Deep Learning Techniques for Analysis of Urban Safety Perception

Felipe Adrian Moreno Vera

Advisor: Dr. Jorge Luis Poco Medina

Committee:
Dr. José Eduardo Ochoa Luna – Universidad Católica San Pablo – Perú
Dr. Guillermo Cámara Chavez – Universidade Federal de Ouro Preto – Brasil
Dr. Rensso Victor Hugo Mora Colque – Universidad Católica San Pablo – Perú

Thesis presented to the
Department of Computer Science
as part of the requirements to obtain the degree of
Master of Computer Science.

Catholic University San Pablo – UCSP
May 2024 – Arequipa – Perú

This work is dedicated to my parents, siblings, and other family members who have always been supporting and placing their trust in me; they will always be with me.

Acknowledgment

I appreciate the support and guidance I received throughout this time from my parents, siblings, advisor, and professors in general. Also, to the National **Council of Science, Technology and Technological Innovation** (CONCYTEC) and the **National Fund for Scientific, Technological, and Technological Innovation Development** (FONDECYT), which, through Management Agreement 234-2015-FONDECYT, enabled the subsidy and financing of my Master's studies in Computer Science at the **Catholic University San Pablo** (UCSP). The university provided me with a diverse environment of knowledge and people to interact with, allowing me to formulate, develop, and implement the present work. I would also like to express my gratitude to the **Getúlio Vargas Foundation** (FGV), Rio de Janeiro, Brazil, where I worked as a researcher using the computational environment provided during my stay in Brazil.

Acronyms

AUC *Area Under Curve*

CAM *Class Activation Map*

CNN *Convolutional Neural Network*

DCNN *Deep Convolutional Neural Network*

DeCAF *Deep Convolutional Activation Feature*

DSVR *Direction based Street View Retrieval*

FOV *Field of View*

GAP *Global Average Pooling*

GAN *Generative Adversarial Networks*

GBP *Guided Back-Propagation*

GEH *Global Edge Histogram*

GSV *Google Street View*

GVI *Green View Index*

HDI *Human Development Index*

HDMiR *Hierarchical Deep Multi-instance Regression*

HOG *Histogram of Oriented Gradients*

HPM *Human Perception Mapping*

KNN *K-Nearest Neighbors*

MLR *Multi Linear Regressor*

MTDRALN *MultiTask Deep Relative Attribute Learning Network*

PRN *Perception Rank Network*

RBF *Radial Basis Function*

RGB *Red-Green-Blue*

SG *SmoothGrad*

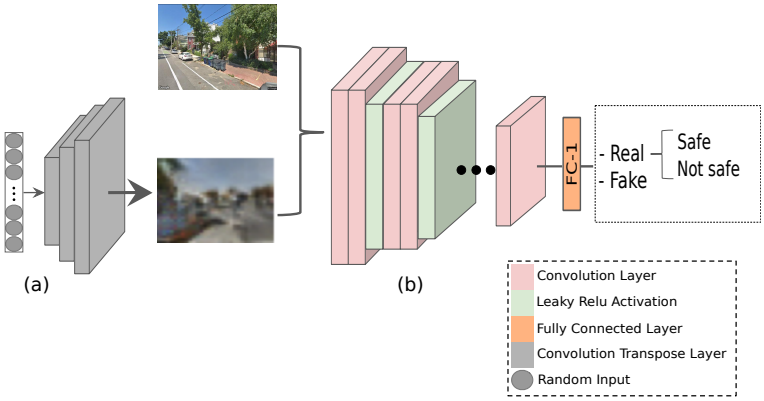
SAPN *Semantic-Aware Perception Network*

