	97,0%	93,4%	91,0%	96,0%	0,1%	1,0%	-1,0%	5599
	89,7%	88,5%	91,0%	89,7%	89,0%	90,5%	0,0%	0,5%	-0,5%	

Figure 25: Comparison of the results for each class from the baseline model and the focal class balanced loss model on the clean dataset.

6.2.2 Results per class

In order to better understand where the `fCBL` effectively overtakes the baseline model, Figure 25 compares the models' precision, recall and F1-score for each category.

The most difficult categories among the rare classes are alpine sport facilities, sport facilities, public buildings, agricultural buildings and parking areas. They receive the worst accuracy metrics and as mentioned earlier, they show complex patterns with high variability and models have trouble to predict them. Interestingly, they are also the classes where the differences in performance between the baseline and the `fCBL` are the largest. For these difficult classes, the `fCBL` overtakes the baseline with a positive difference in F1-score ranging from 3% to 17%. Reversely, the performances of the `fCBL` is beaten by the baseline for the motorways, lakes,

commercial and industrial areas and the damaged forests. On average, the focal class balanced loss exceeds the baseline for the rare classes on the average F1-score by 3.8%, the precision by 5.2% and the recall by 3.1%. These results indicate that the `fCBL` does help in the case of difficult and rare category.

The semi-natural grasslands group obtains a F1-score of 75%, which is a considerable increase compared to the individual results of its sub-categories on the full dataset (semi-natural grassland 52% and the farm pastures 50%, see the Appendix (Figure 30)). The fusion of these two class allows to reduce the error related to their mutual confusion.

Compared to the baseline model on the full dataset, the formation of the group alpine pastures allowed a F1-score value of 83%. This score is driven by the majority class Alpine pastures in general that attained 82% in the full data set. It is not damaged by the addition of the two other classes that previously achieved much lower results (with a F1-score of 23% for the alpine sheep grazing pastures and 40% for the alpine meadows).

To summarise, the focal class balanced loss method reaches the goal of compensating the class imbalance by increasing the rare classes accuracy after a dataset cleaning. Additionally, the grouping of several similar classes into one manages to make them more coherent and easier for the model to identify.

6.3 Comparison with ADELE predictions

We compare the F1-score of the focal class balanced loss with the general results of ADELE in its report (Jordan, Meyer, et al., 2019). Our `fCBL` model overtakes ADELE on a few rare classes such as the motorways, the orchards and the rivers and it has very similar results for the frequent classes. These observations motivated us to proceed to a deeper comparison with ADELE predictions on a similar test set.

This experiment compares the results of the baseline and the `fCBL` models with the predictions from ADELE on a similar test set. Table 4 presents the results of both models with ADELE predictions on the test area. The ADELE metrics for each experiment are computed on the same class or class group as our models.

Results on the full test area (28 classes)

	K	**OA**	**AP**	**AR**	**F1**	**Pr**	**Pc**	**Pf**
ADELE	81,4%	86,3%	**61,1%**	**46,1%**	**49,6%**	**53,7%**	**70,5%**	89,0%
Baseline	**82,0%**	**86,7%**	48,0%	46,0%	45,0%	35,6%	68,0%	**89,8%**

Results on the clean test area (21 classes)

	K	**OA**	**AP**	**AR**	**F1**	**Pr**	**Pc**	**Pf**
ADELE	84,7%	88,9%	**70,0%**	57,7%	**61,6%**	**61,8%**	**81,0%**	**90,3%**
`fCBL`	**85,6%**	**89,4%**	54,4%	**61,1%**	54,4%	40,4%	76,3%	87,0%

Table 4: General results for ADELE and the baseline on 28 classes, and ADELE and the focal class balanced loss on 21 classes

For both the clean and the full test area, ADELE outperforms our models by a large margin for the average and rare class precision. The experiment on the test area with 28 classes leads our

baseline model to some metrics with superior (K, OA, Pf) or similar range of values (Pc, AR) as ADELE. The baseline model a slightly better precision for the frequent classes which leads to better performances on metrics that do not take label imbalance into account such as OA. The `fCBL` surpasses ADELE for the κ, overall accuracy and average recall, but it is exceeded for other metrics (AP, F1, Pr, Pc, Pf). We observe that the `fCBL` favours more average recall than precision contrarily to ADELE.

The large difference in precision regarding the rare classes is not a surprise. Our training set covers a much smaller area than the one from ADELE and our models are trained in absence of auxiliary information. However, the similarity in performances for the frequent and common categories indicates that our model already has a proper understanding of these classes.

	ADELE			fCBL			Difference			test set
Rare classes	F1	P	R	F1	P	R	F1	P	R	size
Alpine sports facilities	0%	0%	0%	0%	0%	0%	0%	0%	0%	1
Public buildings and surroundings	0%	0%	0%	25%	14%	100%	25%	14%	100%	1
Damaged forest	0%	0%	0%	0%	0%	0%	0%	0%	0%	1
Motorways	40%	33%	50%	50%	50%	50%	10%	17%	0%	2
Sports facilities	80%	100%	67%	80%	100%	67%	0%	0%	0%	3
Parking areas	57%	100%	40%	0%	0%	0%	-57%	-100%	-40%	4
Industrial and commercial areas > 1 ha	100%	100%	100%	60%	50%	75%	-40%	-50%	-25%	4
Golf courses	80%	100%	67%	73%	62%	89%	-7%	-38%	22%	9
Agricultural buildings and surroundings	23%	50%	15%	0%	0%	0%	-23%	-50%	-15%	13
Residential areas (blocks of flats)	38%	45%	33%	48%	38%	67%	10%	-7%	34%	15
Orchards	67%	83%	56%	64%	57%	72%	-3%	-26%	16%	18
Rivers streams	48%	73%	36%	44%	36%	55%	-5%	-37%	19%	22
Lakes	96%	98%	95%	82%	76%	88%	-15%	-22%	-7%	40
Arable land in general	58%	71%	49%	35%	82%	22%	-23%	11%	-27%	41
	49%	61%	43%	40%	40%	49%	-9%	-21%	6%	
Common classes										
Roads	78%	82%	75%	65%	72%	60%	-13%	-10%	-15%	121
Residential areas (one and two-family houses)	83%	78%	89%	79%	78%	81%	-4%	0%	-8%	237
Group of semi-natural grasslands	57%	64%	51%	82%	79%	85%	25%	15%	34%	524
	73%	75%	72%	76%	76%	75%	3%	2%	4%	
Frequent classes										
Vineyards	95%	95%	95%	83%	71%	100%	-12%	-24%	5%	62
Group of alpine pastures	80%	80%	81%	86%	89%	83%	5%	9%	2%	1007
Unused	93%	93%	94%	93%	95%	92%	0%	2%	-2%	2654
Forest	92%	88%	96%	95%	93%	97%	3%	5%	1%	1775
	90%	89%	92%	89%	87%	93%	-1%	-2%	2%	

Figure 26: Comparison per class of the focal class balanced loss and ADELE on the test area with 21 classes

Figure 26 gives a deeper insight into the classes that make the difference between the `fCBL` and ADELE. Details for the comparison on the dataset with 28 classes is given in the Appendix (Figure 34).

First, the number of test samples in some classes is very small with less than 5 samples for several classes, meaning that the metrics presented may not be representative of the general model performances. Several of these small classes including alpine sport facilities, public buildings, damaged forests and parking areas are entirely misclassified by one or both models. These results (highlighted in red in Figure 26) are not further analysed.

The rare classes are generally more likely to be misclassified by our model than by ADELE, since the latter ameliorates our average F1-score by 9%. ADELE seems to overtake the `fCBL` by a large margin especially on the different buildings and constructions types such as the parking

areas, the industrial and agricultural buildings and roads. An exception to this trend concerns the residential areas with blocks of flat where our model receives a higher F1-score and recall value. The superior recognition of buildings by ADELE is probably related to the input of the cadastral surfaces categories that significantly helps in the identification of buildings.

For categories with vegetation or nature like golf courses, orchards, rivers and streams both models show more similar performances, but the `fCBL` is still slightly lower than ADELE. This trend is validated by the per-class result from the baseline model on the test area with 28 classes (see detailed results in the Appendix Figure 34): Sheep grazing pastures, alpine meadows, farm pastures and semi-natural grasslands are better classified by our baseline model than by ADELE.

The frequent classes are predicted with very similar accuracy from both models. These classes are among the easiest to recognise and do not require auxiliary data for their identification.

