POLITECNICO DI MILANO
Master Degree in Computer Science and Engineering
Department of Electronics, Information and Bioengineering
School of Industrial and Information Engineering

Suspicious Objects Classification for Illegal Landfills Discovery in Remote Sensing Images

DEIB Politecnico di Milano

Supervisor: Piero Fraternali
Co-Supervisor: Rocio Nahime Torres

Master Degree Thesis by:
Samuele Moscatelli,
Matricola: 940515

Academic Year 2020-2021

Dedicated to my family

Abstract

The problem of waste management has recently gained worldwide relevance, having both an economic and social resonance in every country. One of the most concerning issues is constituted by illegal dumping, consisting of the uncontrolled discharge of waste into the environment. From the 1970s, with the further development of technology, the need for automatic procedures to monitor these types of phenomena increased over and over, bringing researchers to explore many possible options. Until now, however, it has been not possible to go beyond semi-automatic techniques, always requiring human intervention. In particular, the combination of satellite images with geographic information systems (GIS) has been examined far and wide for a long time. Even if it produced very good results, as anticipated, it has never allowed achieving total independence from human expertise. A decisive improvement towards fully automatic systems has been obtained with the adoption of Deep Learning techniques, in particular Convolutional Neural Networks (CNNs). In this context, current experiments focus on the automatic classification of images.

The goal of this research is to exploit one of the neural networks already used in this field, the ResNet50 architecture, to further extend the monitoring capabilities of automatic systems, allowing them to also account for the classification and localization of different types of objects characterizing Illegal Landfills. In particular, firstly the ResNet50 architecture has been used to solve the multilabel classification task, consisting in the recognition of the presence (even concurrently) of the considered classes in the given images. The proposed model reached an F_1 score of 81% on average on the test set. From the trained classifier, Class Activation Maps (CAMs) were produced and analyzed to identify the regions of the images belonging to the different waste types. The obtained results have been evaluated quantitatively using custom metrics based on the calculation of the Intersection over Union, and qualitatively by actually looking at the obtained bounding boxes, thus understanding the practical relevance of the achieved results.

Sommario

Il problema della gestione dei rifiuti ha recentemente acquisito rilevanza a livello mondiale, con una risonanza sia economica che sociale in ogni Paese. Una delle criticità più preoccupanti è costituita dalle discariche abusive, ovvero nello scarico incontrollato di rifiuti nell'ambiente. Dagli anni '70, con l'ulteriore sviluppo della tecnologia, la necessità di monitorare in maniera automatica questi tipi di fenomeni è aumentata, portando i ricercatori a esplorare molte possibili opzioni. Finora, però, non è stato possibile andare oltre a tecniche semiautomatiche, che richiedono sempre l'intervento umano. La combinazione delle immagini satellitari con i sistemi di informazione geografica (GIS) è stata esaminata in lungo e in largo per molto tempo. Nonostante abbia prodotto ottimi risultati, queste tecniche non hanno mai permesso di raggiungere la totale indipendenza dalle competenze umane. Un deciso miglioramento verso i sistemi completamente automatici è stato ottenuto con l'adozione di tecniche di Deep Learning, in particolare le Convolutional Neural Networks (CNN). In questo contesto, gli esperimenti attuali si concentrano sulla classificazione automatica delle immagini.

L'obiettivo di questa ricerca è sfruttare una delle reti neurali già utilizzate in questo campo (ResNet50) per estendere ulteriormente le capacità di monitoraggio di sistemi automatici, consentendo la classificazione di diverse tipologie di oggetti che caratterizzano le discariche illegali. Inizialmente l'architettura ResNet50 è stata utilizzata per risolvere la classificazione di tipo multilabel, e quindi per riconoscere la presenza (anche contemporanea) di oggetti appartenenti a diverse classi. Il modello proposto ha raggiunto un F_1 score del 81% in media sul test set. Successivamente, con il classificatore allenato sono state prodotte le Class Activation Maps (CAMs), al fine di identificare le regioni delle immagini che appartengono ai diversi tipi di rifiuti. I risultati ottenuti sono stati valutati quantitativamente utilizzando metriche personalizzate basate sul calcolo dell'Intersection over Union, e qualitativamente guardando effettivamente i box di delimitazione ottenuti.

Acknowledgments

First of all, I want to express extreme gratitude to Professor Piero Fraternali and Rocio Nahime Torres for giving me the opportunity of working with two people as rich in knowledge and experience as they are. They made me grow both professionally and personally, giving me all the support that I needed in order to succeed in this enthusiastic experience.

I want to thank my family, for their endless support throughout all my life, especially my parents Giampaolo and Annamaria. They always believed in me in every moment, giving all of themselves to make me succeed in everything that I did and allowing me to have the strength of overcoming any problem I had encountered. I learned so much from them that I will never be able to thank them enough. I admire them deeply.

A special mention is due to my grandparents Giuseppe and Mariagrazia. Thank you for being the most beautiful people that I've ever met.

I want to say special thanks to my friend and colleague Andrea for his friendship and support from the very beginning of this incredible journey. Without him, it would not have been the same.

I want to thank all my friends with whom I shared incredible experiences and with whom I grew side by side. Thank you Alessandro, Alessandro, Biram, Christian, Daniele, Davide, Edoardo, Ivan, Fabio and Simone for always pushing me to believe in my dreams, for making me live incredible adventures and for always being by my side.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Waste management is one of the most severe problems of modern times, affecting not only the environment, but also the social and economic spheres of each country, and gaining a worldwide resonance.

From the '80s onwards the development of scientific knowledge and environmental awareness, also at a legal level, have led to a significant increase in the costs of waste disposal and, consequently, to an expansion of illegal landfills, mostly managed by organized crime.

The major issue in this context is in fact given by the illegal dumping phenomenon, that is the illicit action of uncontrolled discharge of waste into the environment. This activity can be in particular restricted to the disposal of a small amount of waste, produced by a single illicit event, or it can be caused by repeatedly taken actions, covering larger-scale sites, namely **Illegal Landfills (ILs)**.

Illegal waste dumping is a matter of recent concern because of its numerous impacts and the consequent troublesome problems for both inhabitants and civic administration all over the world. In particular, the existence of ILs continues to be an unsolved issue in the developed and peripheral countries of Europe. Researches conducted in Germany, Austria (Allgaier and Stegmann [4]), Ireland (Doak et al. [21]), France (Biotto et al. [10]), Italy (Silvestri and Omri [65]), Romania (Apostol and Mihai [7]) and Serbia (Vasiljevic et al. [71]) demonstrates the severity of this problem and indicates that the analysis of ILs is complicated.

Regarding the environmental consequences, the uncontrolled disposal of waste generates fluxes of contaminants in the adjacent areas (as it is schematized in Figure 1.1). The width of the contaminated area depends on the hydrology of the site and the type of pollutant.

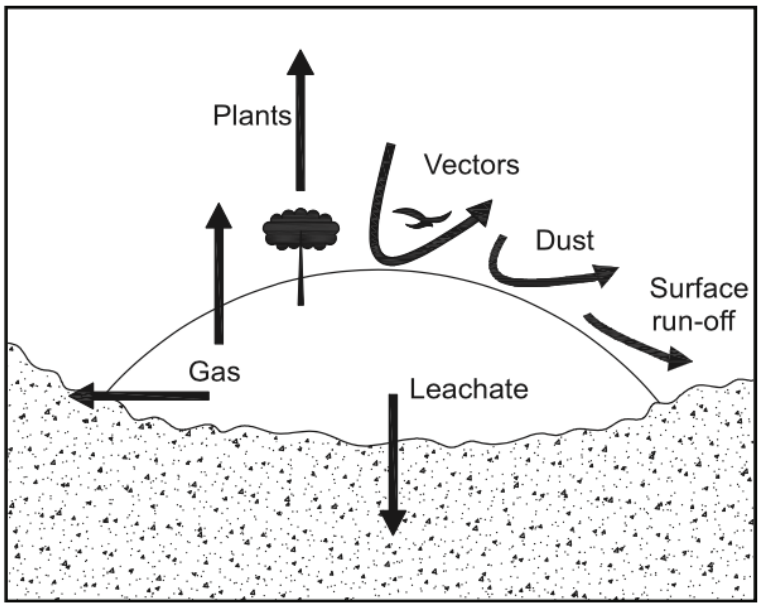

Figure 1.1: Fluxes of emission

The soluble pollutants are more likely to be carried into groundwater, spreading out into the environment easily. In particular, the liquid emissions are generated by the surface run-off and the leachate percolation, compromising the quality of groundwater and the bodies of water. The main effects of this phenomenon are acidification, eutrophication, and oxygen depletion. Moreover, the organoleptic qualities of groundwater can be compromised.

The gaseous pollutants, instead, are generated from illegal combustion and degradation processes in the uncontrolled landfill and are carried by the wind to the surrounding areas. Going deeper into details, the gaseous emissions spread out in the atmosphere or infiltrate into the ground, leading to global warming, acidification, photochemical smog, and formation of ground-level ozone. In addition, persistent pollutants bioaccumulate in animals and crops, also impacting the health of the people.

For what regards the social and health sphere, hazards are mainly linked to the toxicity of some pollutants (freed in the combustion process, released into groundwater, or accumulated in the food chain) and to the infectious diseases carried by insects and animals that can act as vectors. This is particularly dangerous for children, that are more vulnerable and exposed to pollution than adults since they spend more time outside and since their immune system and lungs are not fully developed. Frequent exposure to some chemicals, such as heavy metals and polychlorinated biphenyls, during early fetal development, can cause neurodevelopmental disorders and subclinical

brain dysfunction (Grandjean and Landrigan [30]). As a consequence, living in a community that has visible dump sites, together with the factors previously mentioned, could also wear on mental health, and thus on the quality of life itself.

Regarding instead the economic sphere, ILs have a huge impact on both citizens and governments, with a cause-effect relation. In the Lombardy region, as an example, in recent years, have been paid out 25.9 million euros to clean up 16 illegal waste sites (Affinito Domenico [2]). Waste disposal is in fact very expensive, in particular in the case of illegal landfills, which in this way results in the waste of public money and the increase of taxes. In addition to this, a country characterized by the problem of illegal waste dumping also results to be not attractive to potential buyers and investors, thus causing a decrease in the whole market. These instead results attractive from the point of view of organized criminality, which finds in illegal waste trafficking one of the most profitable sectors, with revenues in the order of 4 billion euros per year (Europol [22]).

Until the '70s the recognition of ILs made use of non-automated methods, mainly entrusted to field research. However, this methodology has been proven to be unfeasible, being very expensive and time-consuming, even more considering that areas of investigation are generally very wide.

In recent years, thanks to the advances of earth observation technology (Qiong Hu et al. [34], Luis Gomez-Chova et al. [28]), **Remote Sensing (RS)** has become more and more a valuable data source for earth observation, allowing to measure and observe detailed structures on the Earth's surface. RS, together with the application of **Geographic Information Systems (GIS)**, allowed the reduction of field research and the improvement of the detection process, making possible a better characterization of the ILs phenomenon both on a geophysical and socio-demographic level and proving the validity of the Multi-Criteria Decision Making (MCDM) approach.

Although this approach opened the way towards automated ILs detection, the differences in terms of geophysical and socio-demographic factors between different countries make the human intervention still required.

In parallel with the advances in RS technologies, in the last decades **Deep Learning (DL)** techniques, and in particular **Convolutional Neural Networks (CNNs)**, demonstrated to be particularly effective, especially when dealing with images. This type of network can be used to solve several different tasks, among these image classification, object detection, and instance segmentation.

In 2019, the regional agency for environmental protection of Lombardy

region in Italy (ARPA Lombardia) started a project called *Savager* (Advanced Waste Management Surveillance) [46] which aims to introduce Artificial Intelligence (AI) techniques in the process of ILs detection, in this way lightening the workload entrusted to the highly specialized human operators and thus significantly speeding up the whole process (see Figure 1.2).

In addition to many other organizations, this project involves the DEIB department of Politecnico di Milano, which is experimenting for this purpose with a series of methods aimed at recognizing ILs in the context of geospatial intelligence. In particular, the baseline neural network used in the *Savager* project is ResNet50.

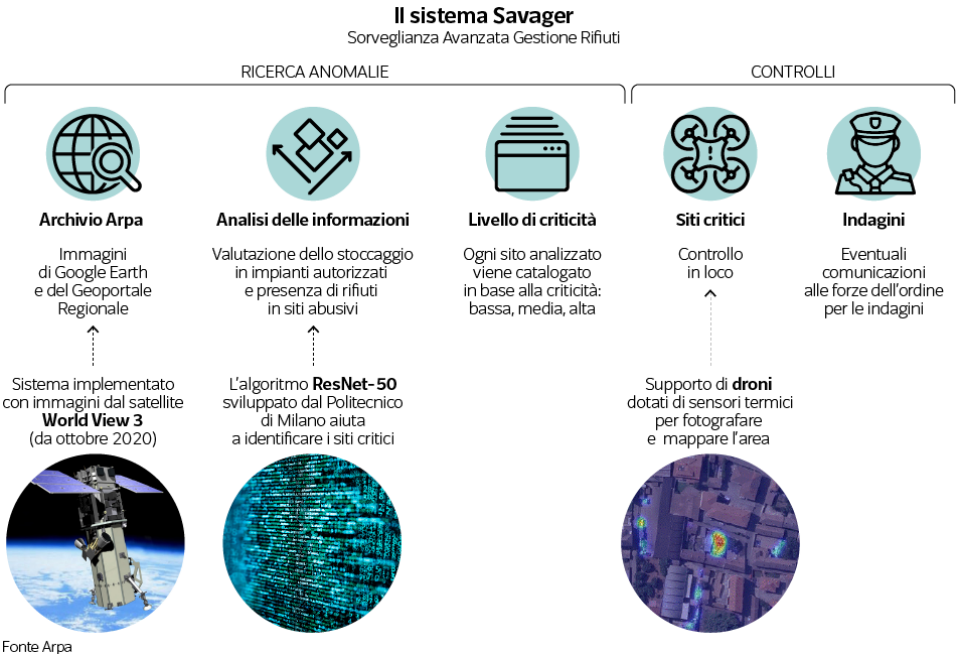

Figure 1.2: ARPA Advanced Waste Management Surveillance

As previously anticipated, Deep Learning techniques demonstrated to be very effective when dealing with images, and not only in the binary classification task but also in the recognition of a higher number of categories. Since landfills are often composed of different types of objects and materials, one of the ways that points towards further automation and refining of the waste management process is that of performing **Multilabel Classification**. This task consists of recognizing the different types of objects that are present in an image, even when there are multiple categories in the same picture at the same time.

Moreover, due to the high dimension of remote sensing images and to

the coexistence of a high number of different ground objects, another task that can reduce in a significant way the need for human intervention is that of not only recognize the different types of objects composing waste but also localize them in the image. This type of task in the context of AI is called **Object Detection**.

Classic (i.e. Supervised) Object Detection, however, requires to manually label a high number of images with not only the names of the classes included in them but also the bounding boxes enclosing the objects that make the image belonging to those classes, which is a long and tedious problem. Since Deep Learning models, if correctly tuned, are already able to learn really good features within the classification task, and since many of them, in particular way ResNet50, demonstrated to maintain a remarkable localization ability until the very final layer, an alternative is that of the **Weakly Supervised Object Detection (WSOD)**. This approach allows in fact to perform the Object Detection task by exploiting the discriminative regions that the model uses while performing the classification task, thus making the process even more independent from the human intervention.

The main goal of this study is to perform the Multilabel Classification task and to prove that the obtained class activation maps are good enough to also localize different categories of objects composing Illegal Landfills in remote sensing images, thus demonstrating that the Weakly Supervised approach of performing Object Detection is very promising. In particular, the purpose is that of designing a dataset on which the ResNet50 architecture can be trained, validated, and tested, such that it can learn good enough features to recognize the different categories of objects of interest included in an image, demonstrating that the model is using the correct discriminative regions.

Initially 12 categories have been considered, however for 6 of them it has not been possible to find a reasonable number of samples for being able to include them in the study (the process that brought to this conclusion is exposed in Appendix A). The six remaining categories faced the problem of high similitude among them. This issue was studied in depth resulting on similar classes being grouped (the process that brought to this conclusion is exposed in Appendix B). Thus, the three pairs have been merged, and so in the end three categories have been taken into consideration: Cisterns and Pallets, Scattered and General waste, and Caissons and Containers.

Once the dataset has been defined, the ResNet50 architecture has been trained, validated, and tested, and after some fine-tuning, it has been reached an F_1 score of 81% on average within the Multilabel Classification task.

The quantitative and qualitative analysis demonstrated the capability

of the model to learn relevant features. This, together with the ability of ResNet50 architecture to retain remarkable localization, has brought to the feasibility study of the Weakly Supervised Object Detection task. In particular, the bounding boxes enclosing the objects of interest have been derived through a thresholding mechanism on the Class Activation Maps (CAMs) produced by the architecture. Subsequently, both a quantitative and a qualitative study of the obtained results have been performed.

1.1 Thesis Structure

The thesis is organized as follows:

- **Chapter 2** gives an overview of the technical background on which the thesis is structured, starting from a description of the computer vision field and then entering in the details of deep learning and artificial neural networks.
- **Chapter 3** presents the researches and methods that have been used till now for ILs detection, giving an idea of the importance assumed by deep learning in this field.
- **Chapter 4** enters in the details of the proposed solution, describing the dataset that has been used to train, validate and test the model and the techniques that have been exploited in order to build it.
- **Chapter 5** describes the details of the adopted architecture, the techniques that have been applied during the process and the way in which the output is used.
- **Chapter 6** provides the quantitative and qualitative analysis of the results obtained within the multilabel classification task. In particular, a detailed analysis of the graphs and metrics is given, also showing the calibration of the model. Regarding the qualitative analysis, the performance of the model is exposed in terms of CAMs (Class Activation Maps), in terms of robustness to image resolution and in terms of predictions of completely unseen territories.
- **Chapter 7** provides the quantitative and qualitative analysis of the results obtained within the feasibility study of the weakly supervised object detection task. In particular, a detailed analysis of the graphs and metrics is given. For what regards the qualitative analysis, examples showing how the model localizes the critical objects are given,

comparing the predictions with the ground truth, and also displaying positive and negative aspects of the model performance.

- **Chapter 8** summarizes the results obtained from the experiments and proposes some possible improvements for the future.

Chapter 2

Technical Background

This chapter will describe the concepts at the basis of Deep Learning and Computer Vision, giving an articulated overview of these topics and the theory behind them.

2.1 Computer Vision

Computer Vision (CV) is the field of study that seeks to develop techniques to help computers "see" and understand the content of digital images, and, at an abstract level, the goal of computer vision is to use the observed image data to infer something about the world.

The problem of Computer Vision appears simple because it is trivially solved by people, even very young children. Nevertheless, it largely remains an unsolved problem based both on the limited understanding of biological vision and on the complexity of visual perception in a dynamic and nearly infinitely varying physical world.

As a consequence, a whole field, homonymously called Computer Vision, has emerged as a discipline in itself with strong connections to mathematics and computer science and looser connections to physics, psychology of perception, and the neurosciences.

CV was born in the early 1960s at universities, that viewed the project as a stepping stone to Artificial Intelligence. Initially, this field of study was mainly focused on geometry, as it could be seen in the very first works such as Larry Roberts' Ph.D. thesis [60]. Here, the *Block World* model is explained, which is the simplification of the visual world into basic geometric shapes to recognize and reconstruct them.

Initially, with their tremendous optimism, researchers had raised public expectations impossibly high while failing to appreciate the difficulty of the

challenge they had set for themselves. When the promised results failed to live up to the hype, the field experienced intense critique and serious financial setbacks. Early computing resources, in fact, could not keep pace technically with the complexity of problems advanced by scientists, and even the most impressive projects could solve only trivial problems. It is famous the episode during which, in 1966, American computer scientist and co-founder of the MIT AI Lab Marvin Minsky received a summer grant to hire a first-year undergraduate student, Gerald Sussman, to spend the summer linking a camera to a computer and getting the computer to describe what it saw, but "Sussman didn't make the deadline".

Thus, by the mid-1970s vision turned out to be one of the most difficult and frustrating challenges in AI, making people lose faith in CV.

As the internet became a mainstay, computer scientists gained access to more data than ever before. Also, Computing hardware continued to improve as costs went down. This, together with the birth of rudimentary neural networks and the improvement of already present algorithms, brought CV into hype again in the 1980s-90s. Of particular importance are the studies made in a seminar group of works [23, 12, 48], where different methods have been proposed to corroborate a similar idea: every object can be decomposed into simple geometric primitives. In this way, the complex structure of the object can be reduced into a collection of simpler shapes and their geometric configurations. This posed the basis to one of the most influential research fields: feature-based object recognition.

Of utmost importance has been the work made by David Lowe on SIFT features [49], who, introducing the concept of invariants in images objects, has introduced the concept of **features**. As a consequence, the task of object recognition began with identifying these critical features on the object, which is an easier task than pattern matching the entire object.

Although up to this point considerable progress has been made, the whole of these researches remains a set of toy examples, without being able of delivering tools satisfying real-world use cases.

As the new century progresses, however, the quality of images and power of resources continued growing, giving researchers the possibility to access an always wider amount of data. This, together with other factors, made machine learning gain momentum in the CV field, and the results were so good that in 2010 the ImageNet team rolled out an international challenge called ImageNet Large-Scale Visual Recognition Challenge (ILSVRC) to test the image classification results for the CV algorithms. ILSVRC is still a de facto benchmark to compare Computer Vision algorithms.

Now more than half a century old, the field of Artificial Intelligence fi-

nally had its breakthrough moment in 2012 at the ImageNet Large Scale Visual Recognition Challenge (ILSVRC), where a team from the University of Toronto entered a Convolutional Neural Network (CNN) called AlexNet that changed the game for Artificial Intelligence and Computer Vision projects, showing the vast potential of Deep Learning (DL).

Computer Vision has already been seamlessly integrated into many aspects of daily lives, and this trend shows no signs of slowing down, opening day by day new and incredibly interesting ways to go.

2.1.1 Computer Vision Tasks

Although in this thesis the main focus is on **Multilabel Classification** and on the feasibility study of the **Weakly Supervised Object Detection**, CV can be used to solve a wide range of problems. In particular, the most basic task is called Image Classification, which consists of the assignment of a label to an entire image by choosing among a fixed set of labels. Further researches, however, permitted the development of models able to go deeper in information extraction from images. In particular, the main tasks that CV can solve besides Image Classification (Figure 2.1) are:

- **Semantic Segmentation**: it consists of labeling each pixel of the input image with a corresponding class from a fixed set.

- **Image Classification with Localization**: it consists of performing the image classification task, also assigning to the image a bounding box that localizes the object. The bounding box is usually provided with the coordinates of a rectangular shape.

- **Object Detection**: it consists of performing an image classification with localization task, but with input images containing more than one object.

- **Instance Segmentation**: it consists of assigning a label to each pixel of the input image in the same way Semantic Segmentation does, but also distinguishing between different instances of the same class.

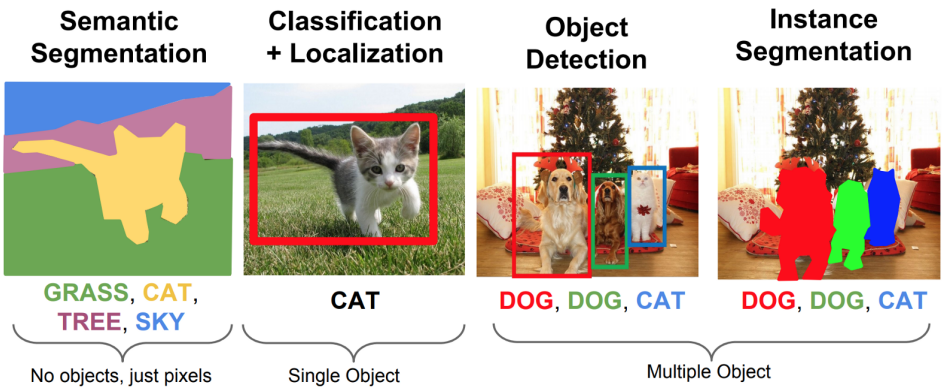

Figure 2.1: Computer Vision tasks

Multilabel Classification is a variant of the image classification task, where more than a single label from the fixed set can be assigned to each input image (Figure 2.2).

Figure 2.2: Multilabel Classification

Weakly Supervised Object Detection (Figure 2.3) is instead a variant of the object detection task, where the model is trained in the same way as when performing the Multilabel Classification task, thus with image-level category labels and without ground truth bounding boxes, and the output bounding boxes are produced using the discriminative regions of the image used for the classification task. These regions can be identified through class activation maps (CAMs).