SVM *Support Vector Machine*

SVR *Support Vector Regressor*

SURF *Speeded Up Robust Features*

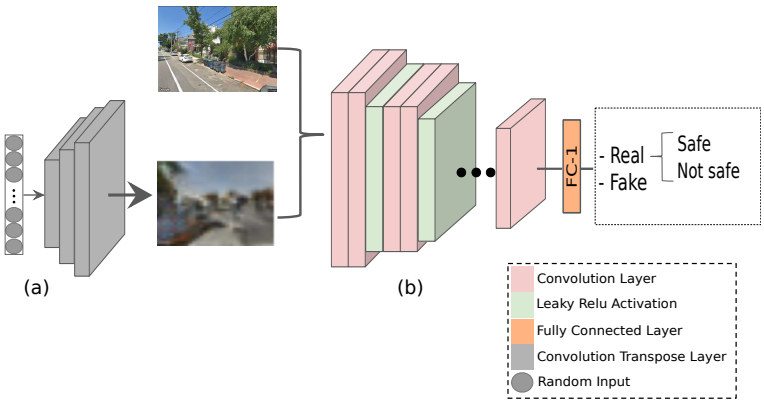
Resumen



La percepción es la forma en que los humanos interpretan y comprenden la información captada después de la interacción con el entorno que les rodea, aprendiendo nuevas experiencias o reforzando otras ya vividas. La percepción de la seguridad urbana se puede describir en cómo los humanos presentan una reacción ante un determinado estímulo proveniente de la apariencia visual o conocimiento previo sobre un cierto lugar (calles, zonas urbanas, etc). A partir de esta idea, diversos estudios buscaron describir dicho fenómeno teniendo como ejemplo más notable la teoría denominada *"The Broken Window"*, la cual estudiaba el comportamiento de las personas frente a ambientes cuya apariencia visual era caótica. Así mismo, recientemente este estudio está siendo implementado utilizando diversos tipos de datos, no solo limitándose a encuestas o experimentos sociales, con el objetivo de determinar la relación entre la percepción urbana y características intrínsecas de los ciudades; de los cuales, uno de los conjuntos de datos más resaltables es *Place Pulse*. En este trabajo, se propone una metodología que permita analizar y explorar los datos de *Place Pulse 2.0*. Como resultados principales, presentamos un análisis exploratorio de los datos, resaltando la organización y comportamiento de los datos. Además, presentamos una comparación entre diferentes técnicas de aprendizaje supervisado y semi-supervisado. Mostrando que un modelo *Generative Adversarial Networks* (GAN) presenta mejores resultados que técnicas convencionales.

Keywords: *Deep Learning*, *Convolutional Neural Networks*, *GAN*, *features extraction*, *urban perception*.

Abstract



Perception is how humans interpret and understand information from some environment. This information is captured after interacting with the environment that surrounds them, learning new experiences, or reinforcing others already lived. The perception of urban security can be described in how humans present a reaction to a particular stimulus from the visual appearance or prior knowledge of a specific place (streets, urban areas, etc.). Based on the previous idea, various studies sought to describe this phenomenon. A very notable example is the theory called *"The Broken Window"* which studied the behavior of people in environments whose visual appearance was chaotic. Likewise, recently this study has been implemented using various types of data, not only limited to surveys or social experiments, to determine the relationship between urban perception and intrinsic characteristics of cities. Which is one of the most noteworthy data sets is *Place Pulse*. In this work, we propose a methodology that allows the analysis and data exploration of *Place Pulse 2.0*. As the main results, we present an exploratory data set analysis, highlighting behavior and outliers. Besides, we show the comparison and training results of supervised and semi-supervised GAN-based models against other techniques. We are showing that our Semi-Supervised GAN approach presents better results in metrics and stability in dealing with this kind of data.

Keywords: *Deep Learning*, *Convolutional Neural Networks*, *GAN*, *features extraction*, *urban perception*.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

In this chapter, we describe the main motivations we have had for the development of our work. Additionally, we will also define the problem that we intend to address and the objectives that we aim to achieve in this work.