To sum up, our model performs similarly to ADELE for the most frequent and easily recognisable classes such as forests or vineyards. When more challenging classes are displayed, our model is outperformed. The limitations of our model come from the relative small dataset size and the absence of auxiliary information that makes the predictions of the rare categories very difficult.

7 Limitations

This section highlights the limits regarding the results presented in this study. First, the study area covers a relative small portion of the Swiss territory and consequently it presents a rather low intra-class variability. For instance, the forest category comprises mostly spruce trees in our study area. However forests of deciduous trees are very common in Switzerland but do not appear in this dataset. Our experiments showed that the category complexity is a key driver of the classifier performances. In consequence, an experiment on an area with more diversified classes and more representative of the overall Swiss land use would give an insight on the generalisation ability of our methodology.

Regarding the ground truth labels, the time gap between the label production and the photography survey introduced some erroneous labels. A more consistent time scale would be beneficial for the model. Moreover, our experiments enhanced the importance of the auxiliary data. Several classes are confounded due to their high similarity. To this end, a new model would need to be designed in order to accept different types of data such as categorical variables.

The training scheme for the equalisation loss, the focal and softmax class balanced loss have been simplified on the clean dataset. In particular, the process of selection of the best hyper-parameters has not been repeated for each experiment. Performances could potentially be improved by finely tuning the hyper-parameters.

Numerous other methods addressing the class imbalance exist. A small selection of them have been applied in this work due to the time constraint. Other methods could lead to enhanced results. For instance, the use of the sigmoid loss function could be an interesting alternative to the softmax function, as it is said to be more robust in case of strong similarities between classes. As future works, an interesting methodology for the scene classification is proposed by Zhang et al. (2019). Their study design a general workflow including a deep neural network that is able to predict both land cover and land use based on aerial photographs. Since the Areal Statistics requires both of them, the training and the predictions of both classes by one model could potentially save time.

8 Conclusion

In this report we discussed the use of convolutional neural networks to automatically predict the land use labels for the Swiss Areal Statistics. Our model exploited a series of about 60'000 aerial images with visible and near-infrared bands in combination with their corresponding digital elevation model. The samples constituting this set exposed a skewed distribution towards some very frequent class such as forests. Through our experiments we tried to address the class imbalance problem that is known to damages the classification results for less frequent categories.

We applied several sampling methods and specific loss functions to improve the classification results of the rare classes. The experiment performed on 28 land use classes showed that high classification accuracy can be obtained on frequent classes, but also on some rare and distinctive categories. Consequently, we discovered that the limited amount of data for a class is not the only factor altering the classification results. The category complexity is another major driver of misclassification of images derived from minority classes. For instance, some classes such sports facilities or unspecified buildings present a large number of different features. The network does not manage to recognise all their variations, especially with a limited number of samples. Moreover, the high level of similarity between different categories also renders their recognition arduous. The distinction of several classes sometimes requires the access to additional geographical information concerning their usage. Another important source of erroneous predictions is related to the fact that the labels are strongly dependant on the objects located at the centre of the reference surface. The model is confused when another class is present on the image.

As a result, we decided to further investigate the class imbalance problem after proceeding to a dataset cleaning. The so-called clean dataset has a reduced number of classes where the most ambiguous samples are either discarded or grouped together. As we solved the class ambiguity problem, we successfully employ the focal class balanced loss on the clean dataset. We obtained a reduction by 4% of the error rate on the rare classes compared to a baseline model. We could verify that this method of network training successfully targets difficult samples from the rare categories.

Finally, we compare our predictions results with ADELE, a model developed by FHNW to produce the land use labels for the Areal Statistics. The predictions results on the rare categories are poor due to our restrained dataset size and the absence of auxiliary data. Nevertheless we reached a level of accuracy similar to ADELE for the more frequent classes.

To conclude, we demonstrated that addressing class imbalance on a realistic dataset does improve classification performances despite the need for a careful data filtering. Auxiliary data remains a requirement to produce more accurate labels for the rare classes. To conclude, we want to enhance the necessity of constructing a model with specific sources of information for each class. A careful selection of training samples regarding their abundance and variability would help to obtain the best performances. Finally, we confirm the necessity and the potential to go towards automatised land use classification, where a significant part of the labelling can be acquired through artificial intelligence.

9 Appendix

9.1 Illustration of the 28 land-use categories

Figure 27: Illustration of the land-use categories: Settlement and urban areas

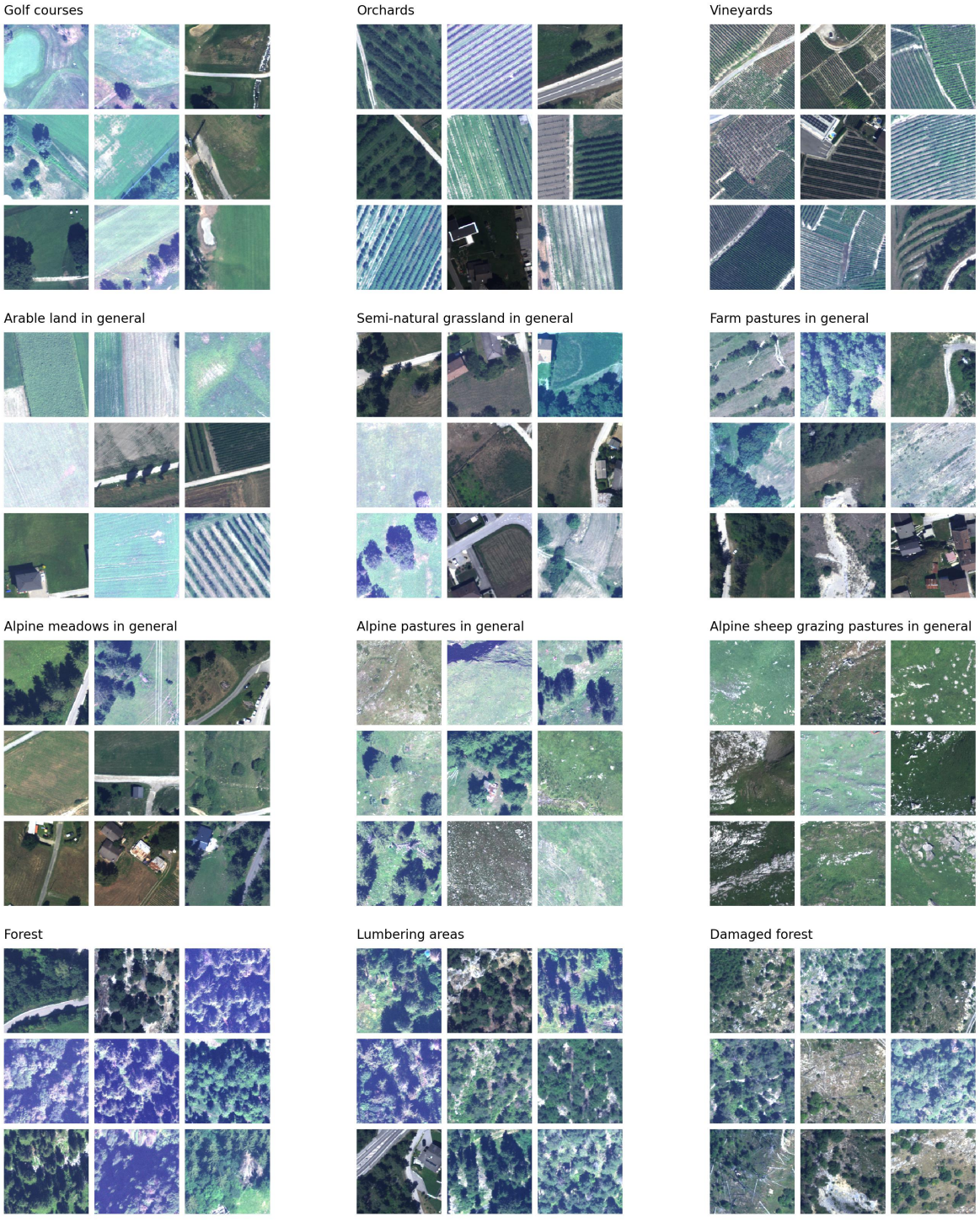


Figure 28: Illustration of the land-use categories: Agricultural areas and forests