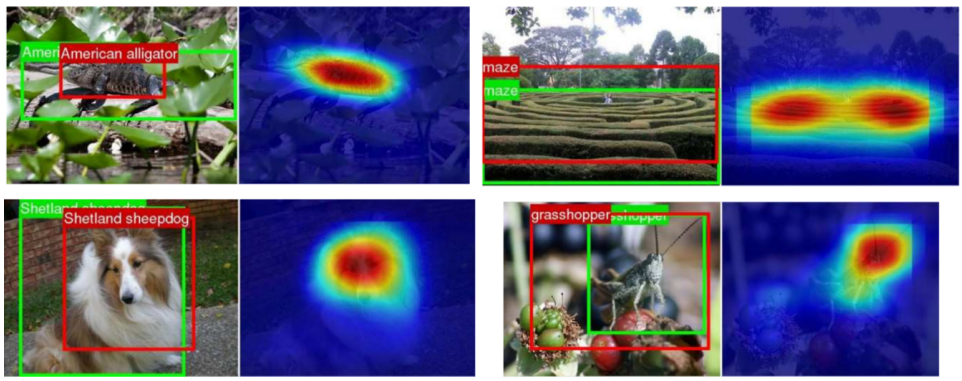

Figure 2.3: Weakly Supervised Object Detection

2.2 Deep Learning

Machine Learning (ML) falls under the Artificial Intelligence (AI) umbrella and consists of providing systems with the ability to automatically learn and improve from experience without being explicitly programmed. In particular, ML constitutes a fusion of algorithms and statistical models to analyze and draw inferences from patterns in data.

Deep Learning (DL) is a subfield of ML that uses models, structures, and algorithms inspired by the human brain. In particular, at the basis of DL, there is the concept of **Artificial Neural Network (ANN)**, a layered architecture that can be compared to the computational model of the human brain. It is in fact distributed among simple non-linear units, it is redundant and thus fault-tolerant, and it is intrinsically parallel.

The development of DL was motivated in part by the failure of traditional ML algorithms in generalizing well on central problems in AI, such as speech or object recognition. The mechanisms used in traditional Machine Learning, in fact, revealed to be insufficient, making exceedingly difficult the learning of complicated functions in high dimensional spaces, suffering from the famous phenomenon called *Course of Dimensionality*.

The main differences between traditional ML and DL can be summarized in three points:

- The model structure
- The lower need for human intervention
- The larger data requirements

The model structure has been already anticipated in the previous paragraphs and will be further detailed in the next sections. The second point refers

to the fact that ML is mainly concerned with finding the model to fit data, making the work of features engineers completely transparent. DL, instead, focuses also on the features extraction part. Deep learning algorithms in fact do not need features engineering, because it is the machine itself that learns not only the classifier but also the best representation of the data. Finally, the fact that Deep Learning models also account for feature extraction makes them very data-hungry.

2.3 Artificial Neural Networks

Artificial Neural Networks (ANNs) are computing systems inspired by the human brain. Their birth historically coincides with the birth of the **Perceptron** model by Frank Rosenblatt in 1958 [62].

ANNs are based on the concept of *Hebbian Learning*, according to which the strength of a synapse increases based on the simultaneous activation of the relative input of the desired target [33]. In particular, as the human brain, an ANN is constituted by hundreds or thousands of artificial neurons called processing units, which are interconnected by nodes and organized in different layers.

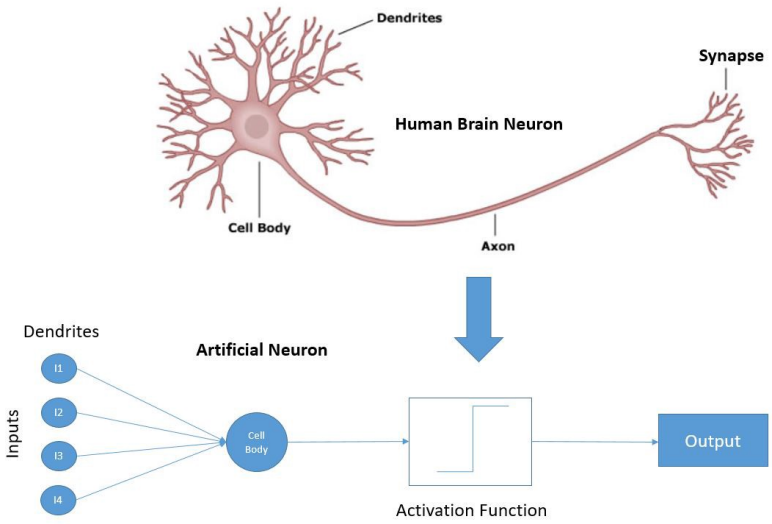

Figure 2.4: Human and artificial neuron comparison

The artificial neuron (Figure 2.4) is the basic block of ANNs. It can receive one or more inputs to compute, producing a single output that can be then sent to other neurons. Usually, each input is separately weighted, either amplifying or dampening their significance, basing on the task that

the algorithm is trying to learn. The weighted inputs are then summed and passed through the so-called activation function, which is usually non-linear, in this way determining whether and to what extent that signal should progress further through the network to affect the outcome. If the signal passes through, it is said that the neuron has been *activated*.

Connecting more of these neurons together, it is possible to extract from the input data only the information that is necessary and important for the task to be solved, acting as a sort of filter that forwards important signals and removes useless ones.

ANNs are organized into **layers**, which are sets of parallel neurons (Figure 2.5). At the highest level there are three types of layers:

- **The Input Layer**, that receives the raw input data.
- **The Hidden Layers**, which are the ones that actually perform the computation.
- **The Output Layer**, that is the one that produces the results for the given input.

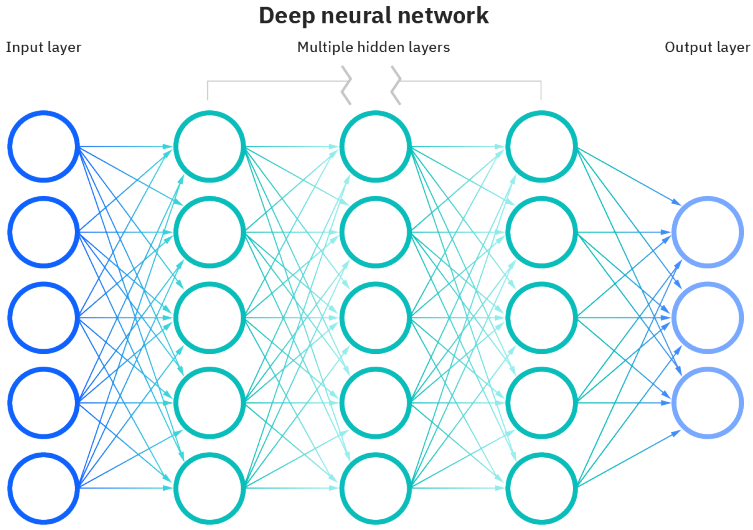

Figure 2.5: Artificial Neural Network structure

The learning process of an ANN is usually structured into two steps. The first step is called **forward pass** (Figure 2.6) and consists of providing the model with the input in the form of a vector. Data will then flow through the hidden layers, which progressively transform it until the output layer

is reached, where the model produces the desired result. This is also the procedure with which the model predicts new unseen data.

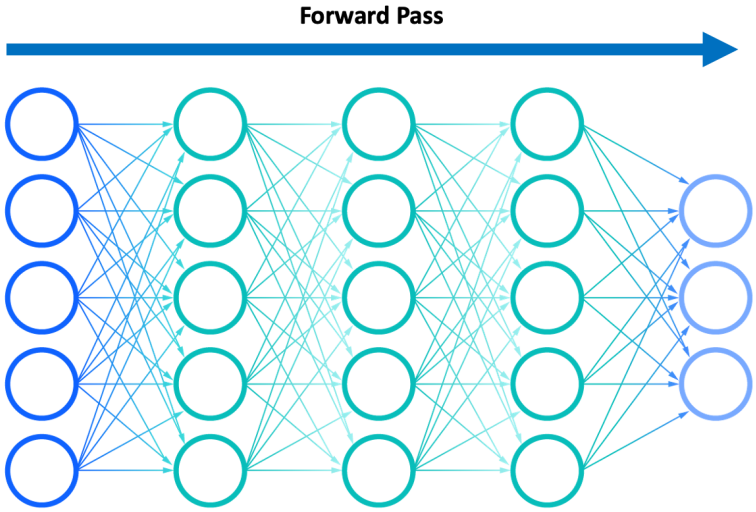

Figure 2.6: Forward pass

During the training phase, however, a second step is needed, called backward pass (Figure 2.7), during which the *Backpropagation* algorithm is performed.

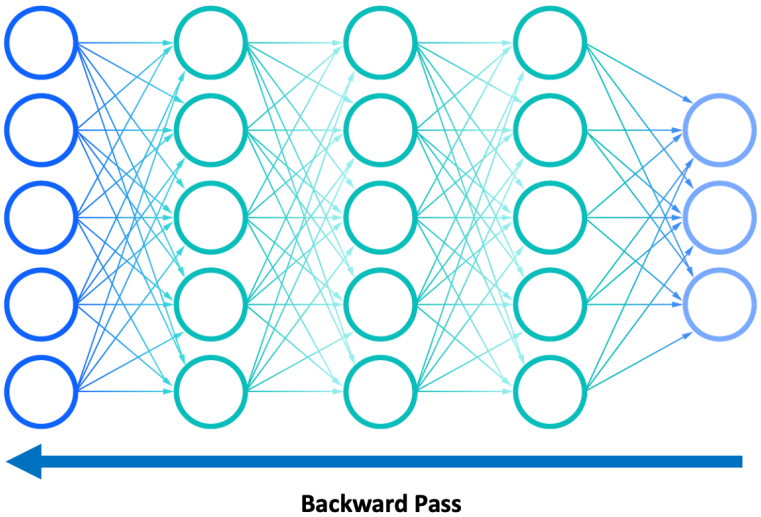

Figure 2.7: Backward pass

Backpropagation is a widely used algorithm for training feedforward neural networks. Several generalizations of it also exist for other types of ANNs and they are usually all referred to as Backpropagation. This algorithm

computes the gradient of the loss function concerning the weights of the network, doing it efficiently. In fact, it allows using gradient methods to train multilayer networks. In particular, it updates the weights to minimize the so-called **Loss Function**. The weights update is performed by another procedure, such as **Gradient Descent**, even if the term Backpropagation is often improperly used to refer to the whole process. At the basis of the Backpropagation algorithm, there is the chain rule, which allows computing the gradient one layer at a time, iterating backward from the last layer to avoid redundant calculations of intermediate terms.

2.4 Convolutional Neural Networks

Images cannot be directly fed to a classifier, but it is necessary some intermediate step to extract meaningful information and to reduce the data dimension: in other words, it is necessary to extract features (Figure 2.8).

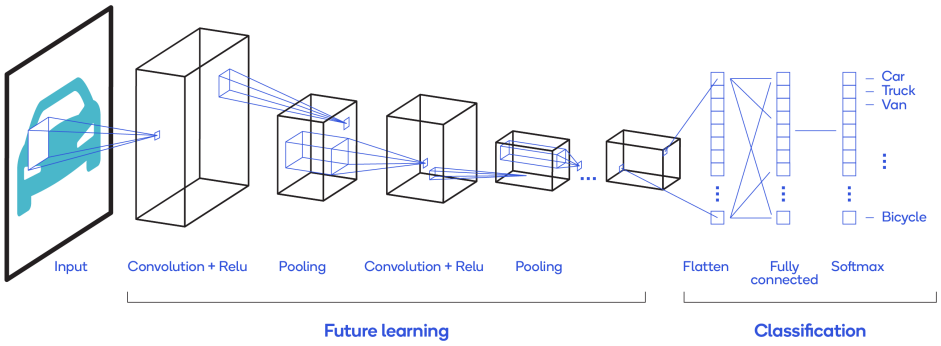

Figure 2.8: Convolutional Neural Network structure

As previously anticipated, deep learning models also account for the features extraction process, thus avoiding the need for human intervention. Images are however highly spatially correlated, thus a traditional ANN built by stacking in sequence a bunch of fully connected layers is not the right choice. The reasons are mainly two:

- In this way the model does not take into account the spatial structure of the image, because it treats all the input pixels in the same way, either if they are far or close to each other.
- Connecting the output of each neuron in a layer with the input of all the ones in the subsequent layer, the network introduces many parameters, often more than needed. In this way, the model will require a long processing time and it will be much more prone to overfitting.

These problems have been solved with *Convolutional Neural Networks (CNNs)*, an innovative architecture introduced for the first time by Yann Lecun [41] in 1998 with the LeNet model, even if the breakthrough moment for CNNs was in 2012 when AlexNet [39] outperformed all the other algorithms. The first big advantage brought by CNNs has been that of being constituted by more lightweight layers, by vastly reducing the number of parameters in the network. This makes the forward pass previously described much more efficient, thus allowing to stack a higher number of layers and so permitting the usage of deeper architecture, which demonstrated to be particularly good with images.

CNNs are typically constituted by blocks that include four different types of layers:

- The Convolutional Layers (the ones that give the name to the architecture)
- The Non-Linear Layers
- The Pooling layers
- The Fully Connected Layers

An image passing through a CNN is transformed in a sequence of so-called **volumes**. As the depth of the volume increases, the height, and the width decrease. Each layer type has a precise and well-defined purpose, but they all take as input a volume and return as output a volume.

2.4.1 Input Layer

Of great importance is the input layer, which constitutes the entry point of the model, holding the raw pixel values of the image. These pixels are organized in matrices of size (H, W, C), where H and W are the height and width of the image, while C represents the color channels (e.g. for color images $C = 3$ like the RGB channels).

2.4.2 Convolutional Layer

Convolutional layers are of central importance in CNNs, so much that they even give it its name. They "mix" all the input components making a linear combination of all the values in a region of the input.

The parameters of this layer are called filters (or also kernels) and they are used through the whole spatial extent of the input. Each filter must have the same depth as one of the input volumes, and this is the reason

why it is usually not specified. Filters have in fact a small spatial extent, but a large depth extent, and each one of them yields a different slice in the output volume. The output is also called **activation maps** and it is the convolution against the filters of the input volume (Figure 2.9).

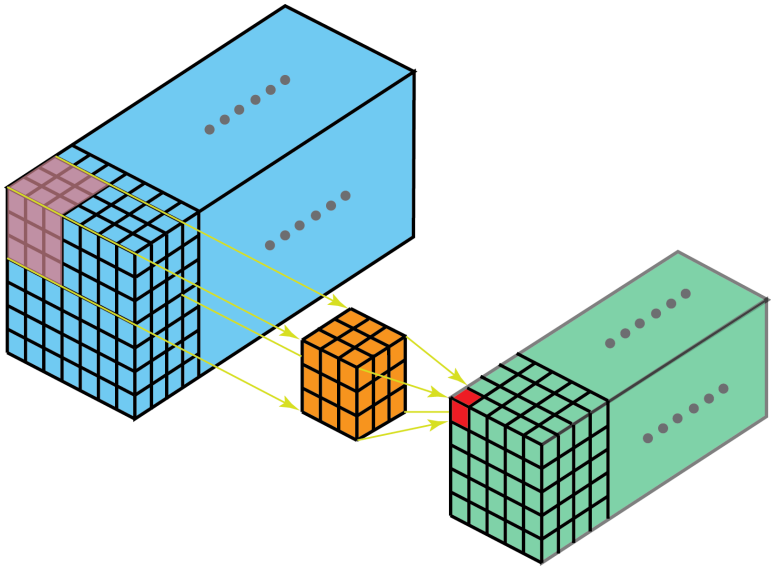

Figure 2.9: Convolution operation

CNNs, in particular, have two characteristics that make them particularly efficient when working with images:

- Sparse Connectivity
- Parameter Sharing

For what regards sparse connectivity, as previously anticipated, the input pixels are not connected to every hidden neuron like in fully connected layers. The connections are made in small, localized regions of the input volume called **local receptive fields**.

Parameter sharing clearly defines the fact that the same parameters are shared across different locations of the input image, thus vastly reducing the number of parameters involved in a CNN. In particular, it relies on the assumption that if a patch feature is useful to compute at some spatial position, then it should also be useful to compute at other positions.

The size of the output volume depends on four parameters: the number of filters, the size of the filters, the stride, and the padding. As previously anticipated, each filter yields a different slice of the output volume, thus its depth is defined by the number of filters applied by the layer. The size

of the filters, the stride, and the padding parameters define together the spatial dimension of the output volume. The second one is the step size with which the filter slides onto the input volume, while the third one refers to the number of pixels added to an input image, such that the border of the images are taken into consideration with the same frequency as the other parts.

Defining the width of the input image as W_{in}, the size of the filters as F_s, the stride as S and the padding as P, the width of the output volume can be calculated as

$$W_{out} = \frac{W_{in} - F_s + 2P}{S} + 1$$

Thus, finally, if the input volume has dimensions $(W_{in} \times W_{in} \times D)$, the dimensions of the output volume will be $(W_{out} \times W_{out} \times F_d)$, with F_d equal to the number of filters.

2.4.3 Non-Linear Layer

Also known as Activation Layers, they have the purpose of introducing non-linearity in the network, otherwise, the CNN might be equivalent to a linear classifier.

They are usually placed right after the convolutional layer and they do not modify the dimensions of the volume. There are many different types of activation functions and, while for all the hidden layers the same function is usually adopted, in the output layer it is important to choose the right one in order to achieve the result that is wanted. The most popular activation functions (Figure 2.10) are the following:

- Linear activation function
- Tanh activation function
- Sigmoid activation function
- ReLu activation function

The most commonly used in the hidden layers is the ReLu activation function (or variants of it) because other non-linear functions such as Sigmoid or Tanh tend to saturate. They have gradients that are close to zero and since the Backpropagation algorithm requires gradient multiplication, the gradient far away from the output vanishes, and consequently the learning in deep networks does not happen.

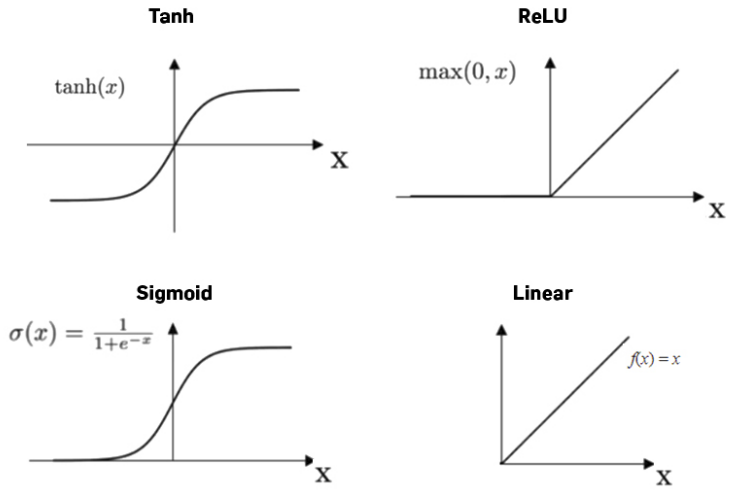

Figure 2.10: Activation functions

2.4.4 Pooling Layer

The pooling layers are usually placed after convolutional layers (and so after non-linear layers). Their purpose is that of reducing the spatial size of the volume, operating independently on every depth slice of the input volume.

Going deeper into details, pooling layers solve an important problem related to convolutional layers. A limitation of the feature map output of convolutional layers is in fact that they record the precise position of features in the input. This means that small movements in the position of the feature in the input image will result in a different feature map. A common approach to address this problem from signal processing is called downsampling, which consists of creating a lower resolution version of the input image, still maintaining the large or important structural elements. Downsampling can be achieved with convolutional layers by changing the stride of the convolution across the image, but a more robust and common approach is to use a pooling layer. This procedure adds to the structure of a CNN another fundamental property called equivariance to translations, which allows the net to recognize features even if they are in different positions.

The pooling operation is specified, rather than learned, this means that this type of layer does not have parameters. In particular, the most commonly used operation is the **Maximum Pooling (or Max Pooling)**, which calculates the maximum value for each patch of the feature map (Figure 2.11).

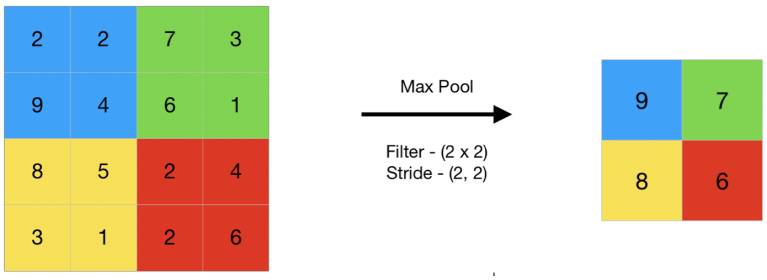

Figure 2.11: Maximum Pooling operation

As done in the case of convolutional layers, also for pooling layers it is possible to determine a function of the input volume size W_{in}, of the filter size F_s and the stride S, to determine the dimension of the output volume:

$$W_{out} = \frac{W_{in} - F_s}{S} + 1$$

An important variant of the pooling layer is defined as **Global Average Pooling (GAP)** layer, which can be used for two principal purposes. On one side, it is a structural regularizer. In fact, being used in place of the final fully connected layer, vastly reduces the number of parameters in the network, making the model less prone to overfitting. It also gives to the model the capacity of handling input images of different sizes, making it more robust to spatial transformations of the images. On the other side, instead, using GAP the network can retain a remarkable localization ability until the final layer, allowing to easily identify the discriminative image regions leading to a prediction. This can be done with the so-called **Class Activation Mapping (CAM)**.

2.4.5 Fully Connected Layer

Finally, a classical CNN is constituted by one or more fully connected layers as output layers. The main purpose is that of compiling data extracted by previous layers to form the final output. In particular, in fully connected layers, as previously anticipated, each neuron is connected to all the neurons in the previous layer.

2.4.6 Other Layers

The layers previously presented are the typical ones in CNNs, but two other types of layers can be used to handle different problems:

- **The Dropout Layer**

- **The Batch Normalization Layer**

The first one randomly sets input units to zero according to a probability defined as a hyperparameter of the model. This is a stochastic regularization technique that results to be particularly useful to mitigate overfitting.

Batch Normalization layers instead can be interpreted as doing preprocessing at every layer of the network, but integrated into the network itself in a differentiable way. This has been shown to improve the gradient flow through the network, to allow using higher learning rates (thus allowing faster learning), to reduce the strong dependence on weights initialization, and to act as a form of regularization slightly reducing the need for dropout.

2.4.7 Class Activation Maps (CAMs)

As previously introduced, it is possible to see where the model focuses its attention when producing the predictions. In particular, it is possible to obtain the heat map representation of an image to highlight the pixels that trigger the trained model to associate the image with a particular class. This representation is called Class Activation Map (CAM), and in practice consists of a matrix having the same dimensions of the input image, and containing a floating-point number for each pixel. Higher values correspond to regions that mostly influenced the model's output predictions. An example is shown in Figure 2.12.

Going deeper into details, CAMs can be obtained through the usage of a Global Average Pooling layer at the end of the network, followed by a fully connected layer and subsequently by a softmax layer. The steps to be taken are the following:

1 Get all the weights connected between the fully connected layer and the softmax class for which it is wanted to predict.

2 Assuming that the number of feature maps passed to the GAP layer is n, for each class take the n feature maps, multiply them serially with the correspondent weight, and finally add them.

The size of the heatmap is the size of the feature map, but it is possible to scale it to the dimensions of the input image thus obtaining the result shown in Figure 2.12.

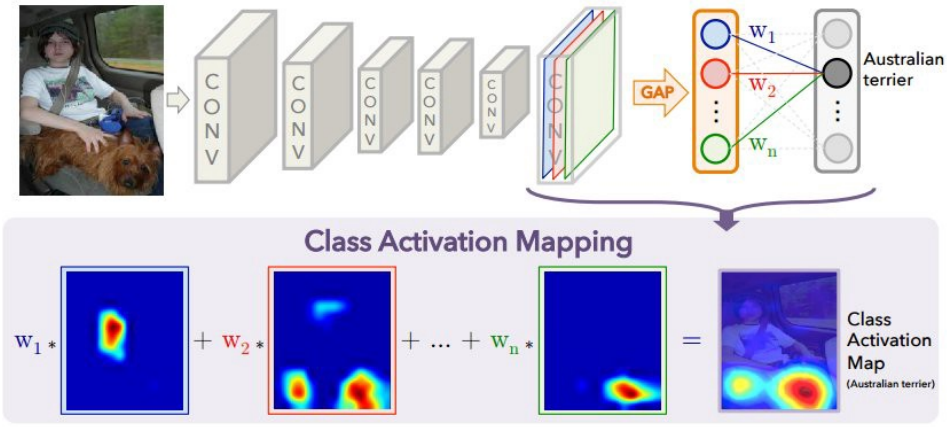

Figure 2.12: Class Activation Maps

The two main limitations of CAMs are related to the fact that the usage of the Global Average Pooling layer is required, and they can only be applied to the final layer, not to the ones before.

2.5 ResNet

Increasing the network depth by stacking an increasing number of layers does not always improve performance. This was demonstrated by He et al. [32], whose main investigation was to verify if it is possible to continuously improve accuracy by stacking more and more layers.

As the depth of the network increases, the accuracy saturates and then degrades rapidly. This however is not due to overfitting, since the same trend is seen on training error, while in the case of overfitting, training and test errors should diverge. Deeper models are in fact harder to optimize than shallower models.