1.1 Context and Motivation

"Cities are designed to shape and influence the lives of their inhabitants" (Lindal y Hartig, 2013). Several studies have shown that the visual appearance of cities plays a central role in human perception and reaction to the environment. A notable example is the Broken Window theory (Wilson y Kelling, 1982) which suggests that visual signs of environmental disorder, such as broken windows, abandoned cars, litter, and graffiti, can induce negative social outcomes and increase crime levels. This theory has had a significant influence on public policy strategies leading to aggressive policing tactics to control manifestations of social and physical disorder. For example, in the study conducted in the city of New York (Keizer et al., 2008), social experiments were carried out on the perceived quality of life on the streets. These experiments compared "impeccable" places such as shopping centers (clean walls, orderly and quiet) with other places where there was the presence of graffiti, old or neglected streets, and litter on the streets, concluding that in places where "rules are violated", in the long term, social norms are not respected. In conclusion, it is determined that the neglected visual appearance negatively influences the environment (e.g., graffiti, litter scattered on the streets, lack of cleanliness in the environment, etc.).

Similarly, other studies have shown that the visual appearance of urban spaces affects the psychological state of its inhabitants (Lindal y Hartig, 2013). For example, a psychological evaluation demonstrated that the presence of green areas in cities tends to produce positive sensations in their inhabitants, such as safety, relaxation, tranquility, etc. (Ulrich, 1979). On the other hand, through the study of 40 psychological reports based on surveys and studies of the mental states of its inhabitants, it was also deduced

that urban disorder induces psychological distress, stress, and constant fear (Sampson et al., 2002). In addition, it has also been shown that graffiti and buildings in poor or abandoned conditions are directly related to the perception of insecurity (Schroeder y Anderson, 1984).

Therefore, through various studies on the impact of the visual appearance of a city on its inhabitants, it becomes of vital importance to understand the perceptions and evaluations of urban spaces by people. In this sense, several studies have been carried out on how the city and its visual appearance influence the behavior of the city. For example, in "The Image of the City" (Lynch, 1984), cities (such as Boston, Jersey, and Los Angeles) were divided into regions of importance (based on data on crime, society, urban or non-urban sections, etc.), generating mental maps about the common characteristics of these cities, indicating that the elements of each city are distinguished among hundreds, thousands, or millions of other artifacts due to their unique shapes, sizes, colors, etc. From this set of studies, a trend began in the psychological aspect to study and evaluate the perception of inhabitants regarding the visual elements of the city. In the work carried out by Nasar (1998), which was strongly related to finding those regions/areas that were most pleasing to citizens, it was shown that in most evaluations, green areas, themed streets, open spaces, shopping centers, and airports predominated. In addition, areas that were rated as "not pleasant" for inhabitants were buildings with unattractive styles, the presence of graffiti, parks without an established form, and abandoned places. Additionally, other studies related to the disorder of the city (Skogan, 1992), focused on the presence of garbage on the streets, abandoned buildings, and cars parked in desolate corners, which contribute to the perception of lack of control, fear, and insecurity in the city.

For this reason, studies have been conducted to understand the behavior of crime and the feeling of insecurity related to the influence of crimes and delinquency in the streets, which have been increasing in crowded places (e.g., tourist destinations). These crimes have a long-term negative impact on how potential tourists perceive the level of security in these places (Mawby, 2014; Mohammed y Sookram, 2015; Glaeser et al., 2018). Additionally, over the years, information on crime rates and trends in various cities has been collected, such as the website Numbeo (Numbeo, 2019), which informs us about crime rates in all countries. Interestingly, it shows that South America has higher crime levels than Asia and Europe (e.g. Caracas, Venezuela is ranked first on the list). With this information on crime rates by city, various applications have been developed, such as crime maps (USA, 2012; Google-Motorolla, 2019), data statistics (EuroStat, 2016), and applications that predict criminal trends in areas (Stalidis et al., 2018).