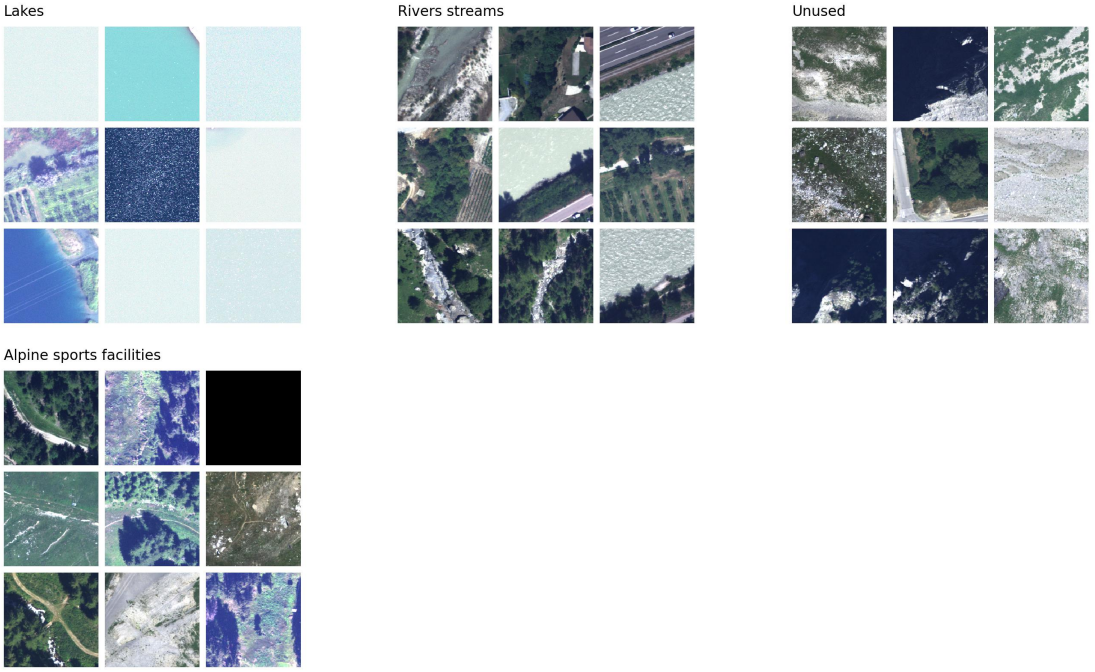


Figure 29: Illustration of the land-use categories: Unproductive areas

9.2 Number of samples per class in the full dataset

Class distribution	Total
Agricultural buildings and surroundings	202
Motorways	102
Construction sites	116
Unexploited urban areas	125
Sports facilities	148
Alpine sports facilities	150
Alpine sheep grazing pastures in general	176
Public buildings and surroundings	181
Unspecified buildings and surroundings	191
Golf courses	198
Parking areas	203
Damaged forest	287
Lakes	291
Industrial and commercial areas 1 ha	338
Residential areas (blocks of flats)	503
Alpine meadows in general	528
Rivers streams	689
Arable land in general	746
Orchards	827
Lumbering areas	876
Roads	1511
Farm pastures in general	2000
Residential areas (one and two-family houses)	2096
Semi-natural grassland in general	2203
Vineyards	3105
Alpine pastures in general	7426
Unused	15171
Forest	18658
Total	**59047**

Classes removed from the dataset	Total
Afforestation	2
Cemeteries	13
Waste water treatment plants	17
Landscape interventions	26
Horticulture	31
Energy supply plants	34
Residential areas (terraced houses)	36
Other supply or waste treatment plants	44
Garden allotments	49
Dumps	51
Camping areas	52
Avalanche and rockfall protection structures	65
Airports and airfields	79
Public parks	82
Flood protection structures	83
Railway surfaces	86
Quarries mines	86
Industrial and commercial areas 1 ha	90
Total	**926**

Table 5: left: Distribution of classes in the full dataset, right: Classes removed from the dataset due to their small size

9.3 List of Experiments

Experiments on the full dataset			**Experiments on the clean dataset**	
Method	Training sets	Test sets	Training sets	Test sets
Baseline	full train	full test	clean train	clean test
`und-100`	100-train	full test	clean 100-train	clean test
`und-1000`	1000-train	full test	clean 1000-train	clean test
`two-phases`	full + 1000-train	full test	clean + clean 1000-train	clean test
`inv_freq`	full train	full test	clean train	clean test
`sq_inv` freq	full train	full test	clean train	clean test
`sCBL`	full train	full test	clean train	clean test
`fCBL`	full train	full test	clean train	clean test
`EQL`	full train	full test	clean train	clean test
Comparison ADELE-Baseline			**Comparison clean ADELE-fCBL**	
Baseline	fhnw train	fhnw test	clean fhnw train	clean fhnw test

Table 6: List of experiments

9.4 Detailed results

9.4.1 Results for several hyper-parameters for the Equalization loss

θ	T_λ	**K**	**OA**	**AP**	**AR**	**F1**	**Pr**	**Pc**	**Pf**	**Epochs**	**Patience**
95%	600	71,5%	76,6%	50,8%	48,4%	44,9%	38,6%	**73,5%**	**89,5%**	36	20
95%	600	72,9%	77,8%	47,3%	48,5%	44,8%	34,2%	70,5%	**89,5%**	100	-
90%	1500	75,6%	80,2%	49,9%	47,0%	46,1%	41,2%	54,0%	89,2%	68	40
90%	900	75,5%	80,1%	51,5%	49,3%	47,9%	41,0%	67,0 %	89,0%	100	-
90%	600	75,7%	80,2%	53,0%	51,4%	48,8%	42,6%	68,2%	**89,5%**	100	-
90%	300	77,1%	81,6%	51,6%	49,5%	49,1%	41,9%	64,5%	87,2%	100	-
75%	600	77,1%	81,5%	55,5%	**52,3%**	51,8%	46,8%	66,0%	88,5%	73	20
75%	300	77,3%	81,8%	58,6%	48,0%	49,2%	52,4%	61,2%	86,8%	31	15
75%	300	**77,8%**	**82,1%**	56,9%	51,5%	**52,7%**	49,6%	62,2%	88,0%	65	20
75%	300	77,4%	81,9%	56,5%	49,8%	51,5%	48,8%	64,0%	87,2%	100	-
50%	300	77,4%	81,8%	**59,2%**	48,5%	49,4%	**53,1**%	60,8%	88,5%	39	20
Baseline	-	**78,2%**	82,5%	**63,1%**	52,0%	54,6%	**58,3%**	62,0%	88,0%	100	-

Table 7: Evaluation of several hyper-parameters for the Equalization loss on the full dataset. θ is the random variable used to randomly maintain the gradient, T_λ is the threshold number for the minority classes, as number of samples in the test set. The patience is the number of epochs without improvement used for early stopping

9.4.2 Results for several hyper-parameters for the Class Balanced loss

	γ	β	**K**	**OA**	**AP**	**AR**	**F1**	**Pr**	**Pc**	**Pf**	**Epochs**
`sCBL`	-	0,9999	74,6%	79,3%	53,1%	54,1%	52,6%	44,6%	61,2%	87,8%	100
`sCBL`	-	0,999	74,2%	79,0%	51,4%	**57,4%**	52,7%	41,4%	**64,5%**	**88,3%**	100
`sCBL`	-	0,99	74.5%	79.2%	53.6%	55.4%	53.3%	45.5%	55.8%	89.0%	100
`sCBL`	-	0,99	**77,3%**	**81,8%**	**61,1%**	50,4%	**53,1%**	**56,0%**	60,5%	86,8%	51(30*)
`fCBL`	0,5	0,99	77,2%	81,6%	61,0%	**51,9%**	**53,8%**	55,9%	59,2%	**88,3%**	100
`fCBL`	1	0,99	**77,3%**	**81,8%**	**62,3%**	50,7%	52,4%	**57,3%**	**62,8%**	86,2%	100
`fCBL`	2	0,99	**77,3%**	**81,8%**	58,3%	47,7%	49,1%	52,1%	61,2%	86,5%	100
`fCBL`	3	0,99	75,4%	80,4%	55,2%	43,3%	44,7%	48,5%	57,7%	86,2%	100
Baseline	-	-	78,2%	82,5%	63,1%	52,0%	54,6%	58,3%	62,0%	88,0%	100

Table 8: Evaluation of several hyper-parameters for the softmax and focal Class Balanced loss on the full dataset. * number of epochs without improvement used for early stopping