In principle, coping the parameters of the shallow network in the deeper one, and then, in the remaining part, setting the weights to yield an identity mapping should not impair the performance of the deeper model. The obtained network in fact should be as good as the shallower ones. The problem is that learning the identity is not easy.

The solution that has been proposed is **residual learning**, which is based on the concept of **residual block** (Figure 2.13), introduced with the so-called **ResNet** model (Figure 2.14).

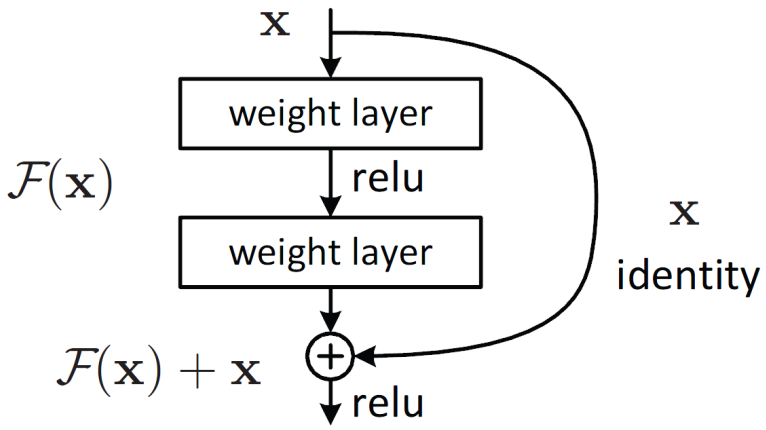

Figure 2.13: The residual block

The idea is that of building a model repeating this type of block which consists of a series of layers with a skip connection that adds the input of the block to its output, also called *identity mapping*. The rationale behind adding this identity mapping is that it is easier for the following layers to learn on top of the input value, mitigating the **vanishing gradient** problem and enabling deeper architectures. In fact, it does not add parameters and in case there are excess blocks their weights will go to zero and the information will be propagated by the identity. This overcomes the difficulty that the weights between the identity mapping have in practice in learning the identity function. The performance achieved by ResNet suggests that probably most of the deep layers have to be close to the identity.

Going deeper into details of residual learning, the intuition is that of forcing the network to learn a different task in each block. If $H(x)$ is the ideal mapping to be learned from a plain network, by using skip connections the network is forced to learn $F(x) = H(x) - x$, where x is the input of the residual block. In particular, $F(x)$ is called the **residual**, something to add on top of identity to improve over the solution that can be achieved by a shallower network.

Typically, at the end of this architecture there are no fully connected layers, but a Global Average Pooling (GAP) layer is used, which fed the final softmax function.

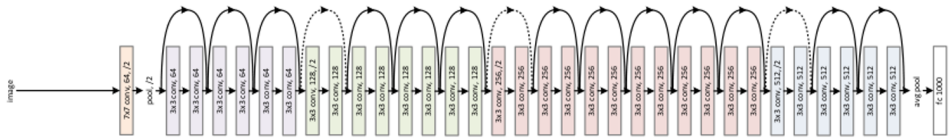

Figure 2.14: ResNet50 architecture

2.6 Data Augmentation

Deep learning models are well known to be very data-hungry, and this is a widespread problem because often the amount of available data is not sufficient. One of the most used solutions that accounts for this problem is called **Data Augmentation**, which consists of increasing the amount of data by adding slightly modified copies of already existing data. When dealing with images, the modifications usually consists of transformations like scaling, cropping, flipping, padding, rotation, and translation, that introduce some variability through which the model could learn better features.

2.6.1 RandAugment

A special type of augmentation technique is the **RandAugment** [17], which is an automated data augmentation method that takes an image and two integers N and M, where N is the number of transformations to be applied and M is the magnitude with which they have to be applied (an example is shown in Figure 2.15). With just these parameters, it can generate drastically different images and improve learners. In this way more random data are introduced into the dataset, making the probability of learning trends in noise very low.

This technique is particularly useful because it allows improving the generalization power of the model, especially in the context of remote sensing images, that are characterized by:

- Big intraclass diversity
- High interclass similarity
- Large variance of object/scene scales
- Coexistence of multiple ground objects.

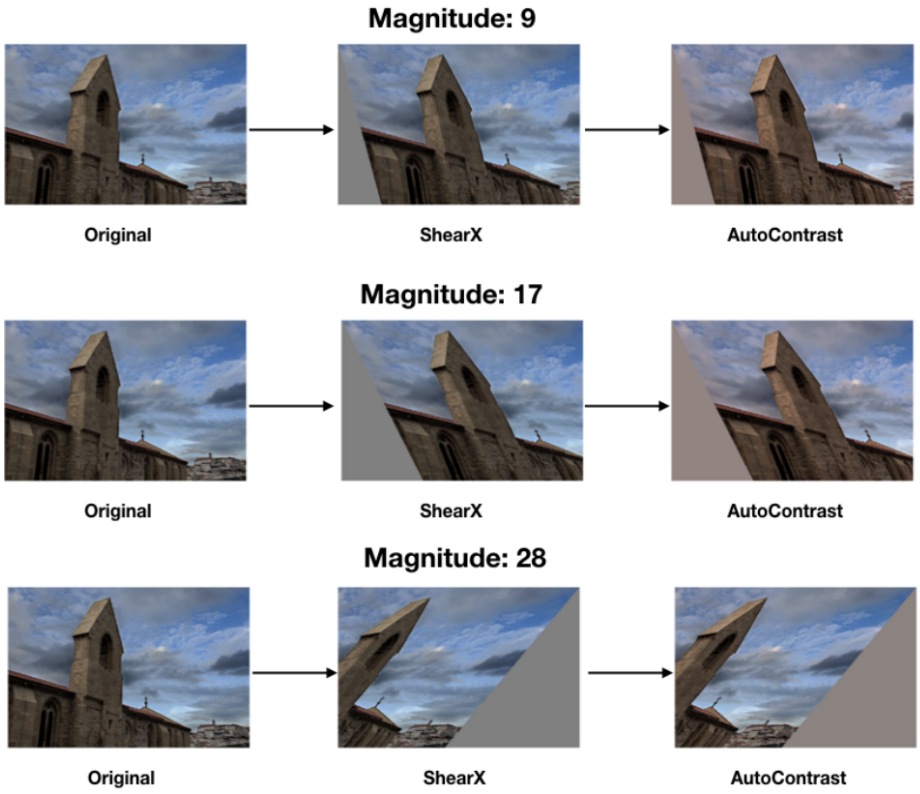

Figure 2.15: Transformation examples

Chapter 3

Related Work

This chapter provides an overview of the techniques that have been used in order to classify and detect Illegal Landfills. In particular, the main focus will be on the most used Deep Learning and Computer Vision methods, giving a background to the scope of this thesis.

3.1 Illegal Landfills Detection

Illegal landfills and uncontrolled waste combustion have always been a severe problem, which is assuming more and more importance in recent years. For this reason, research, even in completely different fields, has been tried to find optimal and automatic ways for many years.

As anticipated, ILs affect negatively the groundwater quality: the soluble pollutants leach freely into the ground and accumulate in the aquifer, severely affecting water's properties [45]. As a consequence, illegal landfills can cause such a significant change in water conductivity, that it can be detected by using coils-based sensors [61]. Thus, the surrounding environment can be harmed by eutrophication and acidification, and moreover, some pollutants are toxic to living beings and can be accumulated into the food chain.

Studies, [8, 72] have also demonstrated that many factors make the waste sector vulnerable to waste crime. These factors include:

- New legislation and its weak regulatory enforcement.
- The economics of waste treatment, where legal and safe treatment of waste can be more expensive than illegal operations.
- The complexity of the waste sector and the different actors who can

have some involvement, directly or indirectly, in the movement of illegal wastes.

- The possibility to hide or disguise waste, which creates an opportunity for illegal businesses to operate alongside legitimate waste operators.

Thus, another path that has been traveled to detect ILs is that of estimating the probability of occurrence of illegal landfills, basing on the concurrent presence of more than one of the cited factors [58, 38, 50], also ranking the possibility of site expansion [59].

A decisive improvement towards fully automatic systems has been obtained with the adoption of deep learning techniques, in particular Convolutional Neural Networks (CNNs). These types of techniques demonstrated in fact of being very effective when dealing with images. Initially, however, the detection focused on images at the ground level, thus still requiring a high level of human intervention. An example is given by A. Dabholkar et al [18], who used AlexNet and GoogleNet as reference models to distinguish six classes of waste: chair, mattresses, table, sofa, furniture, and trash, from ground-level images. An accuracy higher than 70% for all the classes, with a maximum of 95%, has been reached.

Another example is given by Alfarrarjeh et al [3], who performed image classification to determine the level of cleanliness of streets. In this case, in particular, ground-level images have been classified through a geospatial classification approach in order to enhance classification accuracy. An F_1 score of around 0.9 has been obtained.

In order to achieve a higher level of automation, multiple techniques are emerged to also directly detect illegal landfills from remote sensing images (RS). At the beginning, however, RS images were characterized by a very low resolution and thus researches mainly focused on the usage of geographic information systems (GIS), exploiting thermal images [26, 42], brightness surface distribution [68], multi-temporal analysis [20, 19], vegetation indices [13] or also multi-features algorithms [5, 27, 63, 68, 52, 70, 56, 40, 69, 65].

3.2 Deep Learning in Remote sensing

Remote sensing is a valuable data source for Earth observations and can help to measure and observe detailed structures on the planet's surface. In recent years, more and more research is done to find ways that fully exploit the huge potential of RS images [24, 43], also because the advances in technology [35, 29] allowed fast growth in the number of RS data.

Most of the researches done in the last decades has been driven by real-world applications. In particular, most of them focused on scene classification, thus in the task of correctly labeling RS images with predefined semantic categories. This has been in fact applied to a wide variety of tasks, such as urban planning [47, 66], natural hazards detection [53, 51], environment monitoring [36, 25], vegetation mapping [44, 54], and geospatial object detection [16, 15].

Initially, the spatial resolution of RS images was very low, and thus, the size of a pixel was similar to the size of the objects of interest. For this reason, the early researchers were mainly concerned with pixel-level (or subpixel-level) classification [37, 67], labeling each pixel in the remote sensing images with a semantic class.

With the passing of years, however, and the advance of remote sensing imaging, the spatial resolution of this type of images became increasingly finer than common objects of interest, and thus, single pixels lost their semantic meanings. In such a case, recognizing scene images at the pixel level solely demonstrated to be not so efficient [11], and so per-pixel analysis began to be viewed with increasing dissatisfaction. Consequently, the idea that analyzing remote sensing images at the object level is more powerful than per-pixel analysis begun to spread all over, leading to a series of approaches focused at the object level, that dominated remote sensing images analysis for the last two decades.

With the improvement of RS image quality, Deep Learning techniques started spreading also in this field, allowing a higher level of automation. The achieved results demonstrated to be so good that there has been an evolution and a convergence towards Deep Learning methods through time [14]. Initially in fact, when remote sensing images had a quite low resolution, the vast majority of researches were accomplished at the pixel level. Even if there are still researches operating at this granularity [64], the improvement of spatial resolution of remote sensing images and the impressive results obtained with the usage of deep learning have conveyed most of them towards object-level and scene-level detection.

Due to this evolution, which has highlighted the importance of analyzing remote sensing images at the object level, and also thanks to the advances in deep learning theory, deep learning-based algorithms have increasingly prevailed in the area of remote sensing image classification, being able to best exploit the increased availability of remote sensing data and parallel computing resources. This also because deep learning demonstrated to correctly face the main challenges given by object and scene classification on remote sensing images, that are: big intraclass diversity, high interclass sim-

ilarity (also known as low between-class separability), the large variance of object/scene scales and coexistence of multiple ground objects.

Thus, deep learning automatic features extraction has prevailed and numerous deep learning-based classification algorithms have emerged and yielded the best classification accuracy. In particular, they can be grouped into three categories: autoencoder-based methods, CNN-based methods, and GAN-based methods.

For the scope of this thesis, a CNN-based method has been adopted because the development of autoencoder-based methods demonstrated to have reached a bottleneck and the performance of GAN-based methods is relatively low, while CNN's-based methods still dominate and have some upside potential.

An example of what can be achieved with the usage of CNNs in the remote sensing field is given by Biserka Petrovska et al [57], who proposed a two-stream deep architecture, extracting features with a pre-trained CNN and, after feature concatenation, classifying the features using a support vector machine (SVM). The obtained classification accuracies reach a peak of 98.92%, demonstrating that the considered method has competitive results.

Impressive results have been reached also in the automatic classification of crosswalks. Berriel et al [9] obtained an accuracy of 97.11% using free available crowdsourcing data.

This study combines the potential of Deep Learning techniques with the always increasing resolution and availability of Remote Sensing images in order to solve the multilabel classification of objects constituting Illegal Landfills. Thus, differently than previous researches, the recognition is no longer performed at the ground level, but based on satellite images, which allows achieving a higher grade of automation, also being able to inspect larger areas at once.

3.2.1 Weakly Supervised Object Detection

Supervised object detection techniques can achieve astonishing results in many fields of research, and recently this type of task has gained relevance also in the context of automatic Illegal Landfills detection, focusing in particular on the usage of remote sensing images, helped by the growth in quality and quantity of data. An example is given by Shynggys Abdukhamet [1], who proposed a modified version of the state-of-the-art deep learning architecture called RetinaNet in order to perform landfills detection from satellite images. In particular, he used a novel loss function named Focal loss, which solves the class imbalance problem by down weighting the background class,

thus letting the model focus on the regions of interest. Using DenseNet as a backbone, it has been reached an average precision of 84.7%.

Researches in this direction are however still very few because Artificial Neural Networks and Deep Learning, in general, are incredibly data-hungry, thus depending on massive sets of hand-labeled training data. Data are very costly to be obtained, mainly because of the time and human effort required to label them. This is especially valid in the case of supervised object detection, where besides the name of the class it is necessary to also provide the ground truth bounding boxes that precisely enclose the objects of interest. Moreover, in technical fields like the one of ILs detection, domain expertise is also required, thus further increasing the costs both in terms of time and money.

The technique that allows to significantly reduce these costs is called Weakly Supervised Object Detection, which is recently gaining momentum since it is also able to achieve results that are comparable to those obtained with the supervised way. This approach consists in fact in exploiting the regions of the image that the model used to perform the discrimination in order to predict the exact location of the objects of interest, thus not requiring ground truth bounding boxes, but only the labels of the objects contained in the image, as in a generic classification task.

Due to the fact that both the Weakly Supervised Object Detection technique and the quality and amount of remote sensing images are grown only in the last years, there are no precedent studies that tried to combine them in order to perform Illegal Landfills detection. One of the research that is close to this has been made by Mohd Anjum and M. Sarosh Umar [6], who however solved the task at the ground level.

This study also evaluates the possibility of adopting the weakly supervised approach to perform landfills objects detection. In particular, it proves how promising and beneficial is this direction, demonstrating that CNNs can retain a remarkable localization ability while solving the Illegal Landfills multilabel classification task.

Chapter 4

Dataset

This chapter describes the dataset on which the proposed neural network has been trained, validated, and tested. In particular, it specifies how the dataset has been created, the preprocessing that have been performed to obtain the final configuration, and the classes that have been considered.

4.1 Description

The scope of this thesis is that of solving the multilabel classification task on remote sensing images, also proving the feasibility of the weakly supervised object detection to classify and localize dumped wastes characterizing ILs in the Lombardy region.

At the very beginning, twelve categories have been taken into consideration: Cisterns, Scattered waste, Pallets, Caissons, General waste, Containers, Hay bales, Tubes, Wood, Tires, Grouped cars, and Plastic bags.

Images regarding these categories have been extracted from two different data sources: Mapbox and Orthophoto. Mapbox is a location data platform that powers the maps and location services used in many popular applications. Orthophoto images are instead aerial photographs geometrically corrected for topographic relief, lens distortion, and camera tilt. With reference to the tasks to be solved, the difference that mainly concerns the two sources is the change in terms of resolution. Orthophoto images are in fact way higher in resolution, reaching also a pixel size of 0.08 meters, while Mapbox images can arrive at a pixel size in the range of [0.6, 0.3] meters.

In Figure 4.1 it is possible to see an example of an image for each category (in this case Mapbox images have been used).

Figure 4.1: Categories examples - Starting from top left: (1) Cisterns,(2) Scattered waste, (3) Pallets, (4) Caissons, (5) General waste, (6) Containers, (7) Hay bales, (8) Tubes, (9) Wood, (10) Tires, (11) Grouped cars, (12) Plastic bags

The images needed to build the dataset, and thus needed to train and fine-tune the model, have been annotated using a tagger tool [55] that allows to both define the segmentation and the class of the objects in images.

In Table 4.1 it is possible to see the number of objects that have been labeled for each category and each source.

	Ortophoto	*Mapbox*	*Total*
Cisterns	246	241	487
Scattered waste	1179	1061	2240
Pallets	457	503	960
Caissons	555	502	1057
General waste	623	750	1373
Containers	140	127	267
Hay bales	3	152	155
Tubes	8	93	101
Wood	0	113	113
Tires	0	27	27
Grouped cars	0	20	20
Plastic bags	0	47	47

Table 4.1: Initial dataset table

As it can be noticed, the number of samples of the last six categories is much smaller with respect to the first six ones, and deeply studying the performance that the model can reach on those categories with so few images, it has emerged that the amount of available data is not sufficient to be able to include these last six categories into the performed study (the whole procedure that has brought to this conclusion is exposed in Appendix A).

Another important aspect that has been found to affect in a relevant way the performance of the model is the clear similarity between some of the considered classes. In particular, the reference is to the similarity between the following three pairs of categories:

- Cisterns and Pallets
- Scattered waste and General waste
- Caissons and Containers

In Figure 4.2 is reported an example of similarity between the Cisterns and Pallets classes on Mapbox images. The picture on the left could seem to belong to the Pallets class but it belongs to the Cisterns one. Vice versa, the picture on the right belongs to the Pallets class but could seem to belong to the Cisterns one.

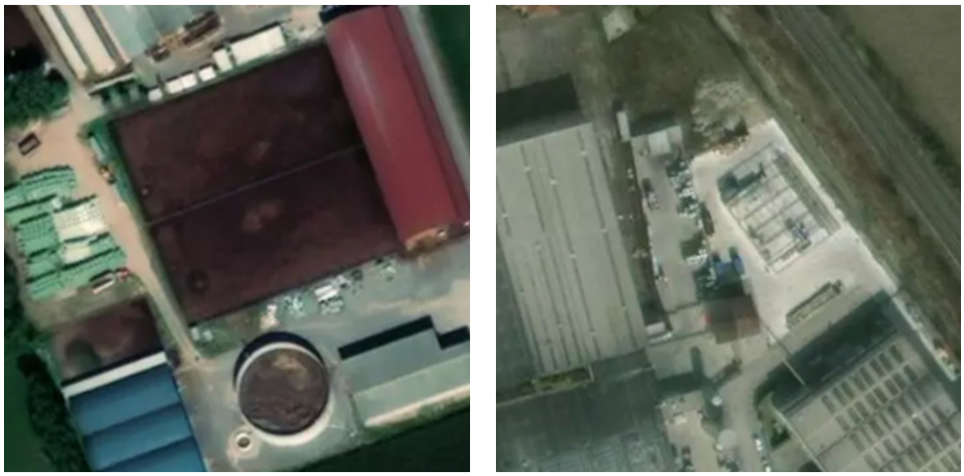

Figure 4.2: Similarity examples - Cisterns (left) and Pallets (right) from Mapbox

Analogously, the same situation is encountered in Figure 4.3, but on Orthophoto images.

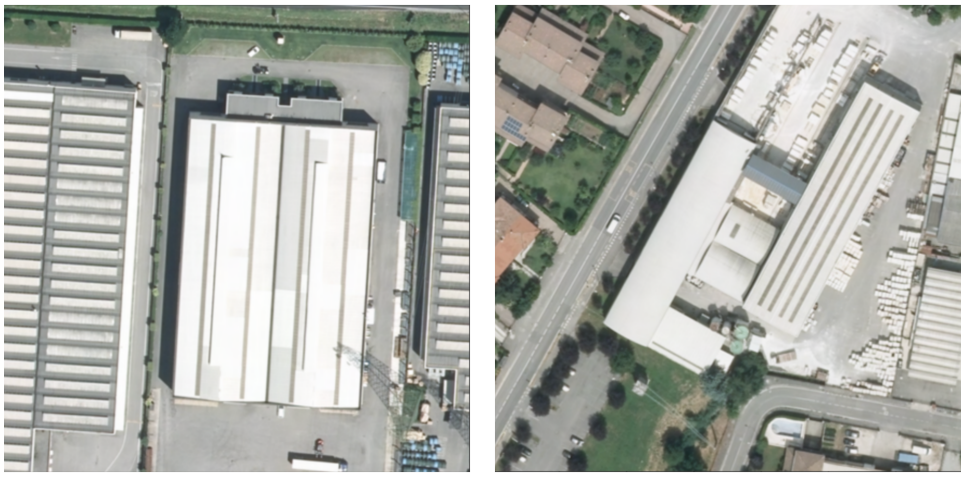

Figure 4.3: Similarity examples - Cisterns (left) and Pallets (right) from Orthophoto

In Figure 4.4, instead, it is possible to see on the left a Mapbox image belonging to the Scattered waste class that could seem to belong to the General waste class, while on the right a Mapbox image that could seem to belong to the Scattered waste category but that belongs to the General waste one.

Figure 4.4: Similarity examples - Scattered waste (left) and General waste (right) from Mapbox

Figure 4.5 shows the same situation on Orthophoto images.

Figure 4.5: Similarity examples - Scattered waste (left) and General waste (right) from Orthophoto

Regarding the Caissons and Containers classes, Figures 4.6 and 4.7 both show on the left an image belonging to the Caissons class but that could seem to belong to the Containers one, and on the right an image belonging to the Containers category but that could seem to belong to the Caissons one. Respectively, Figure 4.6 reports images taken from Mapbox, while Figure 4.7 shows Orthophoto pictures.

Figure 4.6: Similarity examples - Caissons (left) and Containers (right) from Mapbox

Figure 4.7: Similarity examples - Caissons (left) and Containers (right) from Orthophoto

Also in this case a very detailed quantitative analysis of the problem has been carried out, ending up with the conclusion that keeping them separate does not make sense (the whole procedure that has brought to this conclusion is exposed in Appendix B). For this reason, it has been proceeded with their merging.

Thus, in the end, removing the last six classes because of the too-small number of available samples and merging the three pairs of similar categories, three classes have been taken into consideration:

- Cisterns and Pallets
- Scattered and General waste
- Caissons and Containers

Thus, the dataset that has been built and used in order to train, validate and test the chosen model is composed of 3164 total images, of which 1346 positives and 1818 negatives. In particular, the three sets have been organized as follows:

- **Training:** 2233 images, of which 594 for class Cisterns and Pallets, 635 for class Scattered and General waste, 592 for class Caissons and Containers, and 1300 for the negative class
- **Validation:** 471 images, of which 127 for class Cisterns and Pallets, 148 for class Scattered and General waste, 140 for class Caissons and Containers, and 265 for the negative class

- **Testing:** 460 images, of which 138 for class Cisterns and Pallets, 145 for class Scattered and General waste, 127 for class Caissons and Containers, and 253 for the negative class

Going deeper into details, Table 4.2 shows how images are distributed between the different categories.

	Ortophoto	*Mapbox*	*Total*
Cisterns and Pallets	477	382	859
Scattered and General waste	513	415	928
Caissons and Containers	476	383	859
Negatives	1718	100	1818
Total		3164	

Table 4.2: Dataset table

As it can be seen effort has been invested in keeping the positive categories as balanced as possible in order to put the model in the condition of being able to learn equally the features of each class. Unbalanced datasets bring in fact to unbalanced results, with the model predicting with higher frequency and confidence the classes characterized by a larger number of samples. In particular, initially oversampling has been performed to exploit all the available data, and thus the maximum number of samples has been reached in each category. This however demonstrated to introduce overfitting and thus the opposite direction has been chosen, undersampling the categories with a higher number of images to the one with the minimum amount. The Scattered and General waste class has a number of samples a bit higher with respect than the others (despite the undersampling), because it is the most frequently occurring one, and thus it is often present concurrently with the other two categories.

For what regards negatives, the optimal configuration is that of having a number of samples that is almost doubled with respect to the amount of samples for a single positive class.

All the images contained in the dataset have a dimension of 700×700. It has been used such size because it revealed to be the optimal trade-off between too large images, where objects result to be too little to be recognized, and too small images, which do not allow the model to be able

to generalize well on higher-dimensional images. In particular, besides the mentioned one, the following dimensions have been tried: 300×300, 500×500 and 1000×1000.

Also, the RandAugment technique, explained in Section 2.6.1, has been adopted. Moreover, in order to introduce even more randomness, the flip technique has been used, that consists in randomly flipping the image, thus generating images with different orientation.

RandAugment has been performed five times, thus trying, on the one side, to introduce as much randomness as possible to boost generalization, and on the other to avoid the incurring of overfitting.