All these previous studies about the impact of visual appearance and crime rate in cities have generated different approaches to identify which elements of the visual appearance of a certain street influence urban perception (Andersson et al., 2017). For example, quality of life, green areas, and safety, among others. In recent years, with the advancement of various techniques to analyze information (e.g., images) and the evolution of techniques such as deep learning, there has been evidence of not only reports but also the creation of datasets and the trend toward predicting urban per-

ception. An example case is the work "Which looks more safety?" carried out by the MIT-Media Lab, creating the dataset Place Pulse (MIT-Media-Lab, 2013). The data recorded in Place Pulse is about people's urban perception based on an online survey; in the survey, a volunteer must choose between two images of streets the safest one. Based on this dataset, Li et al. (2015b); MIT-Media-Lab (2015) analyzed green areas and their influence on urban perception. Additionally, techniques such as object detection (e.g., graffiti) and the addition of other datasets, such as crime rates, levels of violence, presence of trees, human development index, among others, were analyzed in various works such as Porzi et al. (2015); Tokuda et al. (2019); Arietta et al. (2014); Li et al. (2015b) that we will detail later.

Similarly, studies on the presence of objects and their correlation with urban safety perception have also been conducted, showing that it is possible to divide cities based on the most frequent types of objects (e.g., trees, garbage, buildings, fences, graffiti, etc.) and the respective perception of safety (Zhang et al., 2018; Min et al., 2019). As briefly presented in this section, there is great motivation for the study of urban perception based on the characteristics and visual appearance of street images. Such studies are based on statistics collected over a period of time, as well as models that use these statistical data to make predictions about influential areas, the presence of objects (e.g., graffiti, garbage), or relationships between places and crime statistics. We have identified that in this type of study, the **computational problem lies in how to identify, differentiate, and relate the characteristics of street images with the idea of urban perception**, due to the similarity between images, the small number of samples, etc. In our work, we will first focus on exploring and analyzing the data from *Place Pulse 2.0* (which is composed of street images). In this way, we will propose a model based on *Deep Convolutional Neural Network* (DCNN) that allows us to solve the difficulties mentioned in predicting urban perception effectively.

1.2 Objectives

In this section, we will present the objectives of this work in a concise manner. The main motivation is based on research on how to predict the perception of urban safety. Through the dataset *Place Pulse 2.0*, we propose a methodology that allows us to perform this task. Therefore, we present below the objectives set out in this work:

1.2.1 General Objective

The general objective of this work is to describe, explore, analyze, and present the evaluations of different models based on DCNN to make a prediction of the perception of citizen security (e.g., safe and not safe) using the *Place Pulse 2.0* dataset previously mentioned.

1.2.2 Specific Objectives

In particular, we can list the specific objectives briefly mentioned in the main objective:

- Propose a methodology that allows us to explore and analyze the *Place Pulse 2.0* dataset, which is composed of 1.22 million comparisons of 111,390 images from 56 different cities, containing six different comparison categories: *safety*, *lively*, *beautiful*, *wealthy*, *boring*, *depressing*. Additionally, to analyze and present the characteristics and behavior of the data. The aim is to identify possible limitations that may be present in the dataset (which will be discussed in detail later in this document), and the main category of study will be *safety*.
- Propose and present a *Convolutional Neural Network* (CNN)-based model for the classification task. This model will allow us to efficiently differentiate the urban perception of a street, taking into account the behavior and distribution of the previously analyzed data. For the evaluations, we will use different approaches such as transfer learning, fine-tuning, and generative adversarial networks.

1.3 Contributions

The present work presents two main contributions. The first contribution is the study and analysis of *Place Pulse 2.0*, whose data consists of images of 56 cities associated with an urban perception score (e.g., safety). The objective of this study is to explore and analyze all the characteristics and distribution of the data. The analysis will expose the criteria to be used to divide our data between the safe and unsafe categories, as well as study whether it is possible to divide into regions through perception at different "geographic generalization level" such as city, country, continent, and global.

The second contribution corresponds to a model based on DCNN, which will be evaluated using various techniques and approaches, such as Supervised Learning and Semi-Supervised Learning. This model will be able to differentiate, relate, and identify the characteristics of the images and make the prediction of urban perception. This will be further explained in Chapters 4 and 5, in which the description of the techniques and models will be presented, as well as the obtained results.