9.4.3 Results per class for all models on the full dataset

Label	Baseline			und-100			und-1000			two-phases			Test set size
	F1	P	R	F1	P	R	F1	P	R	F1	P	R	
Agricultural buildings and surroundings	23,3%	43,0%	16,0%	12,9%	9,0%	23,0%	20,0%	30,0%	15,0%	25,0%	31,0%	21,0%	61
Alpine meadows in general	40,2%	54,0%	32,0%	36,0%	25,0%	64,0%	42,8%	38,0%	49,0%	44,2%	41,0%	48,0%	160
Alpine pastures in general	82,0%	80,0%	84,0%	55,5%	75,0%	44,0%	75,5%	75,0%	76,0%	77,1%	72,0%	83,0%	2229
Alpine sheep grazing pastures in general	22,7%	88,0%	13,0%	7,6%	4,0%	78,0%	18,8%	12,0%	43,0%	21,2%	14,0%	44,0%	54
Alpine sports facilities	31,0%	69,0%	20,0%	7,5%	4,0%	67,0%	16,0%	10,0%	40,0%	21,0%	20,0%	22,0%	45
Arable land in general	65,0%	66,0%	64,0%	59,0%	58,0%	60,0%	64,5%	56,0%	76,0%	65,3%	62,0%	69,0%	225
Construction sites	0,0%	0,0%	0,0%	4,7%	3,0%	11,0%	0,0%	0,0%	0,0%	5,7%	50,0%	3,0%	36
Damaged forest	66,7%	75,0%	60,0%	22,2%	13,0%	75,0%	44,3%	32,0%	72,0%	52,2%	42,0%	69,0%	87
Farm pastures in general	55,0%	56,0%	54,0%	49,1%	45,0%	54,0%	55,5%	46,0%	70,0%	52,1%	43,0%	66,0%	600
Forest	91,4%	89,0%	94,0%	80,8%	89,0%	74,0%	82,2%	94,0%	73,0%	84,7%	94,0%	77,0%	5599
Golf courses	84,0%	83,0%	85,0%	80,6%	75,0%	87,0%	74,8%	64,0%	90,0%	82,9%	86,0%	80,0%	61
Industrial and commercial areas > 1 ha	69,5%	64,0%	76,0%	58,5%	60,0%	57,0%	64,0%	51,0%	86,0%	68,9%	67,0%	71,0%	102
Lakes	89,0%	90,0%	88,0%	48,9%	33,0%	94,0%	83,3%	73,0%	97,0%	87,8%	84,0%	92,0%	88
Lumbering areas	22,9%	35,0%	17,0%	16,1%	10,0%	41,0%	22,8%	14,0%	61,0%	25,2%	16,0%	59,0%	264
Motorways	85,8%	82,0%	90,0%	61,8%	46,0%	94,0%	75,7%	81,0%	71,0%	75,7%	71,0%	81,0%	31
Orchards	71,9%	79,0%	66,0%	61,9%	59,0%	65,0%	67,7%	61,0%	76,0%	68,2%	64,0%	73,0%	249
Parking areas	36,0%	43,0%	31,0%	33,8%	25,0%	52,0%	31,7%	29,0%	35,0%	43,0%	39,0%	48,0%	62
Public buildings and surroundings	27,6%	59,0%	18,0%	27,8%	24,0%	33,0%	16,5%	33,0%	11,0%	34,4%	36,0%	33,0%	55
Residential areas (blocks of flats)	46,2%	50,0%	43,0%	31,1%	44,0%	24,0%	45,6%	39,0%	55,0%	44,8%	42,0%	48,0%	152
Residential areas (one and two-family houses)	72,6%	64,0%	84,0%	58,6%	67,0%	52,0%	68,0%	67,0%	69,0%	68,4%	66,0%	71,0%	630
Rivers streams	69,7%	78,0%	63,0%	45,8%	37,0%	60,0%	55,4%	45,0%	72,0%	60,3%	53,0%	70,0%	208
Roads	66,4%	69,0%	64,0%	33,6%	42,0%	28,0%	51,7%	46,0%	59,0%	66,2%	62,0%	71,0%	454
Semi-natural grassland in general	60,5%	59,0%	62,0%	50,8%	62,0%	43,0%	52,2%	72,0%	41,0%	53,5%	62,0%	47,0%	662
Sports facilities	53,7%	86,0%	39,0%	43,0%	43,0%	43,0%	50,2%	89,0%	35,0%	46,0%	67,0%	35,0%	46
Unexploited urban areas	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	38
Unspecified buildings and surroundings	10,6%	22,0%	7,0%	18,2%	12,0%	38,0%	15,5%	15,0%	16,0%	19,2%	22,0%	17,0%	58
Unused	90,0%	90,0%	90,0%	73,1%	96,0%	59,0%	85,6%	92,0%	80,0%	86,6%	93,0%	81,0%	4552
Vineyards	94,5%	93,0%	96,0%	87,4%	84,0%	91,0%	90,5%	90,0%	91,0%	93,5%	94,0%	93,0%	932
Average	55%	63%	52%	42%	41%	54%	49%	48%	56%	53%	53%	56%	

Label	inverse frequency			square root inverse frequency			EQL			sCBL			fCBL			Test set size
	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	
Agricultural buildings and surroundings	21,3%	32,0%	16,0%	17,3%	41,0%	11,0%	0,0%	0,0%	0,0%	31,3%	42,0%	25,0%	15,2%	32,0%	10,0%	61
Alpine meadows in general	48,7%	45,0%	53,0%	27,8%	52,0%	19,0%	26,3%	74,0%	16,0%	39,8%	50,0%	33,0%	36,1%	59,0%	26,0%	160
Alpine pastures in general	79,2%	75,0%	84,0%	81,3%	77,0%	86,0%	80,5%	80,0%	81,0%	81,4%	79,0%	84,0%	81,0%	80,0%	82,0%	2229
Alpine sheep grazing pastures in general	25,0%	24,0%	26,0%	16,2%	83,0%	9,0%	16,0%	71,0%	9,0%	15,5%	56,0%	9,0%	7,7%	100,0%	4,0%	54
Alpine sports facilities	13,7%	9,0%	29,0%	14,1%	33,0%	9,0%	14,4%	36,0%	9,0%	30,2%	48,0%	22,0%	21,0%	55,0%	13,0%	45
Arable land in general	64,9%	67,0%	63,0%	67,9%	65,0%	71,0%	60,0%	75,0%	50,0%	66,9%	64,0%	70,0%	62,9%	66,0%	60,0%	225
Construction sites	8,6%	15,0%	6,0%	10,2%	33,0%	6,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	5,1%	17,0%	3,0%	36
Damaged forest	55,1%	45,0%	71,0%	68,2%	85,0%	57,0%	64,8%	77,0%	56,0%	53,6%	81,0%	40,0%	65,9%	78,0%	57,0%	87
Farm pastures in general	53,8%	47,0%	63,0%	52,5%	53,0%	52,0%	59,2%	55,0%	64,0%	53,9%	60,0%	49,0%	55,0%	61,0%	50,0%	600
Forest	87,7%	93,0%	83,0%	90,5%	89,0%	92,0%	90,8%	87,0%	95,0%	90,5%	89,0%	92,0%	90,9%	88,0%	94,0%	5599
Golf courses	86,0%	85,0%	87,0%	86,2%	98,0%	77,0%	83,7%	79,0%	89,0%	81,0%	96,0%	70,0%	82,0%	84,0%	80,0%	61
Industrial and commercial areas > 1 ha	67,1%	60,0%	76,0%	70,3%	61,0%	83,0%	64,3%	50,0%	90,0%	66,4%	64,0%	69,0%	66,0%	61,0%	72,0%	102
Lakes	87,6%	82,0%	94,0%	88,3%	84,0%	93,0%	91,0%	89,0%	93,0%	86,3%	91,0%	82,0%	88,0%	88,0%	88,0%	88
Lumbering areas	28,1%	20,0%	47,0%	21,2%	24,0%	19,0%	15,2%	48,0%	9,0%	18,6%	41,0%	12,0%	18,2%	38,0%	12,0%	264
Motorways	77,5%	68,0%	90,0%	78,7%	74,0%	84,0%	81,7%	77,0%	87,0%	65,1%	52,0%	87,0%	72,0%	60,0%	90,0%	31
Orchards	69,9%	73,0%	67,0%	66,0%	59,0%	75,0%	71,6%	83,0%	63,0%	73,1%	87,0%	63,0%	71,7%	77,0%	67,0%	249
Parking areas	38,2%	42,0%	35,0%	22,0%	41,0%	15,0%	20,9%	30,0%	16,0%	33,0%	32,0%	34,0%	42,5%	37,0%	50,0%	62
Public buildings and surroundings	37,3%	40,0%	35,0%	12,1%	44,0%	7,0%	6,9%	25,0%	4,0%	35,9%	47,0%	29,0%	33,8%	45,0%	27,0%	55
Residential areas (blocks of flats)	51,4%	54,0%	49,0%	46,9%	49,0%	45,0%	30,5%	64,0%	20,0%	43,1%	59,0%	34,0%	49,9%	52,0%	48,0%	152
Residential areas (one and two-family houses)	70,5%	65,0%	77,0%	69,3%	66,0%	73,0%	73,4%	64,0%	86,0%	72,9%	65,0%	83,0%	72,0%	63,0%	84,0%	630
Rivers streams	60,9%	58,0%	64,0%	66,4%	76,0%	59,0%	66,2%	83,0%	55,0%	64,3%	90,0%	50,0%	63,3%	93,0%	48,0%	208
Roads	56,9%	55,0%	59,0%	64,3%	61,0%	68,0%	57,0%	68,0%	49,0%	63,0%	56,0%	72,0%	64,5%	64,0%	65,0%	454
Semi-natural grassland in general	59,0%	59,0%	59,0%	55,2%	65,0%	48,0%	61,7%	58,0%	66,0%	60,0%	61,0%	59,0%	62,0%	63,0%	61,0%	662
Sports facilities	51,6%	85,0%	37,0%	49,3%	74,0%	37,0%	35,6%	69,0%	24,0%	53,7%	86,0%	39,0%	47,9%	76,0%	35,0%	46
Unexploited urban areas	5,5%	33,0%	3,0%	4,8%	12,0%	3,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	38
Unspecified buildings and surroundings	15,0%	20,0%	12,0%	19,8%	15,0%	29,0%	22,6%	19,0%	28,0%	24,6%	35,0%	19,0%	13,7%	29,0%	9,0%	58
Unused	88,2%	94,0%	83,0%	89,5%	91,0%	88,0%	89,5%	89,0%	90,0%	89,0%	87,0%	91,0%	88,5%	88,0%	89,0%	4552
Vineyards	92,5%	91,0%	94,0%	93,4%	91,0%	96,0%	93,0%	91,0%	95,0%	92,5%	92,0%	93,0%	92,4%	89,0%	96,0%	932
Average	54%	55%	56%	52%	61%	50%	49%	59%	48%	53%	61%	50%	52%	62%	51%	