Chapter 5

Architecture Definition and Output Usage

This chapter provides a deep overview of the architecture that has been used in order to perform the task. In particular, it specifies how classification has been executed, describing which is the output of the model and how it has been used. Analogously, it is also explained the approach with which the feasibility study of weakly supervised object detection has been performed, illustrating what is the required output and its usage.

5.1 Training Environment

Before starting with the description of the architecture, it is important to make a small digression in order to describe the environment in which the model has been developed.

The programming language adopted is Python 3 and it has been used to develop training, evaluation, and visualization scripts. In addition, the PyTorch libraries are used to develop the structure of the previously described models. Furthermore, two NVIDIA GeForce RTX 2080 Ti GPUs were used.

5.2 Training Hyperparameters

In order to further understand how the model has been built and trained it is also important to outline some basic concepts regarding the architecture's parameters and processes.

Batch Size (BS)

Number of samples composing the set used in each iteration of the training process in order to update the gradient.

Early Stopping

Technique for the limitation of the overfitting phenomenon consisting of using a hold-out set (the validation set) to assess the training process. It is an online estimate of the generalization error, which allows stopping the training procedure when the validation error is at its minimum. In particular, if the online validation error does not improve within a specific number of training epochs (defined as patience), the process is stopped.

Early Stopping Minimum Delta

Hyperparameter related to the early stopping procedure defining the minimum variation in the monitored quantity that can be qualified as an improvement.

Early Stopping Patience

Hyperparameter related to the early stopping procedure that defines the number of training epochs to wait for validation error improvement. If the validation error does not improve within the number of epochs specified by the patience, the training procedure is stopped.

Epoch

Full training pass over the entire dataset. It is organized in a number of iterations equal to the total number of samples in the training set divided by the batch size.

Number of Epochs

Amount of epochs through which the model has to be trained. In other words, it defines the number of times the model has to pass through the whole dataset in order to end the training process.

Learning rate

Hyperparameter defining the step size at each iteration while moving towards the minimum of the loss function. It is often in the range between 0 and 1.

5.3 Architecture

The architecture that has been adopted in order to perform both the tasks is the Resnet architecture, whose theoretical relevance has been widely discussed in Section 2.5 within Chapter 2. Here the aim is that of going deeper in the description of the practical implementation of the model, depicting the number and types of layers that constitute it and the values of the hyperparameters that have been used.

5.3.1 Layers

The specific version that has been chosen is the ResNet50 which is a widely used architecture with 48 Convolutional layers along with 1 MaxPooling layer and 1 Average Pooling layer, characterized by 3.8×10^9 floating-point operations.

Going deeper into details, as it can also be seen in Table 5.1 the considered network is constituted by the following layers:

- A convolutional layer with 64 kernels of size 7×7, with a stride of size 2 and a padding of 3.
- A max-pooling layer with a stride of size 2, followed by a block of three convolutional layers with respectively 64 1×1, 64 3×3 and 256 1×1 kernels. This block is repeated three times, thus giving 9 layers in total.
- A block of three convolutional layers with respectively 128 1×1, 128 3×3 and 512 1×1 kernels. This block is repeated four times, thus giving 12 layers in total.
- A block of three convolutional layers with respectively 256 1×1, 256 3×3 and 1024 1×1 kernels. This block is repeated six times, thus giving 18 layers in total.
- A block of three convolutional layers with respectively 512 1×1, 512 3×3 and 2048 1×1 kernels. This block is repeated three times, thus giving 9 layers in total.
- A convolutional layer with 3 kernels of size 1×1 for outputting the classification predictions. In parallel, a global average pooling layer is followed by a convolutional layer with 3 kernels of size 1×1 for outputting the class activation maps.

Layer name	*Parameters*
Convolution 1	64 kernels 7×7, stride 2
Convolution 2	3×3 max pool, stride 2 $\begin{bmatrix} 1 \times 1, 64 \\ 3 \times 3, 64 \\ 1 \times 1, 256 \end{bmatrix} \times 3$
Convolution 3	$\begin{bmatrix} 1 \times 1, 128 \\ 3 \times 3, 128 \\ 1 \times 1, 512 \end{bmatrix} \times 4$
Convolution 4	$\begin{bmatrix} 1 \times 1, 256 \\ 3 \times 3, 256 \\ 1 \times 1, 1024 \end{bmatrix} \times 6$
Convolution 5	$\begin{bmatrix} 1 \times 1, 512 \\ 3 \times 3, 512 \\ 1 \times 1, 2048 \end{bmatrix} \times 3$
Convcolution 6	global average pooling 3 kernels 1×1, stride 1

Table 5.1: Resnet50 layers

Thus at the end a Deep Convolutional Neural Network of $1 + 9 + 12 + 18 + 9 + 1 = 50$ layers has been constituted.

Transfer learning has been adopted and, in particular, the weights of a ResNet50 model pre-trained on the ImageNet dataset have been loaded through Pytorch. This choice has been done due to the benefits that transfer learning technique brings, that are mainly four:

- It gives a good weight initialization and a better starting point for the training of the model.
- It allows a higher learning rate during the training process.
- It allows to converge to a higher performance level, enabling a more accurate output.

- It allows to achieve the desired performance faster than traditional learning methods since it leverages a pre-trained model.

Moreover, batch normalization has been exploited to improve the gradient flow through the network, to allow using higher learning rates (thus allowing faster learning), to reduce the strong dependence on weights initialization, and to act as a form of regularization slightly reducing the need for dropout. Batch Normalization layers in fact can be interpreted as doing preprocessing at every layer of the network, but integrated into the network itself in a differentiable way.

5.3.2 Hyperparameters

For what regards the hyperparameters described in Section 5.2, some of them have been fine-tuned in order to select optimal settings, while others have been fixed a priori. Table 5.2 shows all the hyperparameters that have been used and the respective values that have been chosen.

Hyperparameter	*Value*
Batch size	8
Number of epochs	80
Early stopping minimum delta	0.005
Learning rate	0.005
Early stopping patience	10

Table 5.2: Hyperparameter values - The first group contains the ones that have been defined a priori, while the second the ones that have been fine-tuned

The Batch Size (BS) has been set a priori since it has been found that 8 is a good trade-off between having a small BS that requires less GPU memory, but which is more affected by noise, and a larger BS that gives a more accurate estimate of the error gradient, but which requires a larger availability of GPU memory.

Regarding the number of epochs, it has been chosen 80 because on one side allows a good number of iterations during the training process, letting the model analyze the dataset for a reasonable amount of times. On the other side, it also allows the early stopping procedure to complete its work, blocking the training only when the model is going towards overfitting.

Analogously, setting 0.005 as early stopping minimum delta is a reasonable middle ground between too small or too large variations in the validation error.

The Learning Rate has been fine-tuned adopting a grid search technique, exploring three different orders of magnitude: values in the around of 10^{-2}, of 10^{-3} and of 10^{-4}. The first range immediately revealed to offer too large values, because, in fact, the model demonstrated to be not able of converging to a minimum point during gradient descent, not being able to reach an acceptable accuracy. On the contrary, for values in the order of magnitude of 10^{-4}, the learning process became too slow, not allowing to reach good solutions in a reasonable time. The range of 10^{-3} instead, and in particular 0.005, revealed to be associated with the steepest drop in the loss in the smaller time, thus this value has been adopted.

Finally, the early stopping patience has been set to 10 because it is the best trade-off between a too large value that does not allow the early stopping procedure to actually limit overfitting when the model is at its minimum and a too-small value that stops the training process too early. In particular, the following values have been taken into consideration:

$$\{5, 7, 10, 12, 15\}.$$

5.3.3 Model Output

Substantially, when an image is received as input by the model, it is traduced into two different outputs (Figure 5.1):

- The classification scores
- The class activation maps (CAMs)

The first is used within the multilabel classification task and consists of an array containing floating-point numbers. Going deeper into details, the length of the array must be equal to the total amount of classes that have been defined (three in the scope of this thesis), and each floating-point quantity corresponds to a single category, constituting the prediction of the model for the input image to belong to that category (expressed as a probability).

Once the classification scores have been generated, thresholding is performed to establish if the input image has been predicted or not to belong to a specific class (0.5 has been used for each category). Being a multilabel classification task, each image can belong concurrently to more than one class, and thus the predictions are independent of each other.

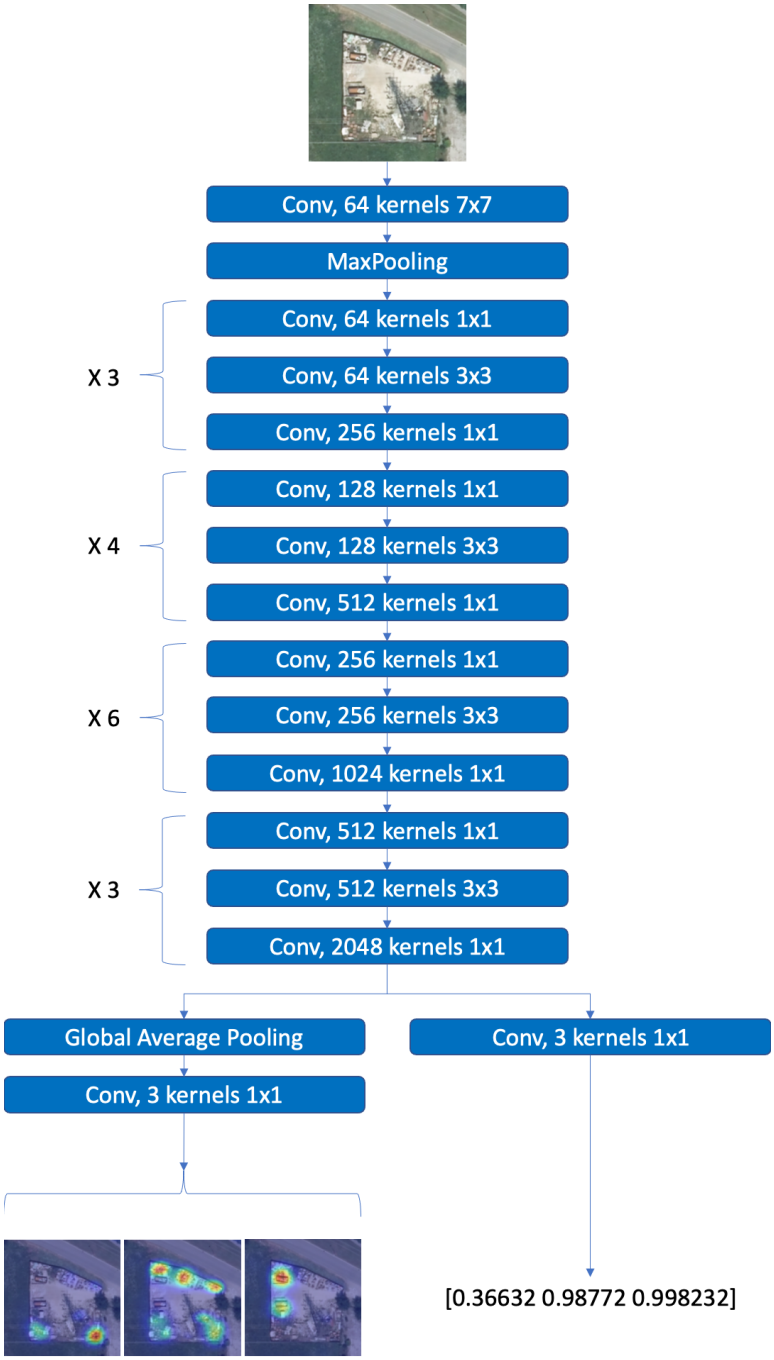

Figure 5.1: Architecture

The class activation maps (CAMs) are obtained thanks to the usage of a global average pooling layer (GAP) at the end of the architecture. In this way the network can in fact retain a remarkable localization ability until

the final layer, thus allowing to identify the discriminative image regions leading to a prediction. In particular, a CAM consists of a 2-dimensional array containing a floating-point value in the range $[0, 1]$ for each pixel in the input image. The higher is the value, the higher is the importance assumed by that pixel in the discrimination process performed by the model. In this case, however, values are not independent of each other. The classification, in fact, is performed at the object level, and thus they are highly correlated, generating clusters with higher values that correspond to the discriminative regions. Also in this case a class activation map is generated for each one of the considered categories.

This second type of output has been exploited to perform the second task in the scope of this thesis, the feasibility study of the weakly supervised object detection. In fact, in case the classification score for a specific class is high, the discriminative regions identified by the correspondent CAM coincide with the locations in which, according to the model, objects of that class are placed. In particular, selecting in advance a proper threshold value as before, it is possible to generate both the bounding boxes and the segmentation localizing the objects of interest.

Chapter 6

Evaluation - Multilabel Classification

This chapter provides a quantitative and qualitative analysis of the results obtained while solving the multilabel classification task by applying the methods and the techniques exposed in the previous chapter.

In particular, for what regards the quantitative analysis, metrics will be shown more in detail, and the model calibration will be investigated.

Regarding the qualitative analysis, firstly some examples of the model's CAMs will be shown, thus having an idea of which features it learns, and subsequently the robustness of the model will be demonstrated through the analysis of the difference in predictions between Mapbox and Orthophoto images.

Before starting with the actual evaluation, however, to better illustrate and frame the analysis, an explanation of the metrics that have been used to guide the study is given.

6.1 Metrics

In order to give a better understanding of how the performance of the model has been measured and improved, in this section the list of all the main metrics is provided, giving both the related definition and formula. In particular, here the focus is on the metrics used in the multilabel classification task.

True Positives (TP)

Amount of samples that have been correctly predicted as belonging to a positive class. In the considered case for example, if an image actually contains cisterns and the model predicts Cisterns and Pallets in that image, then this prediction is counted as a true positive.

True Negatives (TN)

Amount of samples that have been correctly predicted as belonging to the negative class. In the considered case for example, if an image does not contain any object belonging to the three considered categories and the model predicts it as negative, then this prediction is counted as a true negative.

False Positives (FP)

Amount of samples that have been wrongly predicted as belonging to a certain positive class. In the considered case for example, if an image does not contain caissons or containers, but the model predicts it as belonging to the Caissons and Containers class then this prediction is counted as false positive.

False Negatives (FN)

Amount of samples that have been wrongly predicted as belonging to the negative class. In the considered case for example, if an image contains caissons or containers, but the model predicts it as negative, then this prediction is counted as a false negative.

Basing on the previous concept it is possible to define the following metrics:

Accuracy (Acc)

Fraction of predictions that the model correctly predicts on the total number of predictions.

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

Precision (Prec)

Fraction of positive predictions that the model correctly identified out of all the samples that it has predicted as positives. Thus the frequency with which the model is correct when predicts the positive class.

$$Prec = \frac{TP}{TP + FP}$$

Recall (Rec)

Fraction of positive predictions that the model correctly identified out of all the samples that are actually positives.

$$Rec = \frac{TP}{TP + FN}$$

F1 Score (F_1)

Harmonic mean of Precision and Recall. It allows to measure how balanced are the Precision and the Recall of the models, favoring the ones with a high balance between them.

$$F_1 = \frac{2}{Prec + Rec} = 2 \cdot \frac{Prec \cdot Rec}{Prec + Rec} = \frac{TP}{TP + \frac{FN+FP}{2}}$$

Precision-Recall Curve

Plot of the tradeoff between precision and recall for different thresholds. A high area under the curve represents both high recall and high precision, where high precision relates to a low false-positive rate, and high recall relates to a low false-negative rate. High scores for both show that the classifier is returning accurate results (high precision), as well as returning a majority of all positive results (high recall).

Average Precision (AP)

Summary of the Precision-Recall curve as the weighted mean of precisions achieved at each threshold, with the increase in recall from the previous threshold used as the weight.

$$AP = \sum_n (Rec_n - Rec_{n-1}) Prec_n$$

where $Prec_n$ and Rec_n are respectively the Precision and the Recall at the n^{th} threshold.

Reliability Diagram

Plot of the actual distribution of the samples with respect to the prediction probability. In other words, it measures the calibration of the model verifying that the prediction probability reflects the actual frequency with which those samples belong to a positive class. To be more concrete, among all the images that have been predicted with a probability of 70% as belonging to the Cisterns and Pallets class, 70% of them should actually belong to that class.

Confidence Histogram

Plot of the distribution of the predictions in specific intervals. In other words, for each considered interval it tells how many samples have been predicted with a probability contained in that interval. In this case, the [0, 1] range has been divided uniformly into ten intervals, reflecting the same division that has been considered in the Reliability Diagram.

Expected Calibration Error (ECE)

Scalar summary statistic of calibration. It is computed by partitioning predictions into M equally-spaced bins (similar to the reliability diagram) and taking a weighted average of the bins' accuracy/confidence difference.

$$ECE = \sum_{m=1}^{M} \frac{|B_m|}{n} |acc(B_m) - conf(B_m)|$$

where n is the number of samples. The difference between the accuracy and the confidence for a given bin represents the calibration gap.

Maximum Calibration Error (MCE)

Measure of the worst-case deviation between confidence and accuracy. Similar to ECE, this approximation involves binning.

$$MCE = \max_{m \in \{1,\dots,M\}} |acc(B_m) - conf(B_m)|$$

It is possible to visualize both ECE and MCE on Reliability Diagrams, where MCE is the largest calibration gap (red and orange bars) across all bins whereas ECE is a weighted average of all gaps.

6.2 Quantitative Analysis

In this section a quantitative analysis of the results achieved within the multilabel classification task is given, also comparing it with the case in which the six categories have been considered separately, thus without the merging.

6.2.1 Results

In order to go deeper in the analysis of the results achieved in the multilabel classification task, in Table 6.1 are shown the metrics obtained by the chosen model on the validation and test set.

	Validation Set				Test Set			
	Accuracy	*Precision*	*Recall*	*F1-score*	*Accuracy*	*Precision*	*Recall*	*F1-score*
Cisterns and Pallets	0.90	0.79	0.89	0.84	0.88	0.82	0.78	0.80
Scattered and General waste	0.91	0.82	0.93	0.87	0.88	0.77	0.91	0.83
Caissons and Containers	0.90	0.79	0.89	0.84	0.88	0.75	0.83	0.79
Micro Average	0.90	0.80	0.91	0.85	0.88	0.78	0.84	0.81

Table 6.1: Metrics of the chosen model

Here the micro average value has been used to show the overall performance of the model because it is more suited when it is needed to pay attention to class imbalance. In particular, it is measured aggregating the contribution of all the classes, weighting each instance or prediction equally. As it can be seen, it is possible to state that the classes are balanced and this allows to also have balanced Precision and Recall metrics.

For the sake of making a comparison, Table 6.2 shows the metrics achieved by the model on the validation and test sets, considering the six categories separated, and so without the merging. As it can be noticed, in both the validation and test set of the comparative model the metrics are really swinging, not granting a real balance between the Precision and the Recall. This is mainly due to the confusion that the similarity between the categories introduces, thus further validating the choice that has been done. This swinging trend has been in fact encountered in every other experiment performed, except for the one is done with the merged categories, which, as it can be seen from Table 6.1, brings to really high and balanced metrics, overall reaching an F_1 score of 81% on the test set.

	Validation Set				Test Set			
	Accuracy	*Precision*	*Recall*	*F1-score*	*Accuracy*	*Precision*	*Recall*	*F1-score*
Cisterns	0.91	0.73	0.20	0.32	0.89	0.53	0.28	0.36
Scattered waste	0.81	0.60	0.53	0.57	0.80	0.62	0.48	0.54
Pallets	0.85	0.71	0.34	0.46	0.82	0.70	0.28	0.40
Caissons	0.81	0.55	0.57	0.56	0.79	0.60	0.54	0.57
General waste	0.82	0.65	0.59	0.62	0.81	0.60	0.61	0.60
Containers	0.95	0.73	0.25	0.37	0.94	0.67	0.17	0.27
Micro Average	0.86	0.62	0.47	0.54	0.84	0.61	0.44	0.51

Table 6.2: Metrics of the comparative model

The same consideration can be done observing the graphs in Figure 6.1.

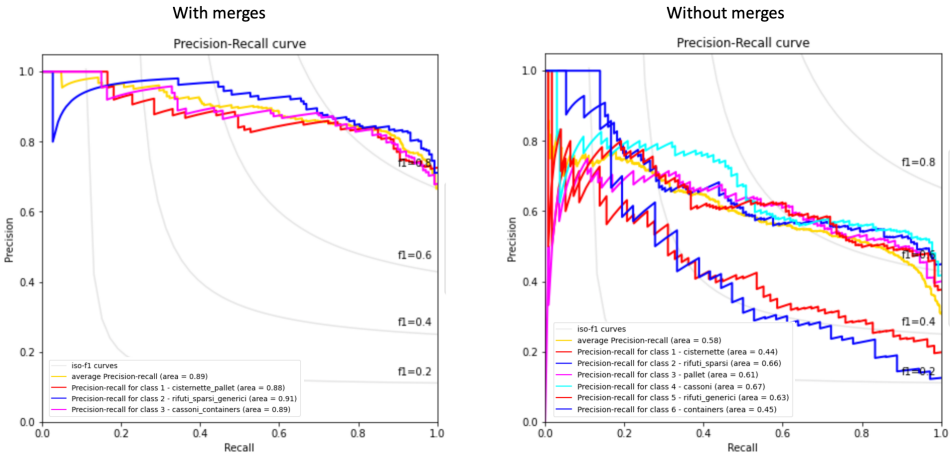

Figure 6.1: Comparison between chosen (on the left) and comparative (on the right) models on Precision-Recall curves

Here the Precision-Recall curves of the two considered models are shown. In particular, the one on the left refers to the model trained on the dataset with the merged categories, while the one on the right is related to the training with the 6 categories separated. It is immediately visible the cleanness obtained in the first case, in contrast with the confusion visible in the second one. This further confirms the relevant influence that the similarity between the classes has on the model performance.

6.2.2 Distribution of Probability and Model Calibration

When performing a multilabel classification task, it is particularly important to also take into consideration the calibration of the model, which defines how much correspondence there is between the model prediction scores and the frequency with which predictions match the actual labels.

To give a complete analysis, the model calibration has been firstly investigated on the single categories, and subsequently on the three classes together. In parallel with the reliability diagram, which shows the calibration of the model, it has been also taken into consideration the confidence histogram, defining the distribution of probability on the samples in the defined prediction intervals.

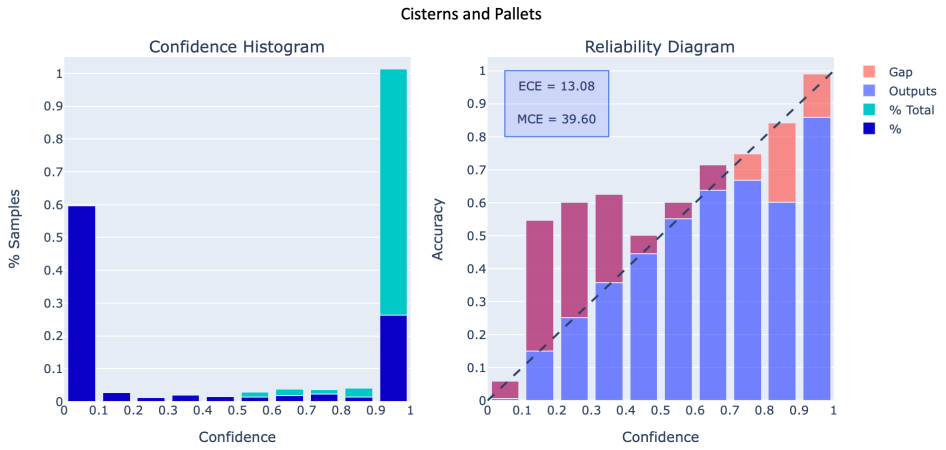

Figure 6.2: Calibration of the model on the on the test set - Cisterns and Pallets

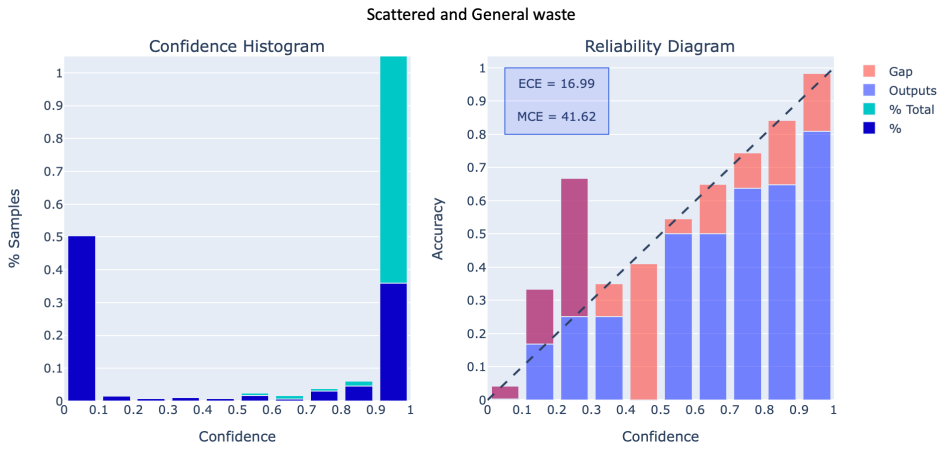

Figure 6.3: Calibration of the model on the test set - Scattered and General waste

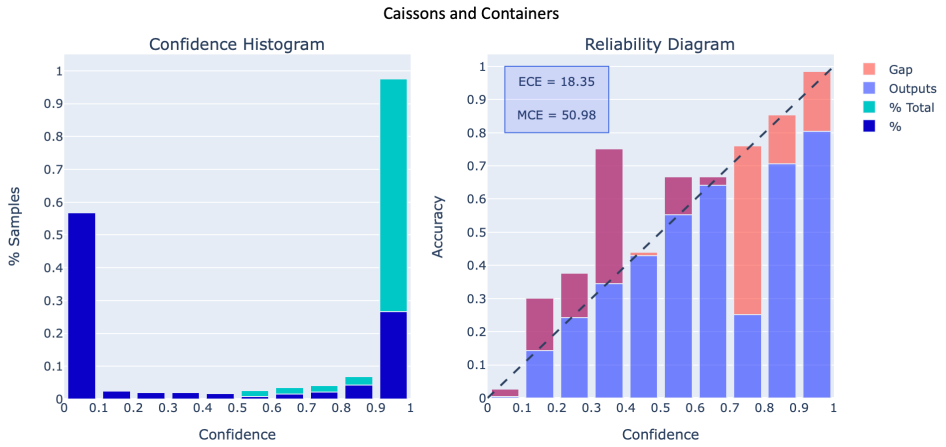

Figure 6.4: Calibration of the model on the test set - Caissons and Containers

In Figures 6.2, 6.3 and 6.4 are reported the confidence histograms and the reliability histograms of the model on the three single categories. The first impression could be that the model is not very well-calibrated, since there are intervals that are a bit far from the 45° line. However, it must be taken into consideration that in all the three cases the number of samples predicted with a score between 0.1 and 0.9 is very few, and thus in most of the cases is not possible to have a realistic measurement of the model calibration. An example is given in interval [0.4, 0.5] in Figure 6.3, referring to the category Scattered and General waste, where, as it can be seen from the confidence histogram, there are very few samples predicted with those scores. In fact, it has emerged that in this interval there is a single sample predicted as negative. As a consequence, in the corresponding reliability diagram, the relative bar is empty. This, of course, does not allow to give a real indication of the model calibration in the middle intervals, and thus an additional investigation of the overall case is needed.