In Figure 1.1, the implicit methodology that encompasses all the work carried out and presented in this document is shown. This methodology allows us to carry out both previously described contributions, starting from the calculation and grouping of images into the two studied categories (safe and not safe) through the associated scores for each one. In (I), we observe this data set with each image and its associated score. From these scores, in (II), an exploratory analysis of the data is carried out with the intention of understanding the behavior of the studied data, showing as a result the disparity of data and distributions of the obtained associated scores. This result corresponds to the article published in Felipe Moreno-Vera (2021b). In (III), the

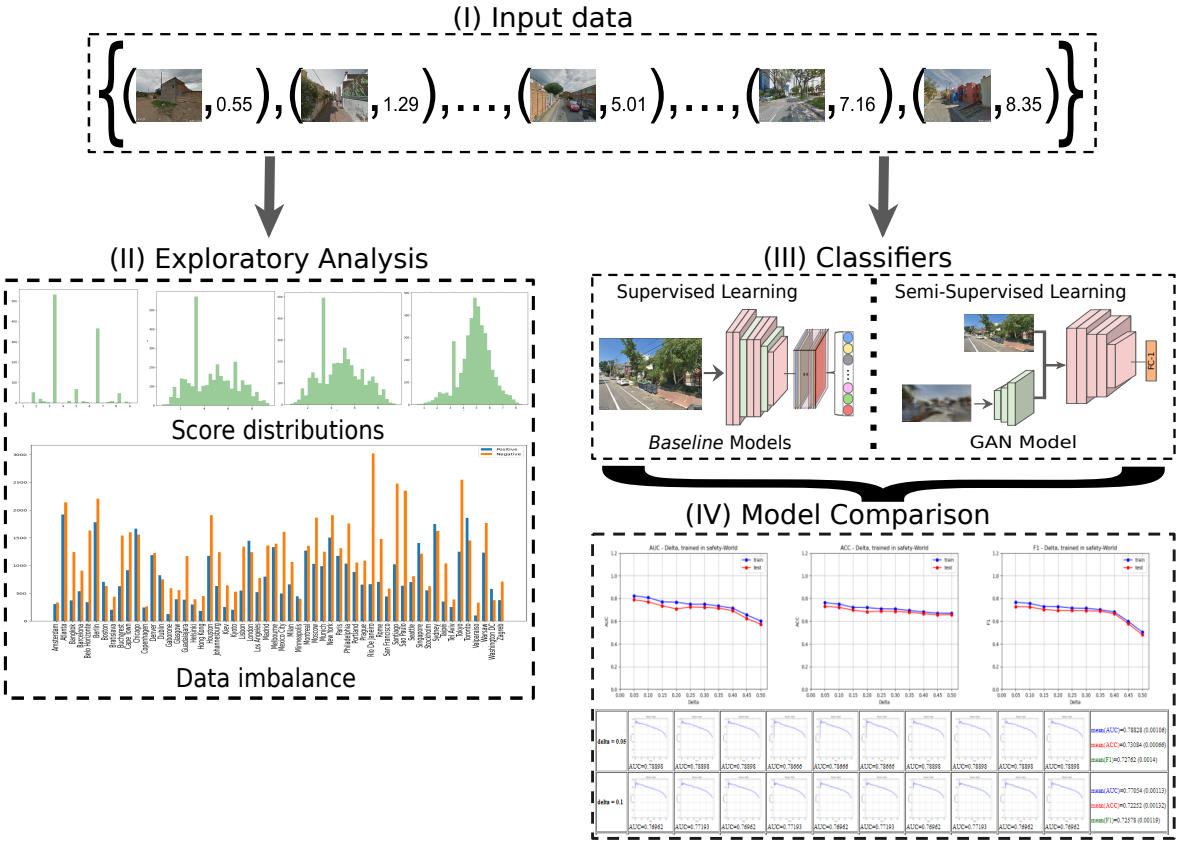


Figure 1.1: The Work Methodology: We present the general steps to achieve the specific objectives described. Starting from the dataset of images and their associated scores, which is located in (I); in (II), we show part of the exploratory analysis carried out on the data, highlighting relevant results such as the score distribution and data disparity; in (III) we train and track the models using supervised and semi-supervised approaches; finally in (IV) we report and compare the evaluation results obtained, representing them through graphs with respect to the evaluated metrics values. The purpose of this outline is to give the reader a general idea of this work, which will be detailed in the next chapters. Source: The author.