Figure 30: Detailed results per class for all the methods tested on the full dataset

9.4.4 Results per class for all models on the clean dataset experiment

Label	Baseline			und-100			und-1000			two-phases			Test set
	F1	P	R	F1	P	R	F1	P	R	F1	P	R	
Agricultural buildings and surroundings	0,0%	0,0%	0,0%	8,7%	6,0%	16,0%	9,4%	75,0%	5,0%	15,8%	38,0%	10,0%	61
Alpine pastures in general	83,0%	82,0%	84,0%	72,9%	75,0%	71,0%	77,1%	72,0%	83,0%	81,0%	79,0%	83,0%	2443
Alpine sports facilities	18,4%	56,0%	11,0%	5,7%	3,0%	69,0%	17,0%	12,0%	29,0%	22,3%	19,0%	27,0%	45
Arable land in general	54,4%	85,0%	40,0%	51,7%	41,0%	70,0%	65,4%	58,0%	75,0%	66,4%	59,0%	76,0%	225
Damaged forest	66,9%	83,0%	56,0%	11,2%	6,0%	85,0%	55,8%	46,0%	71,0%	44,9%	32,0%	75,0%	87
Forest	93,4%	90,0%	97,0%	81,5%	94,0%	72,0%	91,9%	95,0%	89,0%	92,5%	94,0%	91,0%	5599
Golf courses	72,4%	63,0%	85,0%	55,4%	40,0%	90,0%	79,9%	78,0%	82,0%	75,5%	65,0%	90,0%	61
Industrial and commercial areas > 1 ha	71,5%	66,0%	78,0%	60,7%	51,0%	75,0%	71,7%	61,0%	87,0%	70,5%	69,0%	72,0%	101
Lakes	91,0%	93,0%	89,0%	51,3%	36,0%	89,0%	76,5%	66,0%	91,0%	86,0%	80,0%	93,0%	88
Motorways	89,0%	88,0%	90,0%	60,0%	45,0%	90,0%	88,5%	90,0%	87,0%	87,4%	85,0%	90,0%	31
Orchards	72,9%	76,0%	70,0%	62,5%	64,0%	61,0%	67,4%	60,0%	77,0%	72,8%	77,0%	69,0%	249
Parking areas	28,6%	45,0%	21,0%	28,4%	21,0%	44,0%	35,3%	41,0%	31,0%	30,3%	41,0%	24,0%	62
Public buildings and surroundings	22,4%	44,0%	15,0%	21,2%	19,0%	24,0%	17,5%	43,0%	11,0%	28,2%	48,0%	20,0%	55
Residential areas (blocks of flats)	52,5%	54,0%	51,0%	37,3%	28,0%	56,0%	49,7%	38,0%	72,0%	54,5%	53,0%	56,0%	152
Residential areas (one and two-family houses)	75,2%	71,0%	80,0%	60,4%	63,0%	58,0%	68,8%	65,0%	73,0%	75,5%	70,0%	82,0%	630
Rivers streams	67,6%	83,0%	57,0%	35,6%	25,0%	62,0%	47,5%	35,0%	74,0%	64,0%	59,0%	70,0%	208
Roads	67,9%	71,0%	65,0%	30,5%	31,0%	30,0%	49,0%	43,0%	57,0%	60,5%	49,0%	79,0%	454
Semi-natural grassland in general	75,5%	75,0%	76,0%	53,9%	65,0%	46,0%	70,5%	71,0%	70,0%	72,5%	71,0%	74,0%	1262
Sports facilities	34,4%	61,0%	24,0%	32,9%	25,0%	48,0%	51,6%	85,0%	37,0%	38,6%	75,0%	26,0%	46
Unused	89,5%	91,0%	88,0%	75,5%	90,0%	65,0%	82,8%	91,0%	76,0%	87,2%	93,0%	82,0%	4552
Vineyards	93,0%	91,0%	95,0%	86,5%	86,0%	87,0%	91,0%	92,0%	90,0%	93,0%	92,0%	94,0%	932
Average	63%	70%	61%	47%	44%	62%	60%	63%	65%	63%	64%	66%	