Further confirmation of what has been stated is given by the difference between the Expected Calibration Error (ECE) and Maximum Calibration Error (MCE) measurements. In all the three cases the ECE remains low since it averages among all the bins. The MCE, instead, reaches high values because of the absence of a reasonable number of samples in some of the intervals as explained previously.

To have a more realistic measurement of the model calibration it is necessary to take into consideration all the categories together, thus looking at the graph shown in Figure 6.5.

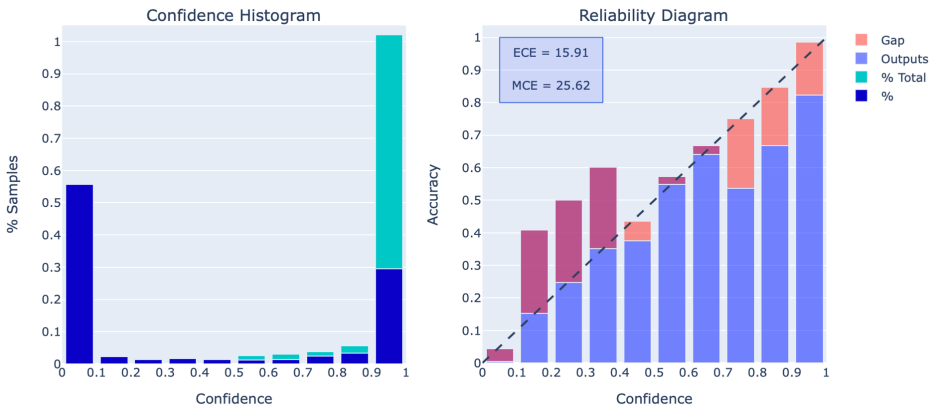

Figure 6.5: Calibration of the model on the test set - All the categories

From what can be seen and knowing that in practical cases perfect calibration is impossible [31], it can be said that the model is very well calibrated. Also in this case, what has been stated is reflected by numbers: both the ECE and MCE are low. The first one is inline with the single categories, while the second one, which measures the maximum gap, has been almost halved.

6.3 Qualitative Analysis

This section gives a qualitative evaluation of the results that have been achieved within the multilabel classification task. In particular, the goodness of the features that the model has learned is assessed by looking at the CAMs, while the generalization robustness is inspected through the comparison between the predictions made on images representing the same place, but with different resolutions and sizes.

6.3.1 Class Activation Maps Evaluation

In order to understand if the model has learned the right features and if it is looking at the right places when making predictions, it is a good practice to look at the class activation maps (CAMs) that it produces, that identify which are the image discriminative regions according to the model. To do this, here below are shown some examples that demonstrate how the model looks at the input image in the process of classifying it.

On the top left of Figure 6.6 is shown an image which contains the second and the third categories, thus whose ground truth includes both the classes Scattered and General waste and Caissons and Containers. On the

top right (1) it is reported the class activation map that the model outputs with respect to the Cisterns and Pallets class (predicted with probability 0.001), on the bottom left (2) the one referred to the Scattered and General waste class (predicted with probability 0.999) and on the bottom right (3) the CAM of the Caissons and Containers class (predicted with probability 0.999).

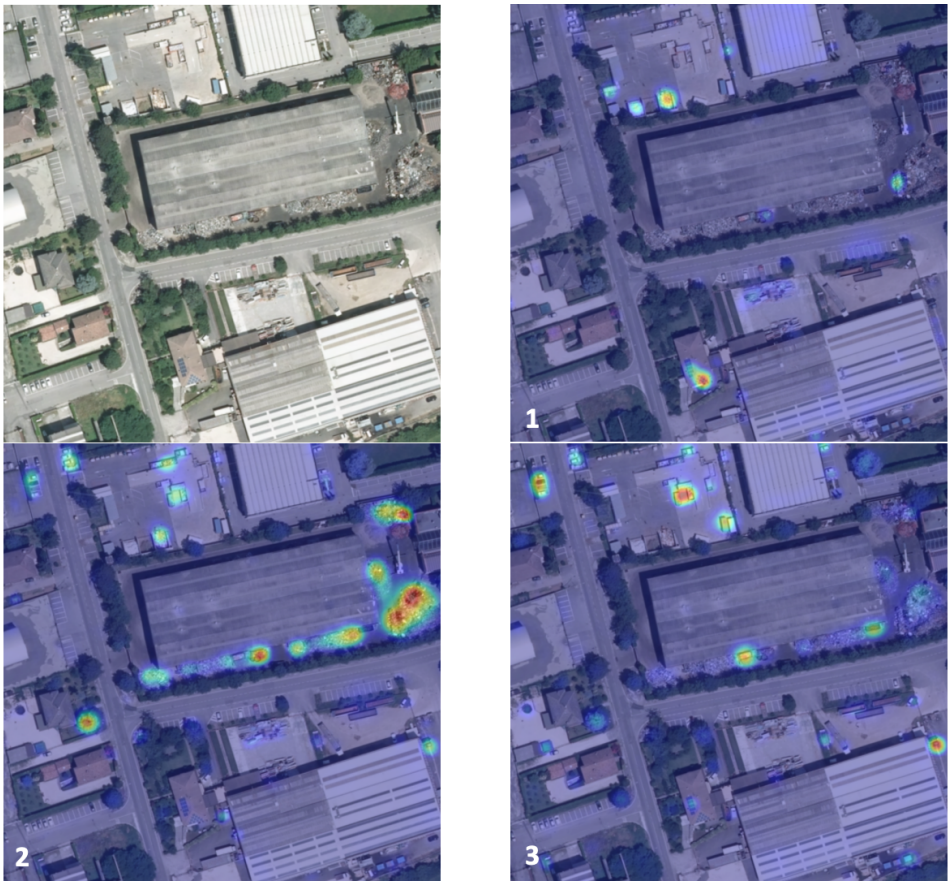

Figure 6.6: CAMs - Example 1

As it can be already seen in this first example, the model can retain a remarkable localization ability until the very final layer, correctly and concisely identifying the areas of the image that make it belonging to the intended classes.

Figure 6.7 shows an additional example where the model correctly predicts the presence of the Scattered and General waste category (with probability 0.999), together with the Caissons and Containers class (with probability 1.000). Even in this case, it is possible to notice a very high precision

and cleanness of the model in the localization of the 2 caissons present in the image. This is a further demonstration of the fact that the right features have been learned.

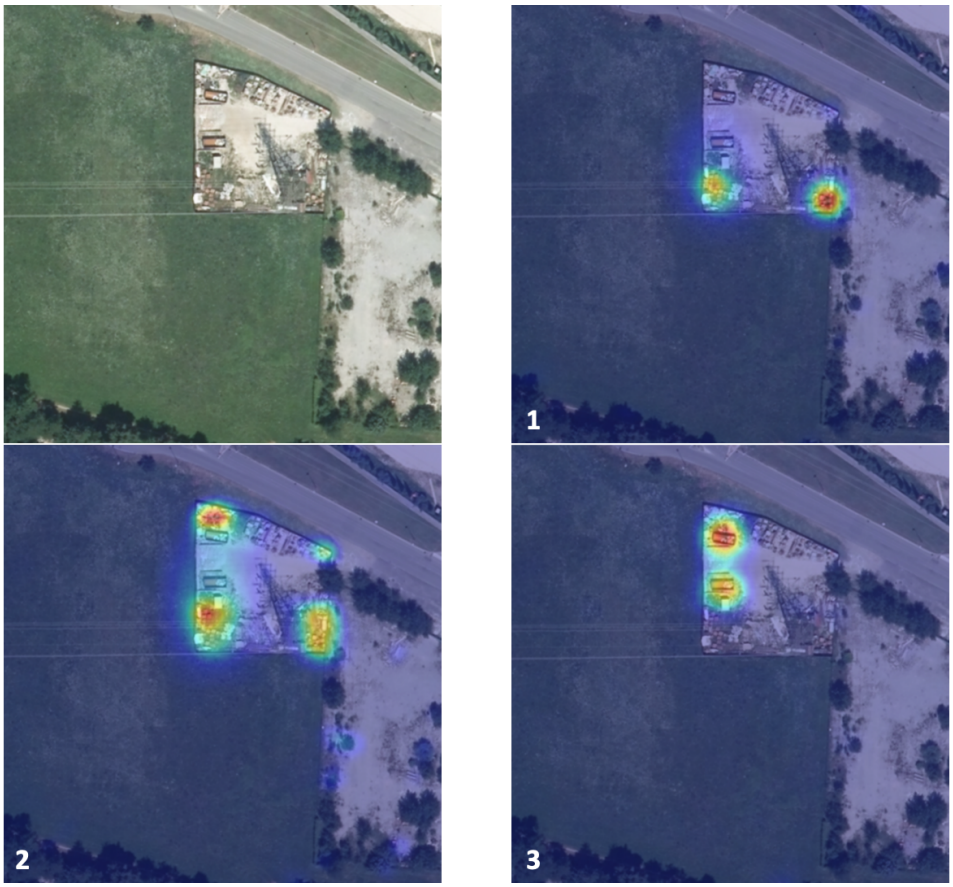

Figure 6.7: CAMs - Example 2

In order to demonstrate the goodness of the model also with respect to the Cisterns and Pallets class, in Figure 6.8 is reported an example of an image belonging to this class and to the Scattered and General waste category (that, as it can be noticed, is really frequent). Looking at the four pictures, it is possible to see that also in this case all the critical regions have been correctly covered with higher values of the CAMs. This sample has been in fact predicted to belong to the first class (1) with a probability of 1.000, to the second class (2) with a probability of 0.999, and to the third (3) with a probability of 0.001.

Figure 6.8: CAMs - Example 3

Finally, it is also important to show an even more difficult case, where objects are very mixed between each other. In particular, the image at the top left corner of Figure 6.9 shows a situation in which there are caissons among general wastes. In these types of conditions, often also the human eye has to pay a bit of attention. Despite this, the image has been correctly predicted to belong to the Scattered and General waste (0.999) and Caissons and Containers (0.999) categories, assigning instead a very low probability (0.010) to the Cisterns and Pallets class. The correspondent class activation maps further demonstrate that the model is focusing on the right regions.

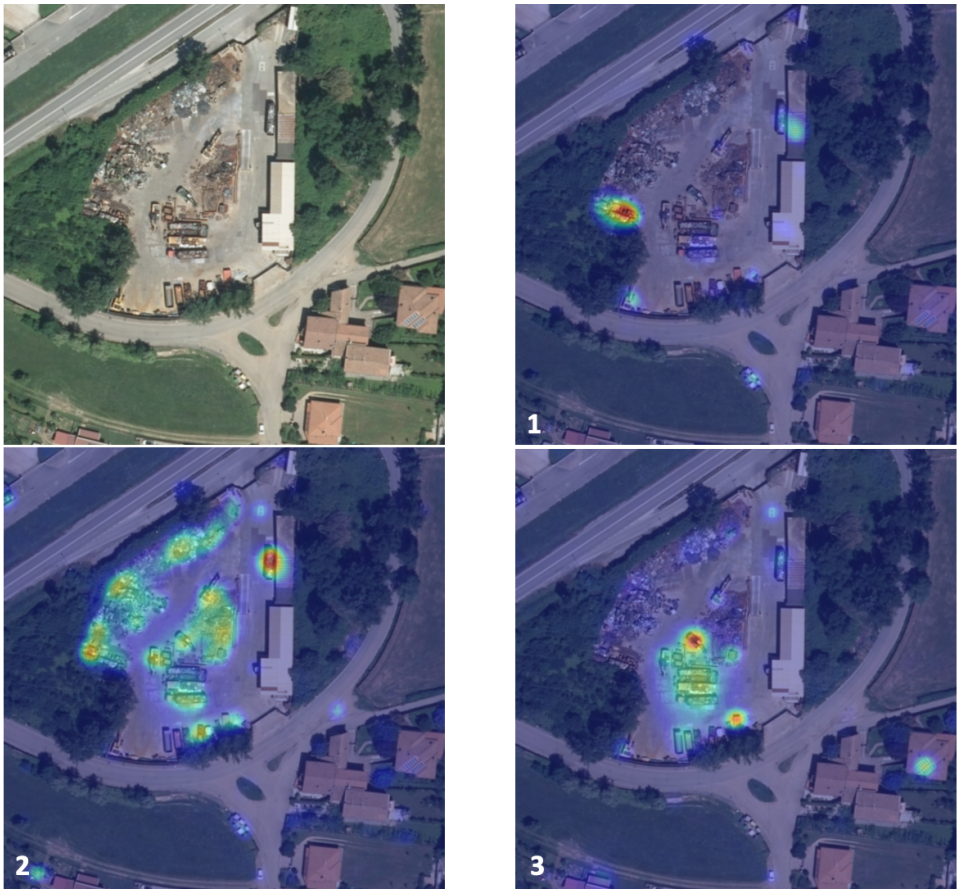

Figure 6.9: CAMs - Example 4

6.3.2 Mapbox and Ortophoto Comparison

As previously anticipated, the main difference between Mapbox and Orthophoto images is the resolution, which is lower in the first one with respect to the second. For this reason, a deeper analysis has been carried out in order to understand the influence of this factor on the performance of the model.

In Figures 6.10, 6.11 and 6.12 are shown the three situations which have been distinguished while comparing the performance of the model on Mapbox and Orthophoto images.

	Mapbox		Orthophoto	
	Ground Truth	*Prediction*	*Ground Truth*	*Prediction*
Cisterns and Pallets	no	no (0.01)	no	no (0.25)
Scattered and General waste	no	no (0.00)	no	no (0.27)
Caissons and Containers	yes	yes (0.73)	no	no (0.25)

Table 6.3: Mapbox and Orthophoto comparison - First situation

Figure 6.10: Mapbox (left) and Ortophoto (right) comparison - Different content

In Figure 6.10, in particular, the images, even if representing the same place, have different resolutions, and different content. This is a situation that happens often because Mapbox and Orthophoto images are taken in time instants that could be even very distant, thus in that time frame of difference one place can change very much. Despite this, most of the time the model correctly predicts in both the sources, as it is possible to see in Table 6.3 referred to the example shown above. The ground truth in fact changes between the two images and the model correctly account for the variation. This means that it was focusing on the right regions for detecting the Caissons and Containers class.

	Mapbox		Orthophoto	
	Ground Truth	*Prediction*	*Ground Truth*	*Prediction*
Cisterns and Pallets	yes	yes (0.99)	yes	yes (0.99)
Scattered and General waste	no	no (0.22)	no	no (0.00)
Caissons and Containers	no	no (0.06)	no	no (0.00)

Table 6.4: Mapbox and Orthophoto comparison - Second situation

Figure 6.11: Mapbox (left) and Ortophoto (right) comparison - Similar content

In Figure 6.11 is shown the second type of situation, where the content is practically identical, thus the only difference is related to the resolution. As it can be seen in Table 6.4, in both cases the image contains only the Cisterns and Pallets category in the ground truth, and the model correctly identifies it in its predictions. The difference in resolution is reflected in the probability values that are produced. In fact, it is possible to notice that while in the Mapbox case the model is more prone to predict positively, in the Orthophoto case, it tends to be more conservative, giving a higher score to the categories that are evidently present and a lower one to those that are less visible.

	Mapbox		Orthophoto	
	Ground Truth	*Prediction*	*Ground Truth*	*Prediction*
Cisterns and Pallets	yes	yes (0.91)	yes	yes (0.99)
Scattered and General waste	yes	yes (0.94)	yes	yes (0.96)
Caissons and Containers	no	no (0.02)	yes	yes (0.99)

Table 6.5: Mapbox and Orthophoto comparison - Third situation

Figure 6.12: Mapbox (left) and Ortophoto (right) comparison - Grainy Mapbox

Figure 6.12 is instead an example of the third situation, where the Mapbox image is even more grainy than usual, making the object distinction task difficult even for the human eye. In the reported Mapbox image for example it is absolutely not trivial to identify the Cisterns and Pallets in the middle (both the blue heap of cisterns and the near sequence of pallets groups), however, the model correctly predicts even in this case. It also accounts for the difference in the content of the two pictures. The Orthophoto image, in fact, contains some containers that are not present in the Mapbox one, and this is reflected in the prediction of the model, as it is possible to see in Table 6.5.

In general, it has been instead noticed that the model tends, in both cases, to confuse some types of buildings with the Caissons and Containers class. A clear example is given in Figure 6.13. The relative predictions are shown in Table 6.6. However, as it can be seen, there are objects that clearly seem to be containers. Only using other sources to investigate (such as Google Maps) it is possible to know that they are small buildings. The

model, overall, tends more to predict positively on Mapbox images, due to the lower resolution.

	Mapbox		Orthophoto	
	Ground Truth	*Prediction*	*Ground Truth*	*Prediction*
Cisterns and Pallets	no	no (0.01)	no	no (0.00)
Scattered and General waste	yes	yes (0.51)	no	no (0.01)
Caissons and Containers	no	yes (0.96)	no	yes (0.89)

Table 6.6: Mapbox and Orthophoto comparison - Caissons and Containers

Figure 6.13: Mapbox (left) and Ortophoto (right) comparison - Caissons and Containers

In conclusion, in order to have a complete idea of how differently the model acts on Mapbox and Orthophoto images, the model has been run on an entire territory, the same on both the sources. In particular, a portion of Nave municipality has been analyzed and the results that have been obtained for each category and each source are reported in Table 6.7.

	Mapbox Images	*Ortophoto Images*
Cisterns and Pallets	108	6
Scattered and General waste	150	24
Caissons and Containers	70	62

Table 6.7: Mapbox and Ortophoto comparison - Model run on a new territory

Considering the differences in the content of the images and the frequency of occurrence of more grainy images, this discrepancy in numbers is expected.

A further example of the robustness of the model is shown in Figure 6.14, which displays two images taken from the analyzed territory. Even if they are almost similar, the Mapbox one (on the left) is evidently more grainy with respect to the Orthophoto image (on the right), and in the bottom part it contains scattered waste. As it can be seen in Table 6.8, the model correctly takes into account the difference, predicting the presence of the Caissons and Containers class in both the images, and of the Scattered and General waste only in the one from Mapbox.

	Mapbox		Orthophoto	
	Ground Truth	*Prediction*	*Ground Truth*	*Prediction*
Cisterns and Pallets	no	no (0.06)	no	no (0.00)
Scattered and General waste	yes	yes (0.99)	no	no (0.11)
Caissons and Containers	yes	yes (0.99)	yes	yes (0.97)

Table 6.8: Mapbox and Orthophoto comparison - Fourth situation

Figure 6.14: Mapbox (left) and Ortophoto (right) comparison - Further example of diversity

Chapter 7

Evaluation - CAMs Analysis for Weakly Supervised Object Detection

This chapter provides a quantitative and qualitative analysis of the results obtained while studying the feasibility of the weakly supervised object detection task, based on the features learned in the multilabel classification task.

In particular, for what regards the quantitative analysis, metrics will be shown more in detail, giving an analytic idea of how good are the discriminative regions identified by the model.

Regarding the qualitative analysis, some examples of the bounding boxes derived from the model predictions will be shown, comparing them with the ones in the ground truth, and showing the strengths and weaknesses of this approach.

Also in this case, before starting with the actual evaluation of the results, the metrics that have been used for the specific task are described, thus framing the analysis.

7.1 Metrics

In order to give a better understanding of the way in which the performance of the model has been measured, in this section the list of all the main metrics is provided, giving both the related definition and formula. In particular, here the focus is on the metrics used in the feasibility study of the weakly supervised object detection task.

Intersection over Union (IoU)

Measure of the extent of overlap between two boxes. The greater is the overlapping region and the larger is the IoU.

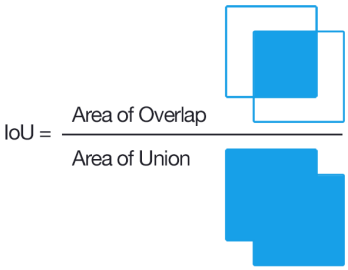

Figure 7.1: Intersection over Union computation

Component IoU

Evaluation of how much the class activation map focuses on the single categories. First, the class activation map foreground area is divided into connected components, which means groups of pixels connected to each other. Then the IoU value is calculated between each ground truth bounding box and the connected components that intersect it. Finally, the average is taken.
This is done for each considered threshold on the class activation maps, which is used to consider only CAMs with a value greater than the threshold. It can be calculated also by considering the bounding boxes derived from the predicted CAMs instead of the CAMs themselves, in order to evaluate the final model performance.

Global IoU

Plot of the measurements of the IoU between the union of all the bounding boxes in the image and the entire foreground area of the class activation map given a threshold. This metric is calculated for all threshold values and assesses how the class activation map focuses on the whole waste region, favoring models that generate wider and more connected areas rather than separated components. This is repeated for every considered threshold. It can be calculated also by considering the bounding boxes derived from the predicted CAMs instead of the CAMs themselves, in order to evaluate the final model performance.

Bounding Box Coverage

Plot of the number of bounding boxes that are covered by class activation maps. This metric alone is not enough to characterize the performance because a trivial class activation map that covers the entire image would have 100% coverage. However, coupled with the two previous metrics, it can give information about which model can generate CAMs that can highlight a large fraction of objects belonging to the considered categories. This is repeated for every considered threshold and can be calculated also by considering the bounding boxes derived from the predicted CAMs instead of the CAMs themselves, in order to evaluate the final model performance.

Irrelevant Attention

Percentage of CAMs area outside any bounding box. When evaluating the global IoU, a low value can occur for two reasons: the two areas have a very small intersection, or the two areas overlap well but one is much larger than the other. Thus, an analysis of how much the class activation maps focus on irrelevant parts of the image helps to characterize low global IoU values. This is repeated for every considered threshold. It can be calculated also by considering the bounding boxes derived from the predicted CAMs instead of the CAMs themselves, in order to evaluate the final model performance.

7.2 Quantitative Analysis

In this section, a quantitative analysis of the results achieved within the feasibility study of the weakly supervised object detection task is given. In particular, the Component IoU, the Global IoU, the Bounding Box Coverage, and the Irrelevant Attention metrics have been measured in three different configurations:

- Measurement of the four metrics between the ground truth bounding boxes and the class activation maps predicted by the model.
- Measurement of the four metrics between the ground truth segmentations and the class activation maps predicted by the model.
- Measurement of the four metrics between the bounding boxes of the ground truth and the ones derived from the CAMs predicted by the model.

7.2.1 Ground Truth Bounding Boxes and CAMs

In the first measurements that have been performed, the four metrics previously exposed have been computed between the bounding boxes given by the ground truth and the CAMs produced by the model. In particular, thresholds are used in order to identify the foreground region of the class activation maps to consider in the calculation.

Besides the estimate of the model performance, in this case, these four metrics have been also computed in order to gather quantitative hints on which threshold is most suited for the derivation of the bounding boxes from the predicted CAMs. The same metrics have been in fact calculated for each threshold in the [0, 1] range with a step of 0.05.