respective training and validation of the data are carried out in different approaches based on CNN models, to finally report, compare, and show the evaluations and metrics obtained in each model in (IV). Each of these steps will be explained and discussed in detail in the next chapters.

Finally, this work has three published papers: The first one was presented at the *IEEE/WIC/ACM International Conference on Web Intelligence (WI-IAT) '21*, which covers all the analysis of the limitations present in this dataset, which will be explained in detail in Chapter 3. The second publication was presented at the *IEEE Mexican International Conference on Artificial Intelligence (MICAI '21)*, which focuses on the analysis of the correlation between the presence of objects and the urban security perception of street images. The third publication was presented at the *International Conference on Intelligent Computing (ICIC '21)*, where some preliminary results of supervised model training and a comparison between two model explanation methods

(looking for the relevant regions for the prediction) were presented. You can find them by following the references:

- **Moreno-Vera, Felipe**, Bahram Lavi, and Jorge Poco. "Quantifying Urban Safety Perception on Street View Images". *In IEEE/WIC/ACM International Conference on Web Intelligence* (WI-IAT '21), December 14–17, 2021, Essedon, Australia. (Felipe Moreno-Vera, 2021b).
- **Moreno-Vera, Felipe**, Bahram Lavi, and Jorge Poco. "Urban Perception: Can We Understand Why a Street Is Safe?". *In Mexican International Conference on Artificial Intelligence* (MICAI '21), October 25-30, 2021, Mexico City, Mexico. (Felipe Moreno-Vera, 2021a).
- **Moreno-Vera, Felipe**. "Understanding Safety based on Urban Perception". *In International Conference on Intelligent Computing* (ICIC '21), August 12-15, 2021, Shenzhen, China. (Moreno-Vera, 2021).

1.4 Document Structure

In Chapter 2, we present related work on (a) urban perception analysis, (b) feature extraction and visual components, and (c) interpretation and visualization of extracted features. In Chapter 3, we present the exploratory analysis of the *Place Pulse 2.0* dataset. In Chapter 4, we describe the architectures of the models we will use for the experiments on the data. In Chapter 5, we present and describe the results of the training and evaluations performed based on our hypotheses. In Chapter 6, we present the discussions and limitations of this work. Finally, in Chapter 7, we present the conclusions obtained from our results.

Chapter 2

Related Works

The studies on urban perception have been increasing due to the availability of more geographic reference information on streets and cities (e.g. Google Street View) and the search for a way to determine the level of perception (e.g. safety) of streets. Related works can be grouped into three main categories: (a) analysis of urban perception, (b) extraction of visual features and components, and (c) interpretation and visualization of extracted features. As an introduction to the reader, we will first have a section on background concepts where we address the minimum knowledge required to understand the content of the document.

2.1 Background Concepts

In this section, we present some basic and necessary concepts to understand this document. Some of the concepts to be addressed are supervised learning techniques, unsupervised learning, and semi-supervised learning. We will also cover model interpretation and some image classification, object detection, and segmentation tasks.

2.1.1 Machine Learning Techniques

We will briefly explain the learning techniques that will be mentioned in this document, which are Supervised Learning, Unsupervised Learning, and Semi-Supervised Learning.