Label	inverse frequency			square root inverse frequency			EQL			sCBL			fCBL			Test set size
	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	
Agricultural buildings and surroundings	22,2%	43,0%	15,0%	14,4%	26,0%	10,0%	0,0%	0,0%	0,0%	11,7%	36,0%	7,0%	16,7%	50,0%	10,0%	61
Alpine pastures in general	82,0%	81,0%	83,0%	81,5%	82,0%	81,0%	82,5%	82,0%	83,0%	81,5%	80,0%	83,0%	82,5%	81,0%	84,0%	2443
Alpine sports facilities	29,3%	32,0%	27,0%	23,0%	19,0%	29,0%	22,0%	22,0%	22,0%	27,6%	37,0%	22,0%	24,2%	50,0%	16,0%	45
Arable land in general	66,4%	69,0%	64,0%	63,8%	56,0%	74,0%	60,2%	56,0%	65,0%	67,2%	63,0%	72,0%	56,7%	74,0%	46,0%	225
Damaged forest	63,0%	62,0%	64,0%	55,2%	46,0%	69,0%	62,0%	61,0%	63,0%	70,2%	81,0%	62,0%	63,1%	74,0%	55,0%	87
Forest	93,0%	92,0%	94,0%	93,0%	91,0%	95,0%	93,0%	91,0%	95,0%	93,4%	91,0%	96,0%	93,4%	91,0%	96,0%	5599
Golf courses	81,0%	82,0%	80,0%	84,3%	89,0%	80,0%	79,8%	76,0%	84,0%	77,1%	68,0%	89,0%	83,0%	81,0%	85,0%	61
Industrial and commercial areas > 1 ha	41,6%	68,0%	30,0%	67,9%	57,0%	84,0%	64,2%	52,0%	84,0%	71,1%	64,0%	80,0%	68,6%	58,0%	84,0%	101
Lakes	87,2%	82,0%	93,0%	86,6%	81,0%	93,0%	87,4%	91,0%	84,0%	92,0%	94,0%	90,0%	88,8%	93,0%	85,0%	88
Motorways	77,5%	66,0%	94,0%	87,4%	85,0%	90,0%	76,7%	88,0%	68,0%	78,7%	74,0%	84,0%	82,0%	83,0%	81,0%	31
Orchards	72,0%	84,0%	63,0%	67,3%	75,0%	61,0%	69,4%	67,0%	72,0%	73,3%	84,0%	65,0%	72,0%	73,0%	71,0%	249
Parking areas	30,9%	38,0%	26,0%	25,0%	41,0%	18,0%	0,0%	0,0%	0,0%	21,2%	57,0%	13,0%	39,1%	46,0%	34,0%	62
Public buildings and surroundings	24,4%	15,0%	65,0%	28,9%	52,0%	20,0%	0,0%	0,0%	0,0%	43,5%	64,0%	33,0%	25,4%	43,0%	18,0%	55
Residential areas (blocks of flats)	47,3%	60,0%	39,0%	45,2%	58,0%	37,0%	43,8%	39,0%	50,0%	53,2%	65,0%	45,0%	53,9%	52,0%	56,0%	152
Residential areas (one and two-family houses)	73,5%	66,0%	83,0%	73,7%	67,0%	82,0%	71,5%	73,0%	70,0%	75,8%	72,0%	80,0%	74,0%	74,0%	74,0%	630
Rivers streams	63,8%	68,0%	60,0%	64,5%	64,0%	65,0%	61,8%	56,0%	69,0%	64,0%	63,0%	65,0%	68,6%	91,0%	55,0%	208
Roads	57,6%	77,0%	46,0%	64,6%	73,0%	58,0%	60,4%	51,0%	74,0%	65,0%	66,0%	64,0%	66,5%	68,0%	65,0%	454
Semi-natural grassland in general	72,5%	72,0%	73,0%	72,9%	76,0%	70,0%	73,1%	79,0%	68,0%	73,9%	77,0%	71,0%	74,9%	73,0%	77,0%	1262
Sports facilities	46,0%	67,0%	35,0%	57,6%	87,0%	43,0%	13,1%	100,0%	7,0%	37,6%	57,0%	28,0%	51,6%	85,0%	37,0%	46
Unused	89,5%	91,0%	88,0%	89,4%	92,0%	87,0%	89,0%	91,0%	87,0%	89,0%	91,0%	87,0%	89,5%	91,0%	88,0%	4552
Vineyards	93,0%	92,0%	94,0%	91,8%	88,0%	96,0%	92,4%	95,0%	90,0%	92,9%	90,0%	96,0%	93,5%	93,0%	94,0%	932
Average	63%	67%	63%	64%	67%	64%	57%	60%	59%	65%	70%	63%	65%	73%	62%	

Figure 31: Results per class for the clean dataset experiment

9.4.5 Confusion matrix of the focal class balanced loss model on the clean dataset

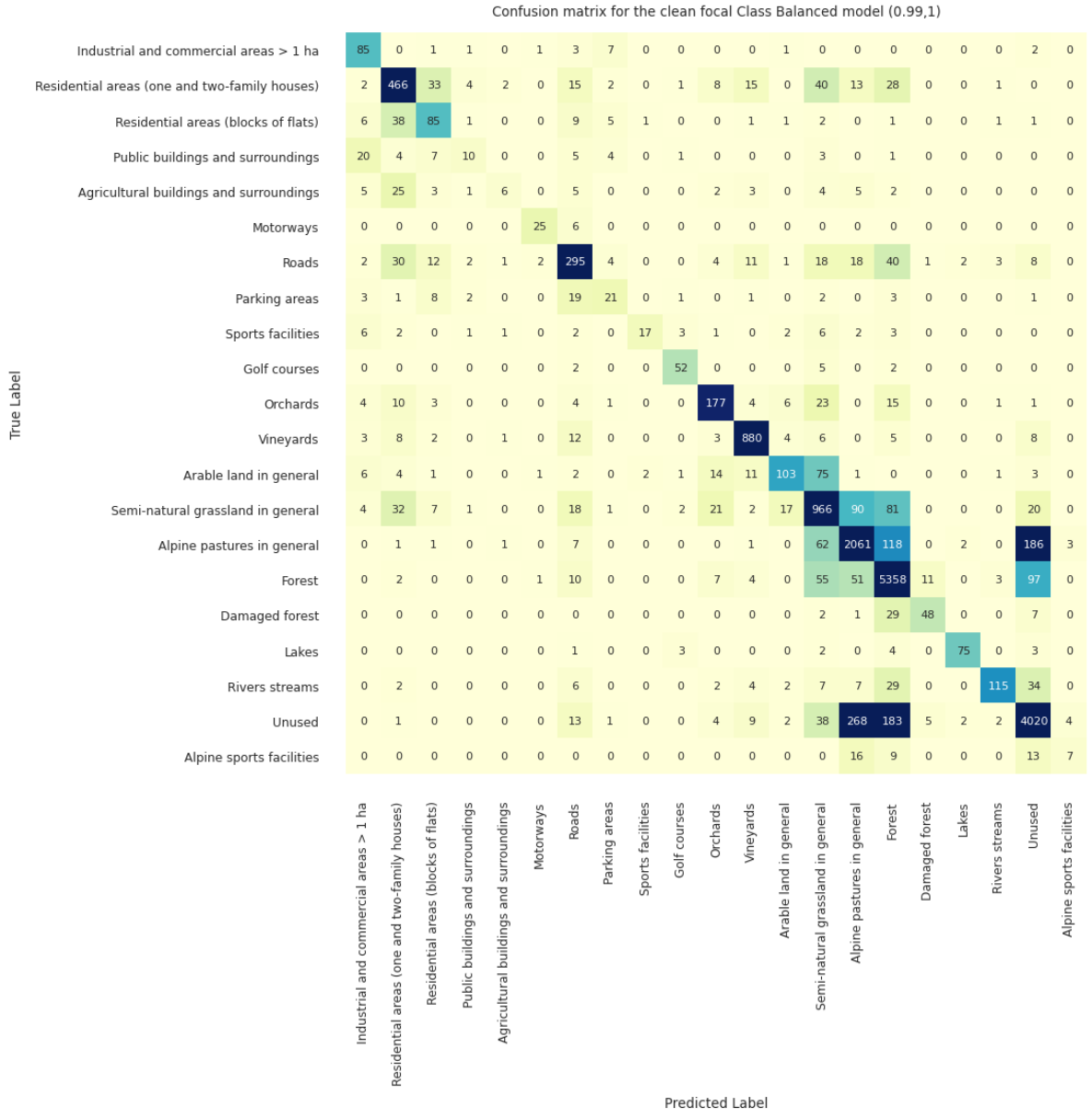


Figure 32: Confusion matrix of the focal class balanced loss model on the clean dataset