Figure 7.2 shows the obtained Component IoU.

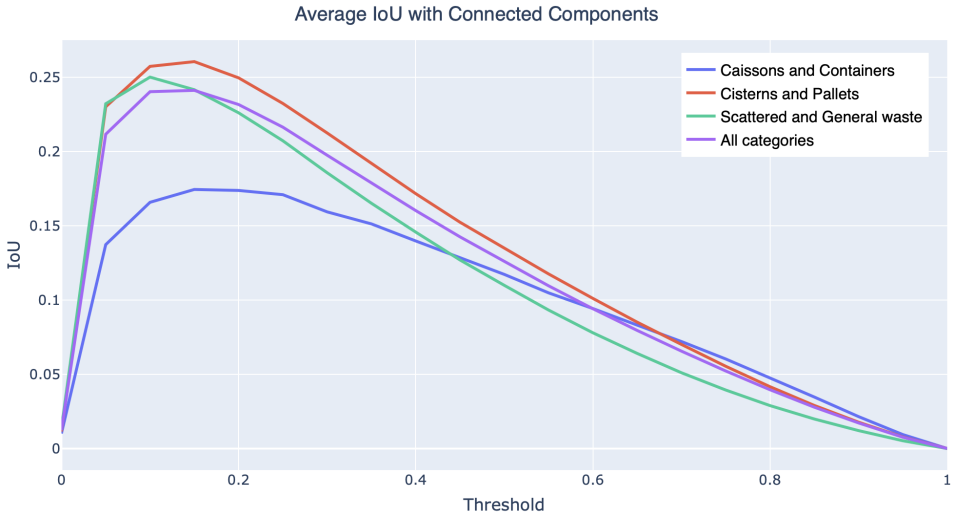

Figure 7.2: Component IoU - GT Bounding Boxes and CAMs

For what regards the classes Scattered and General waste and Cisterns and Pallets it is possible to notice a similar trend. Both reach high values with thresholds in the range [0.1, 0.15], which means that the model is precise at distinguishing the areas of the image that are occupied by the objects of interest, from the regions where the considered categories are absent. In fact, it is enough to use a low threshold value to obtain a very good superimposition between what the model predicts and the ground truth.

Regarding the Caissons and Containers class, the model struggles a bit more, but still reaches good values in the same threshold range. On one side this is due to the fact that, while in the other two categories single objects are small and in most of the cases organized in clusters, in the Caissons and

Containers one the single objects are more distinguishable between each other. Even if in many cases some caissons and containers are close between each other, the model tends to identify them separately, with one bounding box per single object. The ground truth, instead, in the majority of these situations, encloses the entire group of caissons and/or containers within the same rectangle. Of course, this causes a lowering of the intersection over union metric, because ground truth bounding boxes are inevitably bigger than the predicted ones. However, this can be considered as an advantage because demonstrates another time how good are the features that the model has learned, and allows it to also be more precise in the localization task. An example of what has been said is reported in Figure 7.15 in the qualitative analysis Section.

On the other side, this is mainly due to the fact that, as also happened and outlined while solving the multilabel classification task, the model classifies some types of buildings as belonging to the Caissons and Containers class. As it has been already demonstrated, however, in many cases it is difficult even for the human eye to untangle this kind of situation (a clear example has been made in Figure 6.13).

In the end, considering the three categories together, the model achieves a maximum of 24.11% of Component IoU.

Figure 7.3, instead, shows the Global IoU, thus the IoU between the union of all the bounding boxes in the image and the entire foreground area of the class activation map given a threshold.

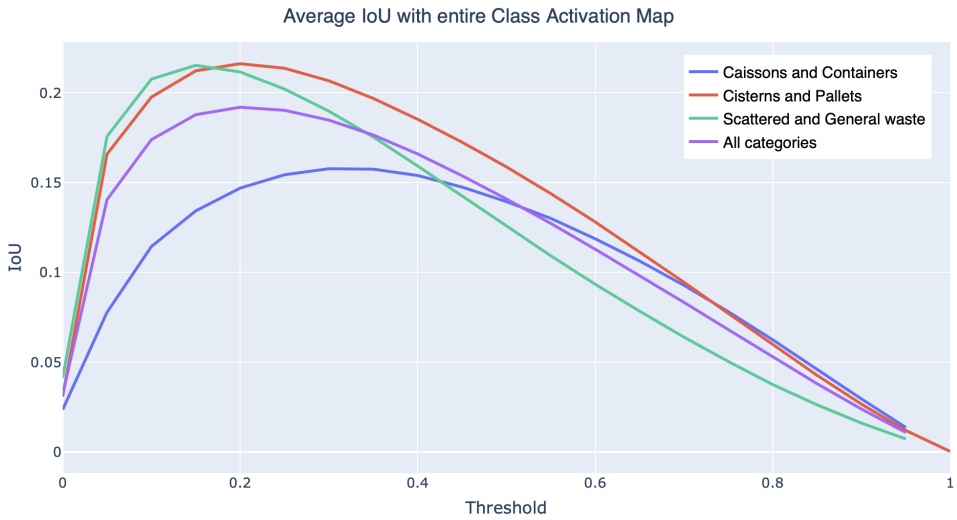

Figure 7.3: Global IoU - GT Bounding Boxes and CAMs

Also in this case, which is an alternative way of measuring the intersection over union, is possible to notice the same trend that has been encountered previously, thus having a further confirmation of the fact that the model reaches the best results on the first two classes of the dataset (Cisterns and Pallets and Scattered and General waste), achieving instead a bit lower performance on the third one, but acting very well in the overall case. Moreover, even in this case, a low threshold is enough to sharply divide the positive regions from the negative ones, being again in the around of 20% of intersection over union between the ground truth bounding boxes and the predicted CAMs.

In order to go further in the current analysis, it is possible to look at the Bounding Box Coverage and at the Irrelevant Attention measurements shown respectively in Figures 7.4 and 7.5. These two metrics are in fact an optimal way of getting a deeper understanding of the curves seen in the previous graphs.

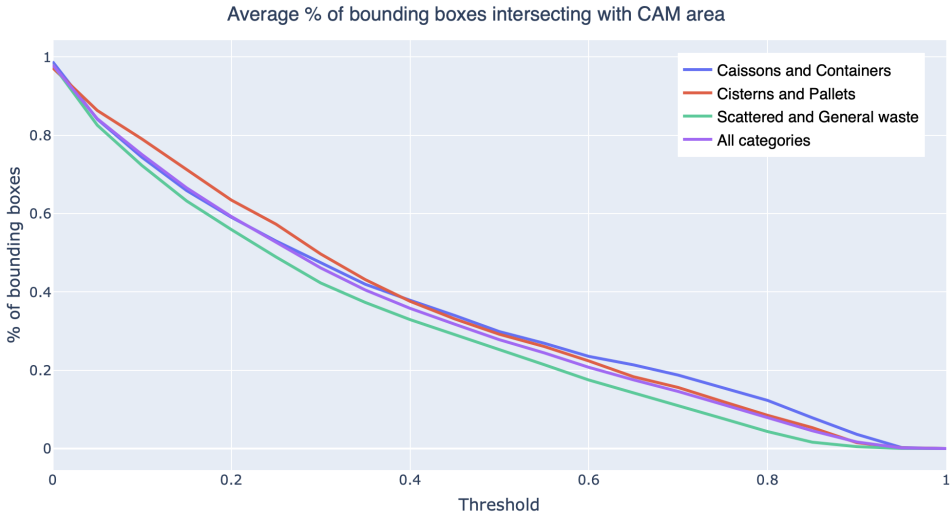

Figure 7.4: Bounding Box Coverage - GT Bounding Boxes and CAMs

The Bounding Box Coverage in fact helps in figuring out how good is the model in covering the ground truth bounding boxes with the predicted CAMs. The first thing that can be noticed is that the architecture works in an almost equivalent way over all the three classes. The deviation between the three curves is in fact really minimal. Additionally, the model retains a good covering of the ground truth bounding boxes even when the threshold gets at high values like 0.4. In fact, even in this case, the predicted CAMs are able of intersecting on average 35% of the ground truth bounding boxes.

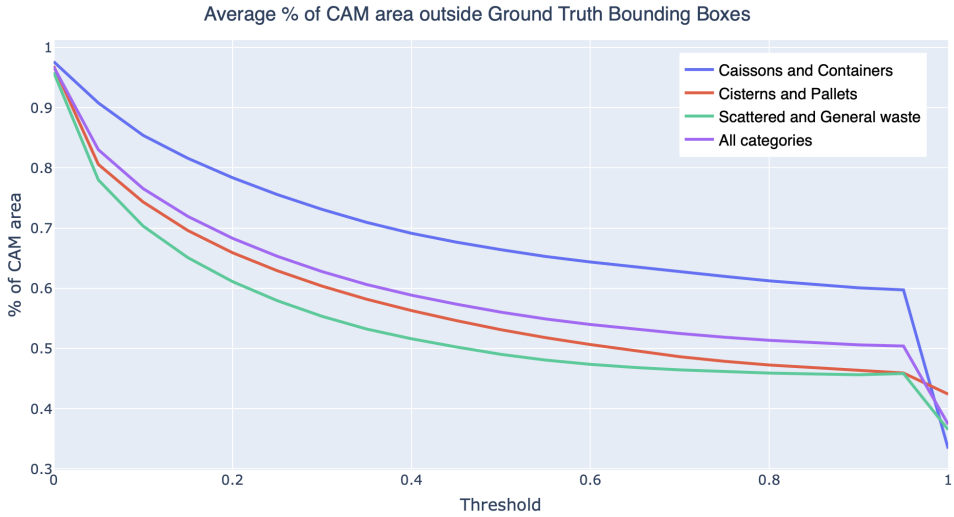

Figure 7.5: Irrelevant Attention - GT Bounding Boxes and CAMs

The analysis on how much the class activation maps focus on irrelevant parts of the image helps instead in characterizing low Global IoU values. In particular, Irrelevant Attention corresponds to the percentage of CAM area outside any bounding box. This metric revealed in fact really useful to understand the reason why the models gain lower values over the Caissons and Containers class. As it can be seen in Figure 7.5, the curve related to the cited category is evidently detached from the others, always reaching higher values. This, in practice, demonstrates what has been said previously regarding the fact that the model tends sometimes to classify some types of buildings as belonging to the Caissons and Containers class, thus including in the CAM also regions that are totally outside of the ground truth bounding boxes. So, when measuring the IoU, these regions contribute to lowering the metric.

7.3 Ground Truth Segmentations and CAMs

In this section are shown the same four metrics that have been analyzed in the previous one, but this time considering the ground truth segmentations instead of the bounding boxes. A general observation that is worth doing before showing any graph is that in this case, the performance is of course lower than before. This is however expected since the segmentations cover a very restricted area of the image, much smaller with respect to the bounding boxes. The metrics calculation is instead very useful also in this case in order to further investigate the performance and the behavior of the model.

Regarding the Component IoU, shown in figure 7.6, the first thing that can be noticed is that even in this case the model reaches its maximum performance within the same range as before ([0.1, 0.15]). Here however there is no longer the same detachment seen in the previous case. Taking into account the scale used on the y-axis, curves are in fact really close between each other. This demonstrates that even if the model tends to consider some types of buildings as belonging to the Caissons and Containers class, in the end, it correctly covers also the right objects in the same way it does with the other two classes.

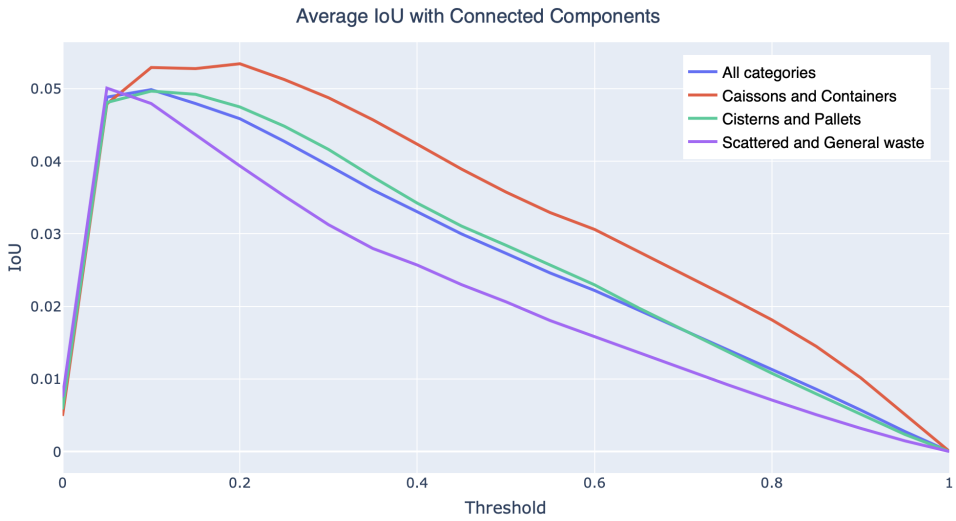

Figure 7.6: Component IoU - GT Segementations and CAMs

The theory is proved also in the Global IoU graph shown in Figure 7.7, where the Caissons and Containers reach higher values with higher thresholds. This in fact means that the model, even if it sometimes incorrectly considers some object, gives higher values to the pixels of the image that actually represent the objects of interest. Consequently, in order to get an improved overlap between the ground truth segmentation and the class activation map, higher thresholds are needed in order to filter out the wrong regions and keep instead only the ones that are interesting for the prediction.

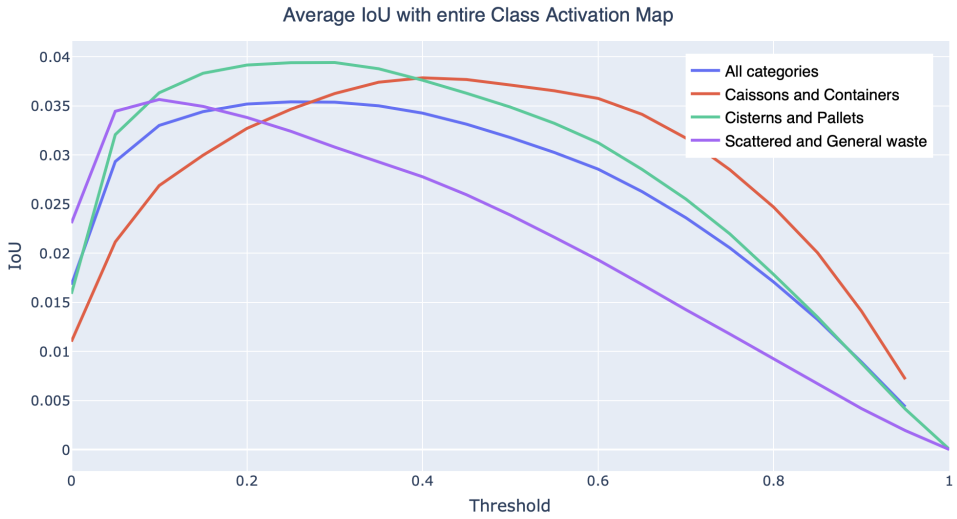

Figure 7.7: Global IoU - GT Segmentations and CAMs

Regarding the Bounding Box Coverage and the Irrelevant Attention, respectively shown in Figure 7.8 and 7.9, in this case, they show an equivalent trend for all the three classes, thus further validating the fact that the model acts in an equivalent way for all the three classes in terms of actually recognizing the pixels of the image occupied by the objects of interest.

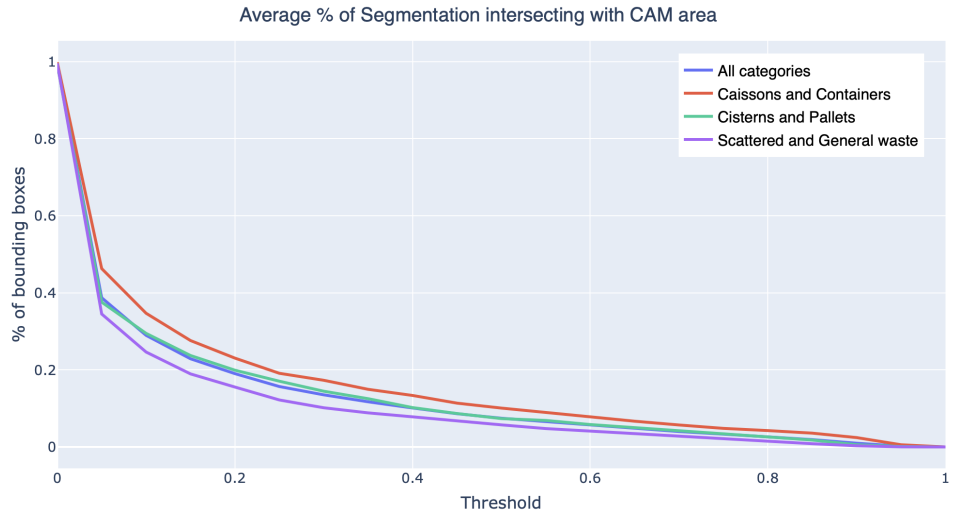

Figure 7.8: Bounding Box Coverage - GT Segmentations and CAMs

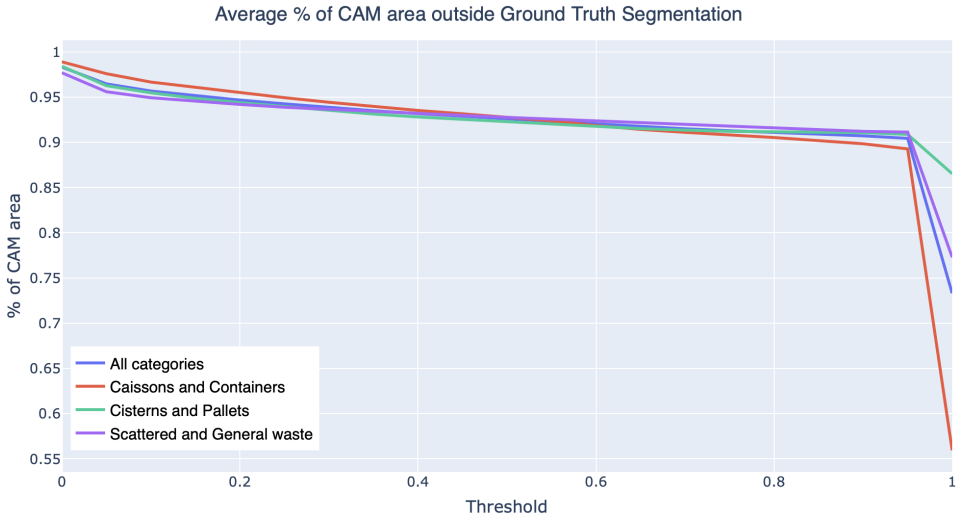

Figure 7.9: Irrelevant Attention - GT Segmentations and CAMs

7.3.1 Ground Truth Bounding Boxes and Predicted Bounding Boxes

The last quantitative measurements are meant for assessing the actual goodness of the model in producing CAMs that correctly localize the discriminative regions of an image. The same four metrics have been exploited to accomplish this evaluation because they demonstrated to be able to give a complete and in-depth quantitative understanding of the model behavior.

Figure 7.10 shows the Component IoU.

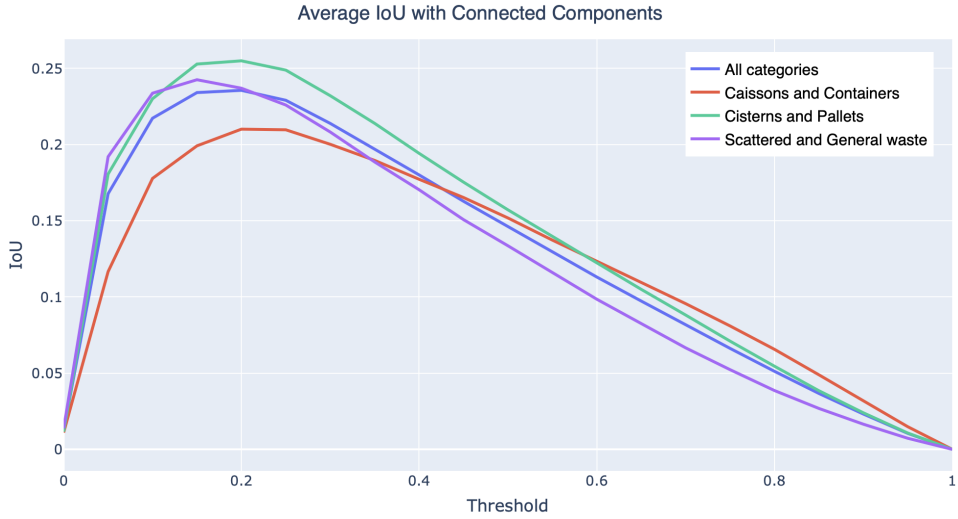

Figure 7.10: Component IoU - GT Bounding Boxes and predicted Bounding Boxes

With respect to the values obtained in the first case, in which CAMs were considered instead of the derived bounding boxes, it is possible to notice that the interval in which the curves reach their maximum value has shifted a bit towards higher threshold values. This mainly because bounding boxes cover a larger area with respect to the corresponding class activation maps and so higher thresholds are needed in order to reduce the regions that fall outside the ground truth bounding boxes.

The second observation that can be made regards the fact that, while the curves corresponding to Cisterns and Pallets class and to Scattered and General waste class get in general lower values than in the first measurements, the curve corresponding to the Caissons and Containers category grows. This enforces what has been said in Section 7.2.1 regarding the fact that the model tends to identify caissons and containers singularly, even if they are close between each other, while the ground truth encloses objects belonging to this class in the same bounding box when they are close. Since the derived bounding boxes enlarge the considered area, the overlap with the ground truth bounding boxes improves, and thus the metric grows.

Regarding the overall performance, the curve still stays in the around of 24%. To be precise it arrives at a maximum of 23.56% with a threshold of 0.2.

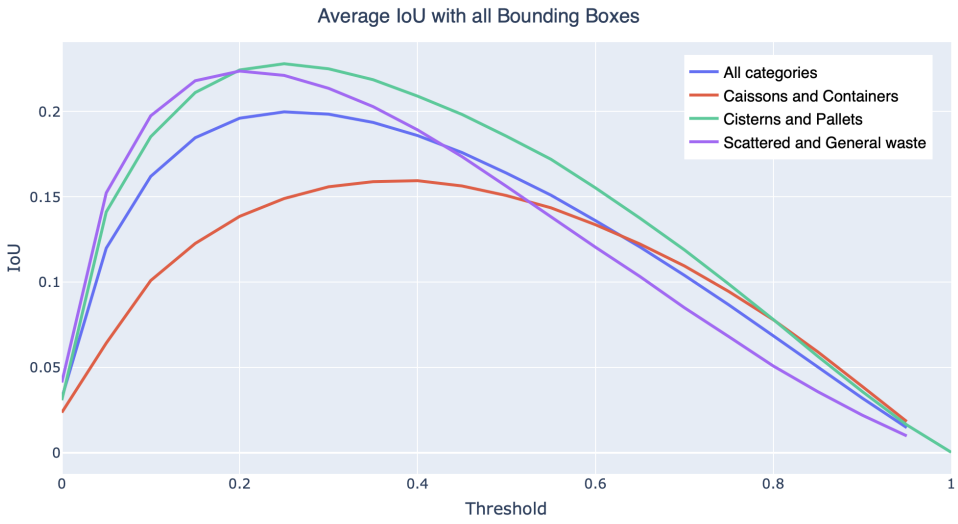

Figure 7.11: Global IoU - GT Bounding Boxes and predicted Bounding Boxes

Figure 7.11 shows the Global IoU. The shift of the peak values towards the left can be seen also here, confirming the fact that higher thresholds are needed to reduce the additional area considered by the derived bounding boxes and that falls outside the ground truth bounding boxes. Differently

than before, in this case, all the three curves related to the single categories grow. This however is due to the fact that the specific metric, as previously said, favors wider and more connected bounding boxes.

For what regards the Bounding Box Coverage, shown in Figure 7.12, as in the previous two cases it is possible to notice that the model behaves similarly for all the three categories. It is worth outlining that also here the curves are slightly higher than the ones obtained with the class activation maps. This because the bounding boxes cover a larger area with respect to CAMs.