2.1.1.1 Supervised Learning

It is a machine learning method in which a data set is composed of "input data" and associated "labels", which can be called "training data" (Abu-Mostafa et al., 2012). The data set of "n" samples can be defined as $I_{supervised} = (x_1, y_1), (x_2, y_2), \ldots (x_n, y_n)$

where each $x_i \in \mathbb{R}^d$ is the i-th "feature vector" and its corresponding label (class) is y_i (Wikipedia, b). Some tasks performed with this technique are binary classification, image classification, sentiment analysis, and linear regressions, among others.

2.1.1.2 Unsupervised Learning

It's a machine learning method in which the training dataset does not contain any labels or associated information (Abu-Mostafa et al., 2012). For this type of learning, the dataset consists only of $I_{unsupervised} = x_1, x_2, \dots x_n$, where each $x_i \in \mathbb{R}^d$ is a "feature vector". Some tasks performed with this technique include clustering, dimensionality reduction, outlier detection, among others.

2.1.1.3 Semi-Supervised Learning

It's a machine learning method in which there is a dataset composed of "n" samples divided into two subsets as $I_{labeled} = (x_1, y_1), (x_2, y_2), \dots (x_p, y_p)$ and $I_{unlabeled} = x_1, x_2, \dots x_q$, such that $I_{semi-supervised} = I_{labeled} \cup I_{unlabeled}$; where each $x_i \in \mathbb{R}^d$, y_i are labels/classes, and $p + q = n$ is the total number of samples. The main objective of this type of learning is to learn relevant representations of the data (Goodfellow et al., 2016). Some tasks performed by this technique can be data augmentation, data generation, pseudo-labeling, as well as supervised and unsupervised learning tasks.

2.1.2 Machine Learning Tasks

We will briefly explain the tasks of machine learning: image classification, object detection, and object segmentation.

2.1.2.1 Image Classification

Image classification is a machine learning problem that defines a set of classes (objects to identify in images) and trains a model to recognize the objects using labels associated with each image (Google-Developers, 2020). Some of the most notable models are LeNet (Y. et al., 1990), AlexNet (Krizhevsky et al., 2012), ZFNet (Zeiler y Fergus, 2013a), GoogleNet (InceptionV1) (Szegedy et al., 2014), VGG-Net (Simonyan y Zisserman, 2014), ResNet (He et al., 2015), InceptionV3 (Szegedy et al., 2015), Xception (Chollet, 2017), among others.

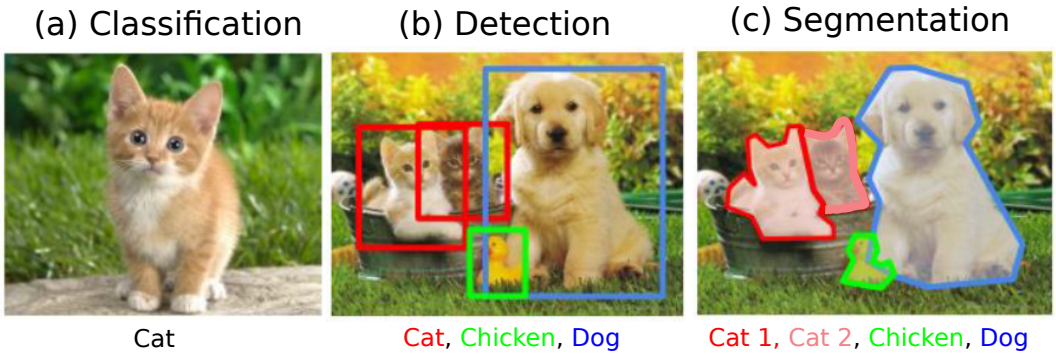


Figure 2.1: We present the different machine learning tasks defined: a) Classification: Identifies the features of a cat in the image. b) Object Detection: Identifies and locates different objects within the same image. c) Object Segmentation: Identifies, locates, and covers all the pixels where each object is found. Source: Stanford-CS231 Deep Learning course (CS231n, 2022).