9.4.6 Confusion matrix of the baseline model on the clean dataset

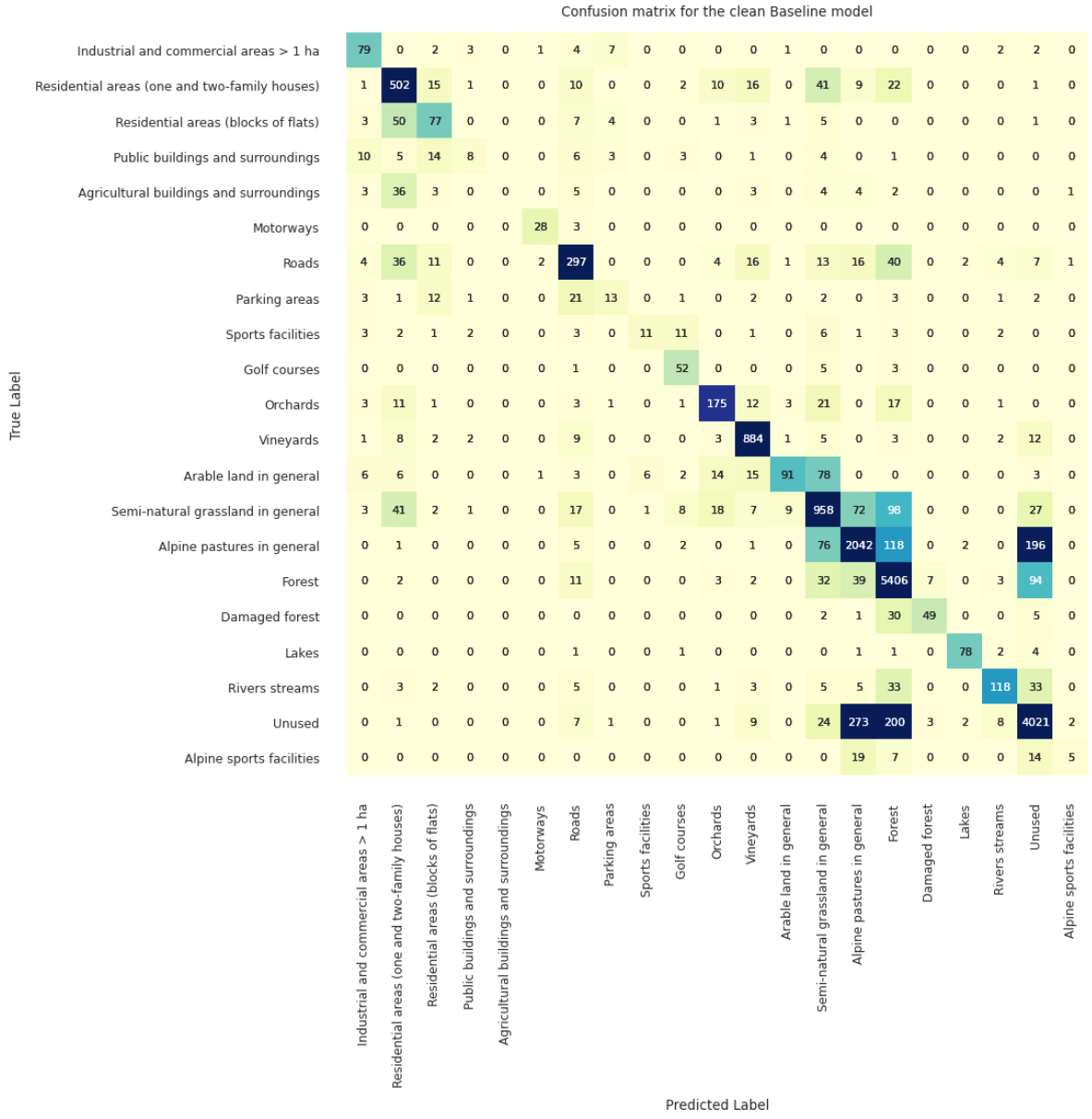


Figure 33: Confusion matrix of the baseline model on the clean dataset

9.5 Detailed results for the comparison experiment

9.5.1 Comparison on the 21 classes

	ADELE			BASELINE			Difference			test set
Rare classes	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	size
Alpine sports facilities	0%	0%	0%	0%	0%	0%	0%	0%	0%	1
Public buildings and surroundings	0%	0%	0%	0%	0%	0%	0%	0%	0%	1
Damaged forest	0%	0%	0%	0%	0%	0%	0%	0%	0%	1
Motorways	40%	33%	50%	50%	50%	50%	10%	17%	0%	2
Unexploited urban areas	0%	0%	0%	0%	0%	0%	0%	0%	0%	2
Construction sites	40%	50%	33%	0%	0%	0%	-40%	-50%	-33%	3
Sports facilities	80%	100%	67%	40%	50%	33%	-40%	-50%	-34%	3
Unspecified buildings and surroundings	0%	0%	0%	0%	0%	0%	0%	0%	0%	3
Parking areas	57%	100%	40%	22%	20%	25%	-35%	-80%	-15%	4
Industrial and commercial areas > 1 ha	100%	100%	100%	67%	60%	75%	-33%	-40%	-25%	4
Golf courses	80%	100%	67%	74%	70%	78%	-6%	-30%	11%	9
Agricultural buildings and surroundings	23%	50%	15%	14%	50%	8%	-9%	0%	-7%	13
Alpine sheep grazing pastures in general	0%	0%	0%	17%	25%	13%	17%	25%	13%	15
Residential areas (blocks of flats)	38%	45%	33%	39%	27%	73%	1%	-18%	40%	15
Orchards	67%	83%	56%	61%	67%	56%	-6%	-16%	0%	18
Alpine meadows in general	10%	100%	5%	40%	29%	63%	30%	-71%	58%	19
Rivers streams	48%	73%	36%	35%	39%	32%	-13%	-34%	-4%	22
Lakes	96%	98%	95%	92%	95%	90%	-4%	-3%	-5%	40
Arable land in general	58%	71%	49%	50%	93%	34%	-8%	22%	-15%	41
Lumbering areas	16%	71%	9%	28%	37%	23%	12%	-34%	14%	57
Average	38%	54%	33%	31%	36%	33%	-6%	-18%	0%	
Common classes										
Roads	78%	82%	75%	67%	76%	60%	-11%	-6%	-15%	121
Farm pastures in general	56%	58%	54%	61%	53%	71%	5%	-5%	17%	256
Residential areas (one and two-family houses)	83%	78%	89%	77%	78%	77%	-6%	0%	-12%	237
Semi-natural grassland in general	57%	64%	51%	63%	65%	61%	6%	1%	10%	268
Average	69%	71%	67%	67%	68%	67%	-1%	-3%	0%	
Frequent classes										
Vineyards	95%	95%	95%	90%	86%	95%	-5%	-9%	0%	62
Alpine pastures in general	80%	80%	81%	85%	88%	83%	5%	8%	2%	973
Unused	93%	93%	94%	93%	94%	93%	0%	1%	-1%	2654
Forest	92%	88%	96%	92%	91%	94%	1%	3%	-2%	1775
Average	90%	89%	92%	90%	90%	91%	0%	1%	0%	

Figure 34: Detailed results on the full test area for the comparison of ADELE and the baseline on 28 classes

References

Algan, G., & Ulusoy, I. (2021). Image classification with deep learning in the presence of noisy labels: a survey. *Knowledge-Based Systems*, *215*, 106771. https://doi.org/10.1016/j.knosys.2021.106771

Beyeler, A. (2018). *Die arealstatistik der schweiz: eine zeitreihe zur dokumentation der bodennutzung basierend auf der interpretation von stichprobenpunkten ab luftbildern.*

Buda, M., Maki, A., & Mazurowski, M. A. (2018). A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*, *106*, 249–259. https://doi.org/10.1016/j.neunet.2018.07.011

Castelluccio, M., Poggi, G., Sansone, C., & Verdoliva, L. (2015). Land use classification in remote sensing images by convolutional neural networks. *arXiv:1508.00092*.

Cheng, G., Han, J., & Lu, X. (2017). Remote sensing image scene classification: benchmark and state of the art. *Proceedings of the IEEE*, *105*(10), 1865–1883. https://doi.org/10.1109/JPROC.2017.2675998

Cui, Y., Jia, M., Lin, T.-Y., Song, Y., & Belongie, S. (2019). Class-balanced loss based on effective number of samples, 9268–9277.

Dong, Q., Gong, S., & Zhu, X. (2019). Imbalanced deep learning by minority class incremental rectification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *41*(6), 1367–1381. https://doi.org/10.1109/TPAMI.2018.2832629

Facchinetti, C. (2019a). *Arealstatistik « Deep Learning » (ADELE):Projets pilotes OFS de la stratégie d'innovation sur les données* (HLG MOS ML Project).

Facchinetti, C. (2019b). Summary: ADELE: visual interpretation of aerial images for the national land use statistics.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *The deep learning book*. MIT Press.

Havaei, M., Davy, A., Warde-Farley, D., Biard, A., Courville, A., Bengio, Y., Pal, C., Jodoin, P.-M., & Larochelle, H. (2017). Brain tumor segmentation with deep neural networks. *Medical Image Analysis*, *35*, 18–31. https://doi.org/10.1016/j.media.2016.05.004

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, *21*(9), 1263–1284. https://doi.org/10.1109/TKDE.2008.239

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition, 770–778.

Hu, F., Xia, G., Yang, W., & Zhang, L. (2018). Recent advances and opportunities in scene classification of aerial images with deep models, 4371–4374. https://doi.org/10.1109/IGARSS.2018.8518336

Hu, F., Xia, G., & Zhang, L. (2016). Deep sparse representations for land-use scene classification in remote sensing images, 192–197. https://doi.org/10.1109/ICSP.2016.7877822

Huang, C., Li, Y., Loy, C. C., & Tang, X. (2016). Learning deep representation for imbalanced classification, 5375–5384.