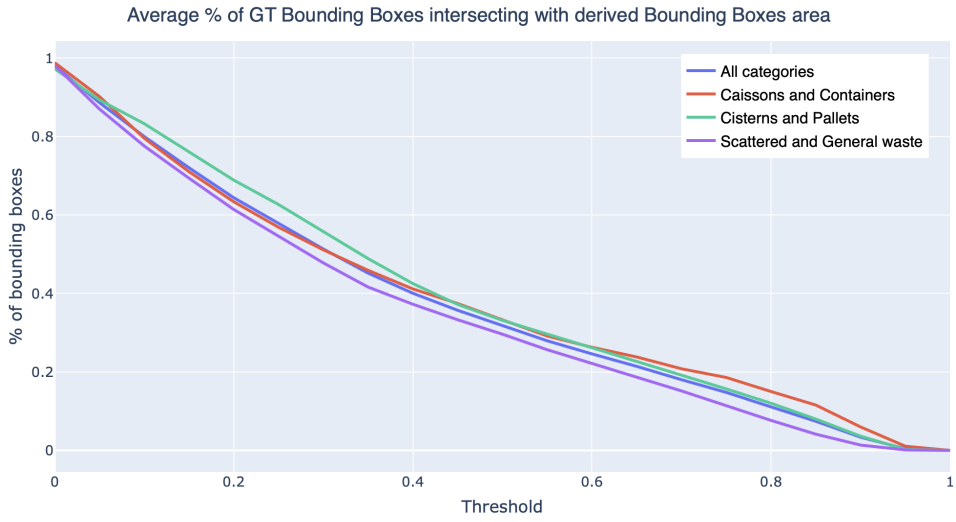

Figure 7.12: Bounding Box Coverage - GT Bounding Boxes and predicted Bounding Boxes

The same behavior has been encountered in the calculation of the Irrelevant Attention, shown in Figure 7.13, where curves slightly grow for the same reason explained before. Moreover, analogously to the first situation, thus the one in which class activation maps have been evaluated against ground truth bounding boxes, the curve referred to the Caissons and Containers class is higher than the other two (and of course of the overall one). The explanation is even in this case related to the fact that the model, in some more complicated situations, positively classifies some objects that have a shape similar to the caissons and containers one, thus increasing the CAMs area that falls outside the ground truth bounding boxes.

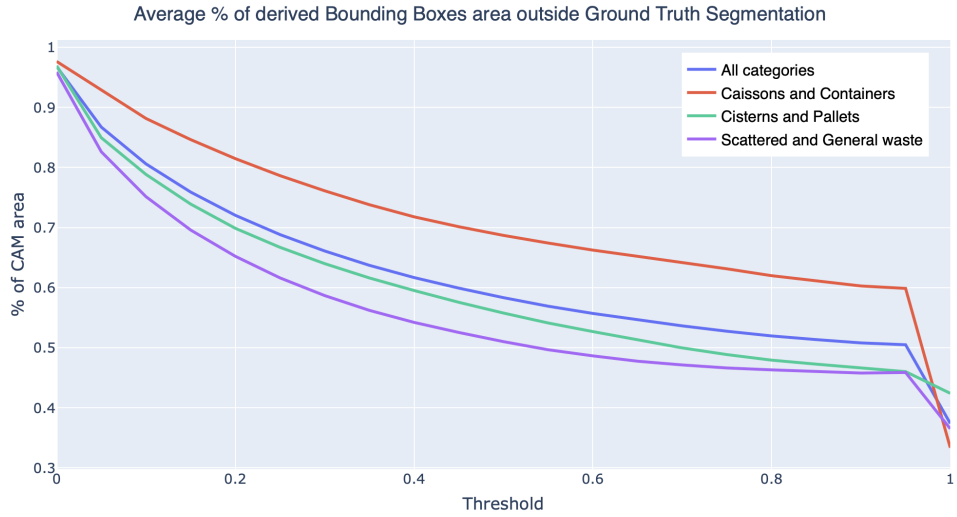

Figure 7.13: Irrelevant Attention - GT Bounding Boxes and predicted Bounding Boxes

In conclusion, according to what has been seen in the entire quantitative analysis, it seems that the best threshold to be used is to be chosen within the interval [0.20, 0.25].

7.4 Qualitative Analysis

This section gives a qualitative evaluation of the results that have been achieved within the feasibility study of the weakly supervised object detection task. In particular, some examples are shown, that clearly display how the model is able to retain a remarkable localization ability with respect to the objects belonging to the three considered classes. All the images that are shown in this Section follow a specific color code in terms of bounding boxes:

- Red bounding boxes enclose Cisterns and Pallets class.
- Green bounding boxes enclose Scattered and General waste class.
- Blue bounding boxes enclose Caissons and Containers class.

Regarding the threshold value, the interval obtained during the quantitative analysis has been tested. Even if those values already give good visual results, a threshold of 0.4 has been adopted, which helped in producing cleaner images, still maintaining good localization capabilities.

In general, from what has been seen, the model acts very well, being able to get very close to the ground truth in many cases. Sometimes it is

even difficult to distinguish between the bounding boxes derived from the prediction and those given by the ground truth. An example is given in Figure 7.14, where the model succeeds in correctly localizing most of the objects of interest. Here in fact two classes have been detected: Cisterns and Pallets and Scattered and General waste. Looking at the prediction (on the left) it is possible to see that for the second category, two out of two groups of waste have been correctly detected and localized. Regarding the first category, two out of three groups of pallets have been individuated. Carefully looking at the image it is possible to notice that even the third one has been detected, however, evidently, the predicted CAM scores are too low to generate a reasonable rectangle enclosing it.

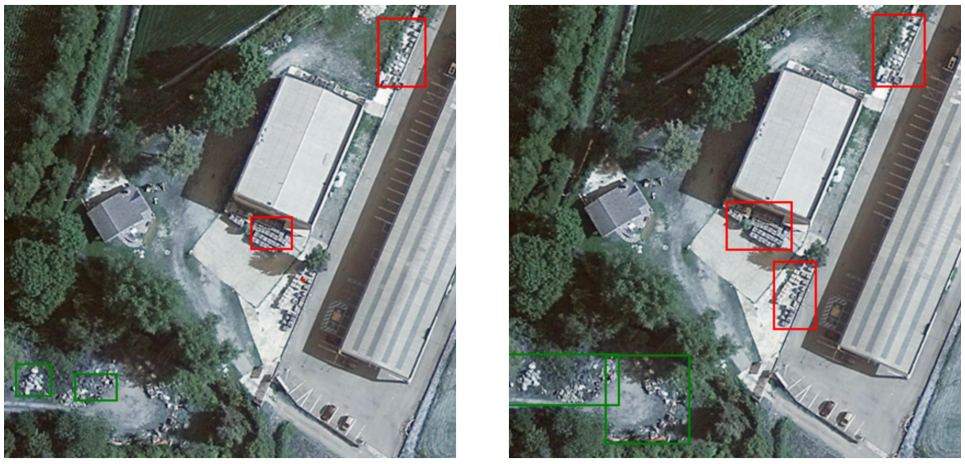

Figure 7.14: CAMs qualitative analysis for Weakly Supervised Object Detection - First example (On the left the derived prediction and on the right the ground truth)

Another good example is given in Figure 7.15, where, again, most of the objects outlined by the ground truth have been correctly predicted. In particular, here there is a bit more difference between the ground truth and the prediction. As previously outlined in the quantitative analysis section, in fact, the model tends to classify containers singularly, or in smaller groups, while in the ground truth sets of caissons and/or containers that are close between each other are enclosed all in the same bounding box. Labeling them singularly, however, would require an even larger amount of time for the dataset creation, resulting in the further increase of costs. On the other side, this is a clear example that gives unequivocal evidence of how much this difference in behavior can affect the Intersection over Union calculation.

The explanation is related to the fact that performing object detection without looking at the ground truth bounding boxes does not allow the model to know the interpretation that the ground truth gives, thus, even if the prediction is equivalently correct, the metrics get lower. For this reason, it is very important to accompany the quantitative analysis with a qualitative investigation of the obtained results.

Figure 7.15: CAMs qualitative analysis for Weakly Supervised Object Detection - Second example (On the left the derived prediction and on the right the ground truth)

Just looking at these first two examples (Figures 7.14 and 7.15), it is possible to notice that in general, the model tends to generate smaller bounding boxes with respect to the ones defined in the ground truth. This is mainly due to the choice of the threshold, that filters out lower values, however, it does not constitute a problem at all, since the main goal of finding the objects of interest is reached. In some cases it can also result to be an advantage because it allows more precise localization of the categories, not risking to include objects that have nothing to do with the category in question only because of a bad disposition that requires a larger bounding box.

A further demonstration of the model robustness is given in Figure 7.16, where there is a bit more confusion between objects and categories. The model, in fact, seems to not have too many problems in correctly finding most of the critical areas. Also in this case it is possible to notice some slight difference in the interpretation of the model with respect to the one used in the ground truth in terms of how to surround the critical objects.

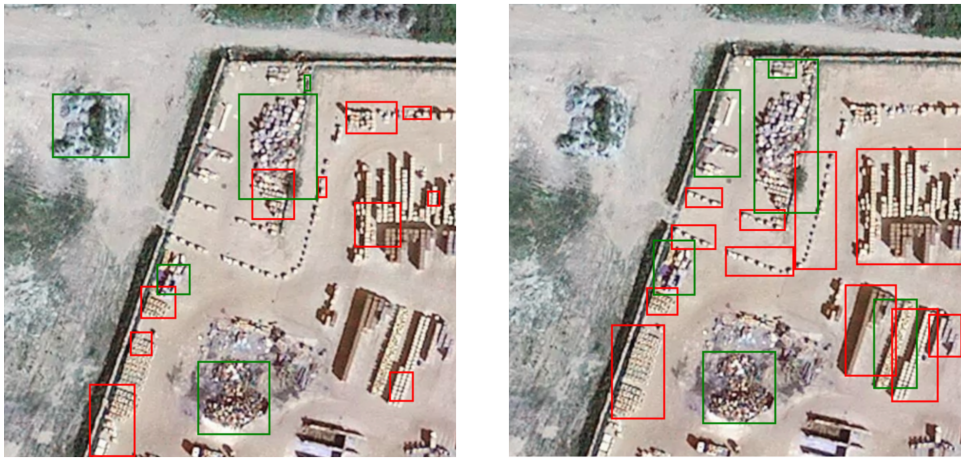

Figure 7.16: CAMs qualitative analysis for Weakly Supervised Object Detection - Third example (On the left the derived prediction and on the right the ground truth)

As it has been already observed in previous investigations, also during the qualitative analysis it has been noticed a bit more of struggle in distinguishing objects actually belonging to the class Caissons and Containers and objects that have only a shape similar to them. An example is shown in Figure 7.17.

Figure 7.17: CAMs qualitative analysis for Weakly Supervised Object Detection - Fourth example (On the left the derived prediction and on the right the ground truth)

Here there is only one real caisson, but beyond that also other areas have been enclosed with blue boxes. As already happened in previous examples, also here the human eye would have struggled a bit in telling whether some of them (apart from the ones that clearly enclose cars) belong or not to the Caissons and Containers class.

Chapter 8

Conclusions and Future Work

The goal of this research was twofold:

- Image classification based on the distinction of suspicious objects that usually constitute Illegal Landfills.
- Evaluation of the potential of the discriminative regions for waste localization.

To be more specific, the first task that has been solved is called multilabel classification, while the second one consists of the feasibility study of the weakly supervised object detection through the analysis of the class activation maps produced by the model.

First of all, the state of the art regarding the application of automatic and semi-automatic methods to the identification of ILs was presented, underlining that the final objective of the current research still lies in the complete automation of these processes. As it has been widely explained, Deep Learning techniques represent the most promising area, achieving always increasing results in the world of Artificial Intelligence applied to images, in particular with the usage of Convolutional Neural Networks (CNNs). Remote sensing images however often represented the Achilles tendon of these types of algorithms due to the numerous factors that contribute to the complexity of the task.

In this work are studied the performances that state-of-the-art techniques can achieve in the particular field of remote sensing images, with a focus on the multilabel classification task, also studying how promising is the weakly supervised approach for objects localization purposes. This study is part of a project promoted by the ARPA of the Lombardy Region for the monitoring

of ILs in the regional territory, which uses ResNet50 as the CNN to perform this task. The proposed solution extends the monitoring capability of the system by applying the same architecture in the concurrent classification of multiple categories and exploiting the goodness of the features learned by the model to also localize the critical objects.

The images composing the dataset used for training, validation and testing were provided by experts, and have been taken from the areas included in the provincial territories of Brescia, Pavia and Lodi. Two data sources have been used, Mapbox and Orthophoto, which mostly differ in the higher resolution of the images coming from the second one with respect to the first one. Also, the data augmentation technique called RandAugment demonstrated to give a significant contribution to the improvement of the generalization power of the model, mostly accounting for the high interclass similarity and intraclass variability.

Initially 12 classes have been considered, however, for half of them it has been not possible to find a number of samples that allowed their inclusion in the study. Also, among the remaining categories, it has been proved the presence of such a marked similarity between pairs of classes that has made it meaningless to keep them separate. In the end, three classes have been considered in the study: Cisterns and Pallets, Scattered and General waste, and Caissons and Containers.

The analysis process started with the multilabel classification task, assessing various configurations in terms of balancing of the categories, image size, and amount of data augmentation.

Given the higher frequency in the presence of certain classes and given the importance of a good balancing among them, the dataset has been populated maintaining the number of images per each positive class as equal as possible. In this sense, in order to exploit all the available data, it has also been tried to use oversampling for the classes with a lower number of samples with respect to the others. However this introduced overfitting, and thus the opposite direction has been traveled, undersampling the classes with too many images.

For what regards the image size, it has been found that the best trade-off is achieved considering 700×700 images, which are neither too large, thus not showing objects that are too little to be recognized, nor too small, thus allowing a good generalization power.

Different experiments have been performed also to understand the optimal amount of data augmentation. It has been found that applying RandAugment five times makes the model increase its performances without incurring into overfitting.

Once the multilabel classification task has been completed, the obtained results have demonstrated that the model learned very good features, being able to recognize the objects of interest even in the worst quality images. Moreover, due to the usage of Global Average Pooling at the end of the network, the model demonstrated to retain a remarkable localization ability until the final layer, thus allowing to identify the discriminative image regions leading to a prediction. This made possible the exploitation of the produced class activation maps to also assess the feasibility of the weakly supervised object detection task, thus without training the model on the ground truth bounding boxes, but only utilizing the features learned within the classification task.

In the end, the proposed solution allowed to reach an F_1 score of 81% on average within the multilabel task, also keeping the values related to the Precision and Recall really balanced, and a maximum Component Intersection over Union of 23.55% within the CAMs analysis for the weakly supervised object detection task. These results, together with the qualitative analysis performed, demonstrate the feasibility of automating the monitoring process for illegal landfills, showing that the combination of remote sensing images with Deep Learning techniques brings the achievement of this goal very close.

8.1 Future Work

In addition to a progressive improvement of the DL architectures in general, advances in the results illustrated in this research can be obtained in these specific areas:

Expansion of the categories list

As demonstrated during the study, this field is rich in categories that can be considered and distinguished among wastes, thus, when having enough data it is possible to experiment with a larger number of classes, also considering the ones that have been left away, such as tires and plastic bags, that are still very dangerous in terms of the environment.

Expansion of the hyperparameters fine-tuning

In this study it has been tried to use the best set of parameters, however, many improvements can be still done in this direction. For example, gradient accumulation techniques could be applied in order to increase the maximum batch size (BS).

Fine-grained classification

This field is highly affected by interclass similarity, and this has brought to the merging of pairs of categories that demonstrated to be nearly indistinguishable in many cases. Fine-grained classification is a sub-field of object recognition that aims at distinguishing subordinate categories within entry-level categories, and thus could be used to try to increase the performance of the model without merging the categories.

Semi-supervised learning dataset expansion

Given the good quality of the features that have been learned by the model and the increasing relevance that the field of semi-supervised learning is assuming, it would be for sure an interesting approach applying an iterative procedure consisting of executing the model on unseen and unlabeled images, to be then used to integrate and enlarge the dataset at each iteration. The combination of soft and hard labels can be in fact exploited to account for potentially wrong labels in the training set, giving the model the possibility to always add something more to learn at each iteration of the process.

Super Resolution

Super Resolution is the process of upscaling and/or improving the details within an image. Often a low-resolution image is taken as an input and the same image is scaled to a higher resolution. Details in the high-resolution output are filled in where they are essentially unknown.

Having seen that the model is more inclined to predict positively on Mapbox images because of the lower resolution, Super Resolution techniques could be adopted to make the model predictions more homogeneous.

Bibliography

[1] Shynggys Abdukhamet. Landfill detection in satellite images using deep learning. Master's thesis, Shanghai Jiao Tong University, 2019.

[2] Domenico Affinito. Lombardia, i furbetti dei rifiuti Â≪pizzicatiÂ≫ dal cielo con il satellite spia. *Corriere della Sera - Data Room*, 2021.

[3] A. Alfarrarjeh, S. H. Kim, S. Agrawal, M. Ashok, S. Y. Kim, and C. Shahabi. Image classification to determine the level of street cleanliness: A case study. In *2018 IEEE Fourth International Conference on Multimedia Big Data (BigMM)*, pages 1–5, 2018.

[4] G Allgaier and R Stegmann. Old landfills in the focus of the urban land management. In *Workshop on Landfill Reclamation and Remediation Technologies (International Waste Working Group and University of Padova), junio*, pages 7–9, 2006.

[5] C. V. Angelino, M. Focareta, S. Parrilli, L. Cicala, G. Piacquadio, G. Meoli, and M. De Mizio. A case study on the detection of illegal dumps with GIS and remote sensing images. In Ulrich Michel and Karsten Schulz, editors, *Earth Resources and Environmental Remote Sensing/GIS Applications IX*, volume 10790, pages 165 – 171. International Society for Optics and Photonics, SPIE, 2018.

[6] M. Anjum and M. S. Umar. Garbage localization based on weakly supervised learning in deep convolutional neural network. In *2018 International Conference on Advances in Computing, Communication Control and Networking (ICACCCN)*, pages 1108–1113, 2018.

[7] Liviu Apostol and Florin Mihai. The process of closing down rural landfills. case study: Neamt county. *Present Environment and Sustainable Development*, 5:167–174, 01 2011.

[8] J Baird, R Curry, and P Cruz. An overview of waste crime, its characteristics, and the vulnerability of the eu waste sector. *Waste Management & Research*, 32(2):97–105, 2014. PMID: 24519223.

[9] R. F. Berriel, A. T. Lopes, A. F. de Souza, and T. Oliveira-Santos. Deep learning-based large-scale automatic satellite crosswalk classification. *IEEE Geoscience and Remote Sensing Letters*, 14(9):1513–1517, 2017.

[10] Giancarlo Biotto, Sonia Silvestri, Lucia Gobbo, Elisa Furlan, Sonia Valenti, and Roberto Rosselli. Gis, multi-criteria and multi-factor spatial analysis for the probability assessment of the existence of illegal landfills. *International Journal of Geographical Information Science*, 23(10):1233–1244, 2009.

[11] Thomas Blaschke and Josef Strobl. What's wrong with pixels? some recent developments interfacing remote sensing and gis. *GIS - Zeitschrift fÃ¼r Geoinformationssysteme*, 14:12 – 17, 06 2001.

[12] R. Brooks, Russell Creiner, and T. Binford. The acronym model-based vision system. In *IJCAI*, 1979.

[13] E. G. Cadau, C. Putignano, R. Aurigemma, A. Melchiorre, P. Bosco, A. Tesseri, and F. Battazza. Simdeo: An integrated system for landfill detection and monitoring using eo data. In *2013 IEEE International Geoscience and Remote Sensing Symposium - IGARSS*, pages 3305–3308, 2013.

[14] G. Cheng, X. Xie, J. Han, L. Guo, and G. S. Xia. Remote sensing image scene classification meets deep learning: Challenges, methods, benchmarks, and opportunities. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 13:3735–3756, 2020.

[15] Gong Cheng, Junwei Han, Lei Guo, and Tianming Liu. Learning coarse-to-fine sparselets for efficient object detection and scene classification. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1173–1181, 2015.

[16] Gong Cheng, Peicheng Zhou, and Junwei Han. Learning rotation-invariant convolutional neural networks for object detection in vhr optical remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 54(12):7405–7415, 2016.

[17] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced

search space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 702–703, 2020.

[18] A. Dabholkar, B. Muthiyan, S. Srinivasan, S. Ravi, H. Jeon, and J. Gao. Smart illegal dumping detection. In *2017 IEEE Third International Conference on Big Data Computing Service and Applications (BigDataService)*, pages 255–260, 2017.

[19] Kh. M. Dewidar. Landfill detection in hurghada, north red sea, egypt, using thematic mapper images. *International Journal of Remote Sensing*, 23(5):939–948, 2002.

[20] Vincenzo Di Fiore, Giuseppe Cavuoto, Michele Punzo, Daniela Tarallo, Marco Casazza, Silvio Marco Guarriello, and Massimiliano Lega. Integrated hierarchical geo-environmental survey strategy applied to the detection and investigation of an illegal landfill: A case study in the campania region (southern italy). *Forensic Science International*, 279:96–105, 2017.

[21] M Doak, S Khan, G Kelly, and S Silvestri. The use of remote sensing to map illegal landfills at the border of ireland/northern ireland. 2007.

[22] Europol. Trash worth millions of euros. 2019.

[23] M.A. Fischler and R.A. Elschlager. The representation and matching of pictorial structures. *IEEE Transactions on Computers*, C-22(1):67–92, 1973.

[24] Paolo Gamba. Human settlements: A global challenge for eo data processing and interpretation. *Proceedings of the IEEE*, 101(3):570–581, 2013.

[25] Fethi Ghazouani, Imed Riadh Farah, and Basel Solaiman. A multi-level semantic scene interpretation strategy for change interpretation in remote sensing imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 57(11):8775–8795, 2019.

[26] Jasravia Gill, Kamil Faisal, Ahmed Shaker, and Wai Yeung Yan. Detection of waste dumping locations in landfill using multi-temporal landsat thermal images. *Waste Management & Research*, 37(4):386–393, 2019. PMID: 30632930.

[27] Katharine Glanville and Hsing-Chung Chang. Remote sensing analysis techniques and sensor requirements to support the mapping of illegal

domestic waste disposal sites in queensland, australia. *Remote Sensing*, 7(10):13053–13069, 2015.

[28] Luis Gomez-Chova, Devis Tuia, Gabriele Moser, and Gustau Camps-Valls. Multimodal classification of remote sensing images: A review and future directions. *Proceedings of the IEEE*, 103(9):1560–1584, 2015.

[29] Luis Gómez-Chova, Devis Tuia, Gabriele Moser, and Gustau Camps-Valls. Multimodal classification of remote sensing images: A review and future directions. *Proceedings of the IEEE*, 103(9):1560–1584, 2015.

[30] P. Grandjean and P. J. Landrigan. Developmental neurotoxicity of industrial chemicals. *The Lancet*, 368(9553):2167–2178, Dec 2006.

[31] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning*, pages 1321–1330. PMLR, 2017.

[32] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[33] Donald Olding Hebb and DO Hebb. *The organization of behavior*, volume 65. Wiley New York, 1949.

[34] Qiong Hu, Wu Wenbin, Tian Xia, Qiangyi Yu, Peng Yang, Zhengguo li, and Qian Song. Exploring the use of google earth imagery and object-based methods in land use/cover mapping. *Remote Sensing*, 5:6026–6042, 11 2013.

[35] Qiong Hu, Wenbin Wu, Tian Xia, Qiangyi Yu, Peng Yang, Zhengguo Li, and Qian Song. Exploring the use of google earth imagery and object-based methods in land use/cover mapping. *Remote Sensing*, 5(11):6026–6042, 2013.

[36] Xin Huang, Dawei Wen, Jiayi Li, and Rongjun Qin. Multi-level monitoring of subtle urban changes for the megacities of china using high-resolution multi-view satellite imagery. *Remote Sensing of Environment*, 196:56–75, 2017.

[37] Minhe Ji and John R Jensen. Effectiveness of subpixel analysis in detecting and quantifying urban imperviousness from landsat thematic mapper imagery. *Geocarto International*, 14(4):33–41, 1999.

[38] Rosa JordÃ¡-Borrell, Francisca Ruiz-RodrÃguez, and Angel LuÃs Lucendo-Monedero. Factor analysis and geographic information system for determining probability areas of presence of illegal landfills. *Ecological Indicators*, 37:151–160, 2014.

[39] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105, 2012.

[40] A. Y. Kwarteng and A. Al-Enezi. Assessment of kuwait's al-qurain landfill using remotely sensed data. *Journal of Environmental Science and Health, Part A*, 39(2):351–364, 2004.

[41] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[42] M. Lega, C. Ferrara, G. Persechino, and P. Bishop. Remote sensing in environmental police investigations: aerial platforms and an innovative application of thermography to detect several illegal activities. *Environmental Monitoring and Assessment*, 186(12):8291–8301, Dec 2014.

[43] Deren Li, Mi Wang, Zhipeng Dong, Xin Shen, and Lite Shi. Earth observation brain (eob): an intelligent earth observation system. *Geospatial Information Science*, 20(2):134–140, 2017.

[44] Xiaoxiao Li and Guofan Shao. Object-based urban vegetation mapping with high-resolution aerial photography as a single data source. *International journal of remote sensing*, 34(3):771–789, 2013.

[45] A. Limoli, E. Garzia, A. De Pretto, and C. De Muri. Illegal landfill in italy (eu)-a multidisciplinary approach. *Environmental Forensics*, 20(1):26–38, 2019.

[46] ARPA Lombardia. Ia nel progetto savager, collaborazione tra arpa lombardia e politecnico di milano. 2020.

[47] Nathan Longbotham, Chuck Chaapel, Laurence Bleiler, Chris Padwick, William J Emery, and Fabio Pacifici. Very high resolution multiangle urban classification analysis. *IEEE Transactions on Geoscience and Remote Sensing*, 50(4):1155–1170, 2011.