Ioffe, S., & Szegedy, C. (2015). Batch normalization: accelerating deep network training by reducing internal covariate shift. *arXiv:1502.03167 [cs]*.

Jordan, D., Lack, N., Hochuli, S., Meyer, A., & SCHÄR, M. (2019). Automatisierte klassifizierung der landnutzung: deep learning basierter ansatz für die arealstatistik der schweiz (Geomatik Schweiz), 260–4.

Jordan, D., Meyer, A., Lack, N., Schonholzer, M., Leiter, R., & Milani, G. (2019, October 5). *Bericht prototyp KI arealstatistik2020*. Office fédéral de la statistique (BFS).

King, J., Kishore, V., & Ranalli, F. (2016). Scene classification with convolutional neural networks. *cs231n. stanford. edu*.

Kingma, D. P., & Ba, J. (2015). Adam: a method for stochastic optimization. *Proceedings of the 3rd International Conference on Learning Representations (ICLR 2015)*.

Krawczyk, B. (2016). Learning from imbalanced data: open challenges and future directions. *Progress in Artificial Intelligence*, *5*(4), 221–232. https://doi.org/10.1007/s13748-016-0094-0

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, *60*(6), 84–90. https://doi.org/10.1145/3065386

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444. https://doi.org/10.1038/nature14539

LeCun, Y., Boser, B., Denker, J., Henderson, D., Howard, R., Hubbard, W., & Jackel, L. (1989). Handwritten digit recognition with a back-propagation network. *Advances in neural information processing systems*, *2*, 396–404.

Lin, T.-Y., Goyal, P., Girshick, R., He, H., & Dollár, P. (2018). Focal loss for dense object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *42*(2), 318–327. https://doi.org/10.1109/tpami.2018.2858826

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Zitnick, C. L. (2014). Microsoft COCO: common objects in context. In D. Fleet, T. Pajdla, B. Schiele, & T. Tuytelaars (Eds.), *Computer vision – ECCV 2014* (pp. 740–755). Springer International Publishing. https://doi.org/10.1007/978-3-319-10602-1_48

Liu, X., He, J., Yao, Y., Zhang, J., Liang, H., Wang, H., & Hong, Y. (2017). Classifying urban land use by integrating remote sensing and social media data. *International Journal of Geographical Information Science*, *31*(8), 1675–1696. https://doi.org/10.1080/13658816.2017.1324976

Liu, Z., Miao, Z., Zhan, X., Wang, J., Gong, B., & Yu, S. X. (2019). Large-scale long-tailed recognition in an open world, 2537–2546.

Lutz, N. (2019). *Deep learning für die arealstatistik der schweiz: ein multimodaler zugang mit maschinellem lernen* (Bachelor Thesis). Fachhochschule Nordwestschweiz-Institut Geomatik (IGEO).

Mahajan, D., Girshick, R., Ramanathan, V., He, K., Paluri, M., Li, Y., Bharambe, A., & van der Maaten, L. (2018). Exploring the limits of weakly supervised pretraining, 181–196.

Nogueira, K., Penatti, O. A. B., & Santos, J. A. d. (2017). Towards better exploiting convolutional neural networks for remote sensing scene classification. *Pattern Recognition*, *61*, 539–556. https://doi.org/10.1016/j.patcog.2016.07.001

OFS. (2015). Les développements économiques et sociaux transforment aussi le paysage - L'utilisation du sol en Suisse 1985-2009: exploitations et analyses | Communiqué de presse (0351-1502-30). *Office fédéral de la statistique*, 3.

OFS. (2016). *Catégories standard : nomenclature standard NOAS04* (2018th ed., Vol. 1).

OFS. (2017). *Description de données GEOSTAT, Statistique de la superficie selon nomenclature 2004 – Standard* (Office fédéral de la statistique OFS).

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., & Lerer, A. (2017). Automatic differentiation in PyTorch.

Patino, J. E., & Duque, J. C. (2013). A review of regional science applications of satellite remote sensing in urban settings. *Computers, Environment and Urban Systems*, *37*, 1–17. https://doi.org/10.1016/j.compenvurbsys.2012.06.003

Picterra, S. D. A. (2017, February 11). Picterra feasability study - arealstatistik2020.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., & Fei-Fei, L. (2015). ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, *115*(3), 211–252. https://doi.org/10.1007/s11263-015-0816-y

Schar, M., Nebiker, S., & Jordan, D. (2017, August 18). *Einsatz von deep learning zur aktualisierung der arealstatistik der schweiz - erste untersuchungen.pdf* (Bachelor Thesis). Fachhochschule Nordwestschweiz-Institut Geomatik (IGEO). Muttenz.

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

Swisstopo. (2018, March). swissALTI3d le modèle de terrain à haute résolution de la suisse.

Swisstopo. (2020a). swissALTI3d rapport sur la publication 2020.

Swisstopo. (2020b, May). SWISSIMAGE la mosaïque d'orthophotos de la suisse.

Tan, J., Wang, C., Li, B., Li, Q., Ouyang, W., Yin, C., & Yan, J. (2020). Equalization loss for long-tailed object recognition, 11662–11671. https://doi.org/https://doi.ieeecomputersociety.org/10.1109/CVPR42600.2020.01168

Ulyanov, D., Vedaldi, A., & Lempitsky, V. (2017). Instance normalization: the missing ingredient for fast stylization. *arXiv:1607.08022 [cs]*.

Van Horn, G., & Perona, P. (2017). The devil is in the tails: fine-grained classification in the wild. *arXiv:1709.01450 [cs]*.

Wang, S., Liu, W., Wu, J., Cao, L., Meng, Q., & Kennedy, P. J. (2016). Training deep neural networks on imbalanced data sets, 4368–4374. https://doi.org/10.1109/IJCNN.2016.7727770

Wang, T., Li, Y., Kang, B., Li, J., Liew, J., Tang, S., Hoi, S., & Feng, J. (2020). The devil is in classification: a simple framework for long-tail instance segmentation. *arXiv:2007.11978 [cs]*.

Wang, Y.-X., Ramanan, D., & Hebert, M. (2017). Learning to model the tail, 11.

Wu, Y., & He, K. (2018). Group normalization, 3–19. https://doi.org/10.1007/978-3-030-58568-6_43

Xia, G., Hu, J., Hu, F., Shi, B., Bai, X., Zhong, Y., Zhang, L., & Lu, X. (2017). AID: a benchmark data set for performance evaluation of aerial scene classification. *IEEE Transactions on Geoscience and Remote Sensing*, *55*(7), 3965–3981. https://doi.org/10.1109/TGRS.2017.2685945

Yang, Y., & Newsam, S. (2010). Bag-of-visual-words and spatial extensions for land-use classification. *Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems*, 270–279. https://doi.org/10.1145/1869790.1869829

Zhang, A., Lipton, Z. C., Li, M., & Smola, A. J. (2020). Dive into deep learning, 997.

Zhang, C., Sargent, I., Pan, X., Li, H., Gardiner, A., Hare, J., & Atkinson, P. M. (2019). Joint deep learning for land cover and land use classification. *Remote Sensing of Environment*, *221*, 173–187. https://doi.org/10.1016/j.rse.2018.11.014

Zhao, B., Zhong, Y., Xia, G., & Zhang, L. (2016). Dirichlet-derived multiple topic scene classification model for high spatial resolution remote sensing imagery. *IEEE Transactions on Geoscience and Remote Sensing*, *54*(4), 2108–2123. https://doi.org/10.1109/TGRS.2015.2496185

Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., & Torralba, A. (2018). Places: a 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *40*(6), 1452–1464. https://doi.org/10.1109/TPAMI.2017.2723009

Zhu, X. X., Tuia, D., Mou, L., Xia, G., Zhang, L., Xu, F., & Fraundorfer, F. (2017). Deep learning in remote sensing: a comprehensive review and list of resources. *IEEE Geoscience and Remote Sensing Magazine*, *5*(4), 8–36. https://doi.org/10.1109/MGRS.2017.2762307

Zimmermann, M. (2018). *Visual speech recognition: from traditional to deep learning frameworks* (Doctorate Thesis). EPFL.

Zou, Q., Ni, L., Zhang, T., & Wang, Q. (2015). Deep learning based feature selection for remote sensing scene classification. *IEEE Geoscience and Remote Sensing Letters*, *12*(11), 2321–2325. https://doi.org/10.1109/LGRS.2015.2475299