[48] David G Lowe. Three-dimensional object recognition from single two-dimensional images. *Artificial intelligence*, 31(3):355–395, 1987.

[49] David G Lowe. Proceedings of the seventh ieee international conference on computer vision. In *Proceedings of the Seventh IEEE International Conference on Computer Vision*, volume 1, pages i–, 1999.

[50] Angel LuÃs Lucendo-Monedero, Rosa JordÃ¡-Borrell, and Francisca Ruiz-RodrÃguez. Predictive model for areas with illegal landfills using logistic regression. *Journal of Environmental Planning and Management*, 58(7):1309–1326, 2015.

[51] Zhi Yong Lv, Wenzhong Shi, Xiaokang Zhang, and Jón Atli Benediktsson. Landslide inventory mapping from bitemporal high-resolution remote sensing images using change detection and multiscale segmentation. *IEEE journal of selected topics in applied earth observations and remote sensing*, 11(5):1520–1532, 2018.

[52] C. Manzo, A. Mei, E. Zampetti, C. Bassani, L. Paciucci, and P. Manetti. Top-down approach from satellite to terrestrial rover application for environmental monitoring of landfills. *Science of The Total Environment*, 584-585:1333–1348, 2017.

[53] Tapas Ranjan Martha, Norman Kerle, Cees J van Westen, Victor Jetten, and K Vinod Kumar. Segment optimization and data-driven thresholding for knowledge-based landslide detection by object-based image analysis. *IEEE Transactions on Geoscience and Remote Sensing*, 49(12):4928–4943, 2011.

[54] Niti B Mishra and Kelley A Crews. Mapping vegetation morphology types in a dry savanna ecosystem: Integrating hierarchical object-based image analysis with random forest. *International Journal of Remote Sensing*, 35(3):1175–1198, 2014.

[55] Alberto Federico Pagani. A system for annotating and analysing multisource geo-referenced images for environmental applications. 2019.

[56] Giuseppe Persechino, Massimiliano Lega, Gianpaolo Romano, Francesco Gargiulo, and Luca Cicala. Ides project: an advanced tool to investigate illegal dumping. *WIT Transactions on Ecology and the Environment*, pages 603–614, 05 2013.

[57] Biserka Petrovska, Eftim Zdravevski, Petre Lameski, Roberto Corizzo, Ivan Å tajduhar, and Jonatan Lerga. Deep learning for feature extraction in remote sensing: A case-study of aerial scene classification. *Sensors*, 20(14), 2020.

[58] Lorenzo Carlos Quesada-Ruiz, Victor Rodriguez-Galiano, and Rosa JordÃ¡-Borrell. Characterization and mapping of illegal landfill potential occurrence in the canary islands. *Waste Management*, 85:506–518, 2019.

[59] Amy Richter, Kelvin Tsun Wai Ng, and Nima Karimi. A data driven technique applying gis, and remote sensing to rank locations for waste disposal site expansion. *Resources, Conservation and Recycling*, 149:352–362, 2019.

[60] Lawrence Roberts. *Machine Perception of Three-Dimensional Solids*. 01 1963.

[61] J. Rocher, D. A. Basterrechea, M. Taha, M. Parra, and J. Lloret. Water conductivity sensor based on coils to detect illegal dumpings in smart cities. In *2019 Fourth International Conference on Fog and Mobile Edge Computing (FMEC)*, pages 324–329, 2019.

[62] Frank Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386, 1958.

[63] Showmitra Kumar Sarkar and Md Esraz-Ul-Zannat. *Managing Municipal Waste : Application of Spatial Tools and Techniques*. PhD thesis, Khulna University of Engineering & Technology, Bangladesh, 06 2019.

[64] L. Selani. *Mapping Illegal Dumping Using a High Resolution Remote Sensing Image Case Study: Soweto Township in South Africa*. University of the Witwatersrand, Faculty of Science, School of Geography, Archaeology & Environmental Studies, 2017.

[65] S Silvestri and M Omri. A method for the remote sensing identification of uncontrolled landfills: formulation and validation. *International Journal of Remote Sensing*, 29(4):975–989, 2008.

[66] Amin Tayyebi, Bryan Christopher Pijanowski, and Amir Hossein Tayyebi. An urban growth boundary model using neural networks, gis and radial parameterization: An application to tehran, iran. *Landscape and Urban Planning*, 100(1):35–44, 2011.

[67] Devis Tuia, Michele Volpi, Loris Copa, Mikhail Kanevski, and Jordi Munoz-Mari. A survey of active learning algorithms for supervised remote sensing image classification. *IEEE Journal of Selected Topics in Signal Processing*, 5(3):606–617, 2011.

[68] Sergij Vambol, Viola Vambol, Muniyan Sundararajan, and Iqbal Ansari. The nature and detection of unauthorized waste dump sites using remote sensing. *Ecological Questions*, 30(3):43–55, 2019.

[69] Andrea Viezzoli, Anders Edsen, Esben Auken, and Sonia Silvestri. The use of satellite remote sensing and helicopter tem data for the identification and characterization of contaminated. *Near Surface 2009 - 15th European Meeting of Environmental and Engineering Geophysics*, 09 2009.

[70] Noor Yasmin Zainun, Ismail Abdul Rahman, and Rosfazreen Azwana Rothman. Mapping of construction waste illegal dumping using geographical information system (GIS). *IOP Conference Series: Materials Science and Engineering*, 160:012049, nov 2016.

[71] Tamara Zelenovic, Zorica Srdjevic, Ratko Bajcetic, and Mirjana Miloradov. Gis and the analytic hierarchy process for regional landfill site selection in transitional countries: a case study from serbia. *Environmental Management*, 01 2011.

[72] Barbora Å edovÃ¡. On causes of illegal waste dumping in slovakia. *Journal of Environmental Planning and Management*, 59(7):1277–1303, 2016.

Appendix A

Low Occurring Classes

It has been deeply analyzed the possibility of taking into consideration also the categories that have a very low number of samples in the dataset:

- Hay bales
- Tubes
- Wood
- Tires
- Grouped cars
- Plastic bags

The aim is that of finding enough samples to allow the model to learn good features also for them. In particular, in Table A.1 is shown the initial situation of the dataset. As it is possible to see, the last six categories have a very low number of samples with respect to the first 3 and considering that it is also very important the balance between categories in a dataset (it directly affects the performance of the model), ideally, it is necessary to have a number of samples that is as near as possible to the numbers of the first 6 categories.

	Ortophoto	*Mapbox*	*Total*
Cisterns	246	241	487
Scattered waste	1179	1061	2240
Pallets	457	503	960
Caissons	555	502	1057
General waste	623	750	1373
Containers	140	127	267
Hay bales	3	152	155
Tubes	8	93	101
Wood	0	113	113
Tires	0	27	27
Grouped cars	0	20	20
Plastic bags	0	47	47

Table A.1: Initial dataset table

In this Appendix, the Tires category will be taken under consideration as an example. The metrics that have been obtained for this category before performing the following study are shown in Table A.2.

Average Precision	*Precision*	*Recall*	F_1*-score*
1.00	0.67	1.00	0.80

Table A.2: Initial model performance on the Tires class - Test set

To improve the model performance four main paths have been explored:

- Look at all the available images still unlabeled and manually label them in case the 6 categories of interest appear.
- Look for coordinates of places where those categories might appear in Google Maps.
- Train a binary classification model for each of the 6 categories and execute it on new territory.
- Try a particular data augmentation technique.

A.1 Look at all the available images still unlabeled and manually label them in case the 6 categories of interest appear

It has been used the tagger tool in order to analyze 7 campaigns. Table A.3 shows the number of unlabelled images that have been checked to find other examples of the low occurring categories.

	Images per campaign
Pavia Ortophoto	636
Lodi Ortophoto	122
Brescia Ortophoto	156
Piemonte Mapbox	806
Pavia Mapbox	636
Lodi Mapbox	122
Campania Mapbox	244
Total	2722

Table A.3: Total number of unlabeled images that have been checked

The results of this analysis, and so the number of samples per category that have been found, are shown in Table A.4.

	New images per category	*Frequency*
Hay bales	29	1.06%
Tubes	3	0.11%
Wood	7	0.26%
Tires	25	0.92%
Grouped cars	3	0.11%
Plastic bags	7	0.26%

Table A.4: Total number of added samples per category

As it can be seen from the numbers obtained, the frequency of the considered categories is very low with respect to the total number of images

seen. It is thus evident that even having twice as many images as those considered, the number of obtained samples would be not sufficient for even having acceptable results in terms of performance. Another consideration is that having seen such a large portion of the Lombardy region without finding an even slightly significant number of samples of these categories could mean that the disposal of these types of wastes is not a very diffused problem until now.

As an example, in Table A.5 are shown the metrics that have been obtained on the Tires category after having added the new images (in the next Section the same model will be run on an unseen territory showing that these high metrics are due to complete overfitting of the dataset).

Average Precision	*Precision*	*Recall*	F_1*-score*
1.00	1.00	1.00	1.00

Table A.5: Model performance on the Tires class after first experiment - Test set

A.2 Look for coordinates of places where those categories could appear in Google Maps

The procedure has been that of looking in Google Maps for places where it's likely that those categories appear, such as, for example, for Tires, it has been looked for tire dealers.

Even in this case the results have been really poor: looking in three different regions (Lombardy, Piedmont, and Campania) an average of only 20 pairs of coordinates have been found for each category. In most of the cases, however, considering the corresponding images from Mapbox and Orthophoto, most of them did not show the same content seen in Google Maps, thus further reducing the number of additional samples available for each category to about 4 on the initial 20. Thus, also in this case, there is a really low number of these sites and also not all of them contain what is needed.

An example for the category Tires is shown in image A.1.

Figure A.1: Images from Google Maps example - Google image on the left, Orthophoto image in the middle and Mapbox image on the right

In conclusion, also this path revealed not to be a good way to find the needed number of samples.

A.3 Train a binary classification model for each of the 6 categories and execute it on a new territory

In this experiment, 6 different binary classification models have been trained, one for each category of interest. Each dataset was composed of all the available images in the respective category and the number of negatives doubled with respect to the number of positives. Once trained, the models have been run on new territory in order to see if it was possible to automatically detect new samples.

Also here however results were really bad since most of the models predicted all the new images as positives. In all the cases the obtained results were of no sense at all, thus further demonstrating the problem of having so few images even in a simpler task like binary classification. In conclusion, even this path has been discarded. As an example, running the binary classification model for the Tires category, the results shown in Table A.6 have been obtained.

Total images	7843
Images with prediction over 0.9	964
Images with prediction in range (0.8, 0.9]	1445
Images with prediction in range (0.7, 0.8]	5401
Images with prediction in range (0.3, 0.7]	33
Images with prediction in range (0.2, 0.3]	0
Images with prediction in range [0.1, 0.2]	0
Images with prediction under 0.1	1 0

Table A.6: Number of predictions per interval

The metrics are the same as in Table A.5, since the same model has been used. Figure A.2 shows some examples of positively predicted images.

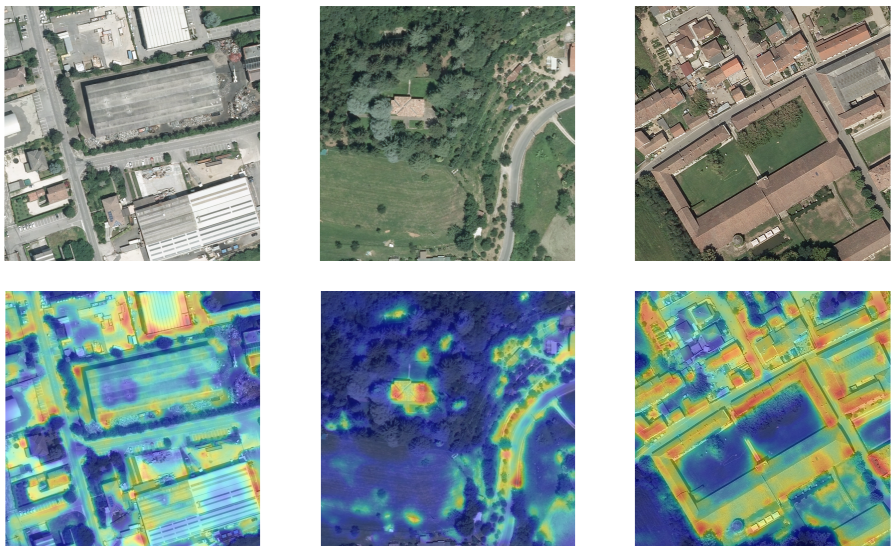

Figure A.2: Examples of positively predicted images

A.4 Try a particular data augmentation technique

In order to leave no stone unturned, a particular augmentation technique has also experimented that consists of cropping the objects of the considered categories from the available images and pasting them on other images, paying attention to not superpose them on other annotations. This allows a sort of data augmentation giving a wider variability in terms of context,

which might help the model to see more varying images and so to learn better features.

Figure A.3 shows an example of an image created through this technique for the Tires class.

Figure A.3: Example of augmented Tire image

In Table A.7 is shown the number of samples that have been added using this technique.

	New images per category	*Total images*
Hay bales	136	291
Tubes	143	244
Wood	189	302
Tires	158	185
Grouped cars	164	184
Plastic bags	164	211

Table A.7: Total number of added samples per category with the particular augmentation technique

The datasets have been built by using all the available images for each of the six categories and adding a number of negative samples doubled with respect to the number of positive ones. After having created again 6 different datasets for performing binary classification on the single categories and after having trained the model on them, the results demonstrated the poor validity, in the case under analysis, of this technique. The model in fact completely overfits the dataset, not being able at all to predict new territories. As already happened in the previous experiment, images are always predicted as positives, thus making the model completely useless.

In order to give a more concrete example, Table A.8 shows the numbers obtained running the Tires model on new territory.

Total images	1507
Images with prediction over 0.9	1494
Images with prediction in range (0.8, 0.9]	8
Images with prediction in range (0.7, 0.8]	1
Images with prediction in range (0.3, 0.7]	2
Images with prediction in range (0.2, 0.3]	1
Images with prediction in range [0.1, 0.2]	1
Images with prediction under 0.1	0 0

Table A.8: Number of predictions per interval with the particular augmentation technique

The metrics obtained are shown in Table A.9.

Average Precision	*Precision*	*Recall*	F_1*-score*
1.00	1.00	1.00	1.00

Table A.9: Model performance on the Tires class after experiment with the particular augmentation technique - Test set

Figure A.4 shows some examples of images predicted as positives with the respective CAMs, demonstrating that the model is learning nothing good.

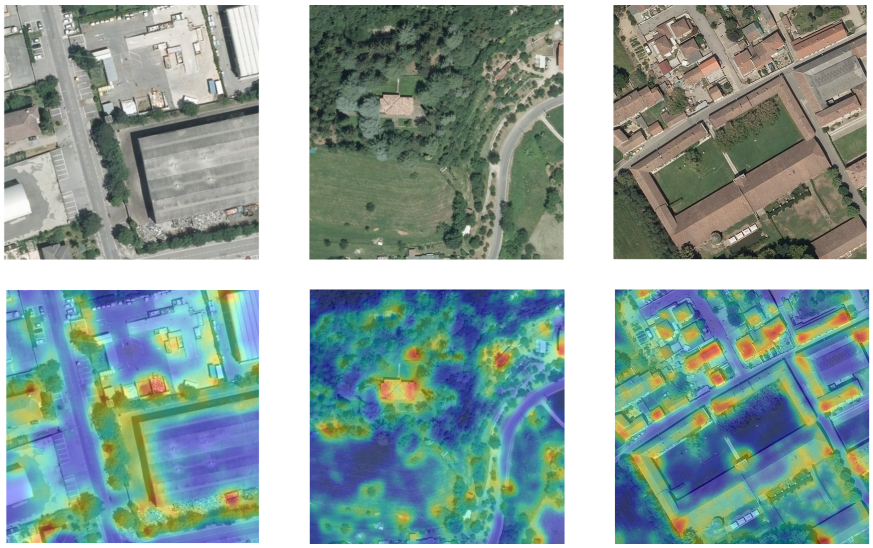

Figure A.4: Examples of positively predicted images with the particular augmentation technique

A.5 Conclusions

After having performed a very deep analysis, trying every possible path and tool available and without achieving any good result that would give a minimum of hope, it is necessary to conclude that for the moment it is not possible and not feasible to consider the categories mentioned at the beginning of the Appendix (Hay bales, Tubes, Wood, Tires, Grouped cars, Plastic bags) for the Illegal Landfills multilabel classification task.

Appendix B

Similarity of Classes

During the evolution of the study, it has been found that some of the considered classes are very similar to each other. The categories in question are:

- Cisterns and Pallets
- Scattered and General waste
- Caissons and Containers

This has been demonstrated quantitatively by performing eight specific experiments, merging with different configurations the pairs of classes:

- No merges and General waste category removed
- Only Cisterns and Pallets merged
- Only Scattered waste and General waste merged
- Only Caissons and Containers merged
- Cisterns - Pallets, Scattered waste - General waste merged
- Cisterns - Pallets, Caissons - Containers merged
- Scattered waste - General waste, Caissons - Containers merged
- Cisterns - Pallets, Scattered waste - General waste, Caissons - Containers merged

In the following experiments, the obtained results will be always compared with the ones obtained without making any modification of the ones listed above, thus just considering the six separated categories. Table B.1 shows the metrics obtained on the test set in the case described.

	Test Set			
	Accuracy	*Precision*	*Recall*	*F1-score*
Cisterns	0.89	0.53	0.28	0.36
Scattered waste	0.80	0.62	0.48	0.54
Pallets	0.82	0.70	0.28	0.40
Caissons	0.79	0.60	0.54	0.57
General waste	0.81	0.60	0.61	0.60
Containers	0.94	0.67	0.17	0.27

Table B.1: Metrics of the comparative model

B.1 No merges and General waste category removed

Since the category General waste includes many different types of objects, the first investigation has been that of understanding if considering this category could bring much confusion to the model. So, the experiment carried out is that of maintaining the same configuration used in Table B.1, but removing the General waste category. The obtained results on the test set are shown in Table B.2.

	Test Set			
	Accuracy	*Precision*	*Recall*	*F1-score*
Cisterns	0.87	0.47	0.43	0.45
Scattered waste	0.83	0.67	0.50	0.57
Pallets	0.85	0.63	0.66	0.65
Caissons	0.82	0.61	0.54	0.57
Containers	0.95	0.58	0.41	0.48

Table B.2: No merges and General waste category removed

Comparing Table B.1 and B.2, it is not possible to see a significant improvement, either in terms of single metrics nor in terms of balancement between Precision and Recall.

B.2 Single pairs of categories merged

In the second, third, and fourth experiments, a single pair of categories has been merged at a time, with the aim of understanding whether there is a pair that is responsible for most of the confusion introduced.

The results on the test sets are respectively shown in Tables B.3, B.4, B.5.

	Test Set			
	Accuracy	*Precision*	*Recall*	*F1-score*
Cisterns and Pallets	0.87	0.75	0.67	0.71
Scattered waste	0.84	0.72	0.70	0.71
Caissons	0.82	0.58	0.62	0.60
General waste	0.80	0.66	0.38	0.48
Containers	0.96	0.78	0.28	0.41

Table B.3: Single pairs of categories merged - Cisterns and Pallets

	Test Set			
	Accuracy	*Precision*	*Recall*	*F1-score*
Cisterns	0.87	0.45	0.49	0.47
Scattered and General waste	0.87	0.74	0.78	0.76
Pallets	0.84	0.59	0.74	0.66
Caissons	0.83	0.58	0.59	0.59
Containers	0.96	0.67	0.43	0.52

Table B.4: Single pairs of categories merged - Scattered waste and General waste

	Test Set			
	Accuracy	*Precision*	*Recall*	*F1-score*
Cisterns	0.90	0.42	0.23	0.30
Scattered waste	0.83	0.77	0.44	0.56
Pallets	0.84	0.61	0.58	0.59
Caissons and Containers	0.82	0.61	0.67	0.64
General waste	0.80	0.59	0.58	0.59

Table B.5: Single pairs of categories merged - Caissons and Containers

Looking at the results, it is possible to state that there is not a pair of categories that, with the merging, predominantly improves the performance of the model. In fact, it seems that the most significant improvement is in the merged categories, while for what regards the others, they tend to have metrics really close to the ones achieved in the experiment without merges.

B.3 Two pairs of categories merged

In the fifth, sixth, and seventh experiments, a combination of two pairs of categories have been merged at a time.

The results on the test sets are respectively shown in Tables B.6, B.7, B.8.

	Test Set			
	Accuracy	*Precision*	*Recall*	*F1-score*
Cisterns and Pallets	0.85	0.82	0.65	0.73
Scattered and General waste	0.84	0.78	0.73	0.75
Caissons	0.84	0.65	0.88	0.75
Containers	0.95	0.60	0.25	0.35

Table B.6: Two pairs of categories merged - Cisterns - Pallets and Scattered waste - General waste

	Test Set			
	Accuracy	*Precision*	*Recall*	*F1-score*
Cisterns and Pallets	0.84	0.74	0.64	0.68
Scattered waste	0.82	0.67	0.55	0.60
Caissons and Containers	0.85	0.65	0.75	0.69
General waste	0.82	0.62	0.58	0.60

Table B.7: Two pairs of categories merged - Cisterns - Pallets and Caissons - Containers

	Test Set			
	Accuracy	*Precision*	*Recall*	*F1-score*
Cisterns	0.91	0.53	0.41	0.46
Scattered and General waste	0.84	0.70	0.81	0.75
Pallets	0.86	0.69	0.79	0.74
Caissons and Containers	0.85	0.83	0.64	0.72

Table B.8: Two pairs of categories merged - Scattered waste - General waste and Caissons - Containers

Also in this case the results demonstrate that only the categories that have been merged have a significant improvement, while the other ones maintain similar metrics as in experiments where categories have been merged. This confirms the fact that all three pairs of categories have an equivalent influence on the model performance and that each pair introduce confusion independently with respect to the others.

B.4 All the pairs merged

The last experiment has been that of merging all three pairs of categories. Here, as expected, the most significant results have been obtained, achieving metrics that are both high and balanced between each other. Results obtained on the test set are reported in Table B.9.

	Test Set			
	Accuracy	*Precision*	*Recall*	*F1-score*
Cisterns and Pallets	0.88	0.82	0.78	0.80
Scattered and General waste	0.88	0.77	0.91	0.83
Caissons and Containers	0.88	0.75	0.83	0.79
Micro Average	0.88	0.78	0.84	0.81

Table B.9: All the pairs merged

After having seen the results, and considering also the qualitative analysis made on examples of images, which has shown the high difficulty that also the human eye makes to distinguish between categories belonging to the same pair, this configuration has been adopted. The decision is also justified by the fact that, in most of the cases, if a category is predicted, also the corresponding one belonging to the same pair tends to be predicted, thus making the choice of keeping them separated of no sense.

In order to provide a concrete demonstration of what has been said, Table B.10 displays the *co-occurrence matrix*, that shows the frequency with which categories are predicted together.

	Cisterns	*Scattered waste*	*Pallets*	*Caissons*	*General waste*	*Containers*	*Support*
Cisterns	100%	45%	41%	36%	40%	10%	375
Scattered waste	19%	100%	41%	57%	44%	10%	866
Pallets	21%	49%	100%	40%	40%	9%	724
Caissons	16%	58%	34%	100%	40%	11%	859
General waste	17%	44%	34%	40%	100%	11%	859
Containers	18%	41%	29%	43%	44%	100%	219

Table B.10: Co-occurrence matrix

As it can be seen in the reported co-occurrence matrix, where each row defines the category under consideration, while columns define the frequency with which the other categories appear together with the one considered, in most of the cases the similar categories are both present in the same image. In particular, apart from Containers and Cisterns that have a low number of samples with respect to other categories (as it can be noticed from the support column), and thus a lower co-occurrence frequency in each case, nearly half of the times, if a category is present, also the similar one appears.

In order to also give an overall vision of the confusion introduced keeping the categories separated, a variant of the *confusion matrix* is shown in Table B.11. This variant focuses on false positives predictions, telling which are the categories that are actually present in the image when the one under consideration is predicted as false positive.

	Cisterns	*Scattered waste*	*Pallets*	*Caissons*	*General waste*	*Containers*	*Support*
Cisterns	0%	9%	30%	16%	17%	13%	375
Scattered waste	52%	0%	61%	50%	63%	33%	866
Pallets	54%	41%	0%	47%	51%	33%	724
Caissons	48%	45%	59%	0%	59%	47%	859
General waste	50%	59%	54%	56%	0%	47%	859
Containers	9%	11%	8%	11%	8%	0%	219
Negatives	1%	0%	0%	0%	0%	7%	219

Table B.11: Variant of confusion matrix

Here it is also more clear that when a category is predicted as false positive (row), in most of the cases there is its similar category in the image (column). As an example, from the table is clearly visible the difficulty that the model does in distinguishing Scattered and General waste. Looking at the second row in fact, when the category Scattered waste is predicted, in 63% of the cases the category General waste is actually present in the image.