2.1.2.2 Object detection

The task of object detection is a computer vision technique to locate instances of objects within images or videos (MatLab-Developers, 2020). Some well-known methods include R-CNN (Girshick et al., 2014), Fast R-CNN (Girshick, 2015), Faster R-CNN (Ren et al., 2017), Single Shot MultiBox Detector (SSD) (Liu et al., 2016), and You Only Look Once (YOLO) and its derivatives (up to the current version 7) (Redmon et al., 2016).

2.1.2.3 Semantic Segmentation and Instance Segmentation

The task of image segmentation is a pixel clustering technique that groups pixels belonging to the same object within an image, also known as "pixel-level classification." In other words, it involves dividing an image into multiple regions (pixel groups) called segments (Viso-AI, 2020). Some notable methods include DeconvNet (Noh et al., 2015), U-Net (Ronneberger et al., 2015), DeepMask (Pinheiro et al., 2015), Dilated Convolutions (Yu y Koltun, 2015), Pyramidal Scene Parsing Network (PSP-Net) (Zhao et al., 2017), Mask RNN (Hu et al., 2017), Mask R-CNN (He et al., 2017), DeepLab (and its derivatives) (Chen et al., 2017), among others.

2.1.3 Machine Learning Models

We will briefly explain linear and non-linear models, as well as their main components. Additionally, we will cover model interpretation.

2.1.3.1 Linear Models

Let a dataset of "n" samples be $I = (x_1, y_1), (x_2, y_2), \dots (x_n, y_n)$ where each $x_i \in \mathbb{R}^d$ is called an independent variable and each associated y_i are dependent variables. A linear model studies the linear relationship between dependent variables y and independent variables x and predicts values for new samples. These models are called linear regressions, which describe the dependent variable $y_i = f(x_i) = \sum_{j=1}^{d} \phi(x_{ij})\beta_j + \beta_0$, where $\beta = (\beta_0, \beta_1, \dots, \beta_d) \in \mathbb{R}^{(d+1)}$ are the slopes of each variable $\phi(x_{ij})$. Likewise, the linear relationship between each y_i and x_i is observed through the variables β and the functions $\phi()$ which can be linear or non-linear functions (Wikipedia, a).

One way to evaluate the efficiency of these models is by using the error obtained from the defined cost function of the form $Err = L(y_i, f(x_i)) + R(\beta_i)$, where $L(y_i, f(x_i))$ is the cost function, $R(\beta_i) = \frac{(1-\rho)}{2}||\beta_i||_2 + \rho||\beta_i||_1$, where $\rho \in [0, 1]$ is the regularization parameter. Furthermore, it is observed that: (a) If $\rho = 0$ it is **L2** regularization or also called **Ridge** (Tikhonov, 1943), (b) If $\rho = 1$ it is **L1** regularization or also called **LASSO** (Tibshirani, 1996) and (c) If $0 < \rho < 1$ it is **Elastic Net** regularization (Zou y Hastie, 2005).

Likewise, the cost function $L(y, f(x))$ changes depending on the type of task to be performed. In the classification task we have the following cost functions:

- The **Logistic Regression** method has the cost function defined by $L(y, f(x)) = \sum_{i=1}^{n}[y_i \log(f(x_i)) + (1 - y_i)\log(1 - f(x_i))]$, this function is also called **Binary crossentropy**.
- The **Support Vector Classification** method has the cost function defined by $L(y, f(x)) = \sum_{i=1}^{n} \max(0, 1 - y_i f(x_i))$, this function is also called **hinge**. It can also be defined as $L(y, f(x)) = \sum_{i=1}^{n} \max(0, 1 - y_i f(x_i))^2$ (**huber** or **squared hinge**).

In the regression task, we have the following cost functions:

- The **Linear Regression** method has the cost function defined by $L(y, f(x)) = \sum_{i=1}^{n} ||y_i - f(x_i)||2^2$, this function is also called **Least Squares**.
- The **Support Vector Regression** method has the cost function defined by $L(y, f(x)) = \sum i = 1^n \max(0, |y_i - f(x_i)|)$, this function is also called **epsilon-sensitive**.

2.1.3.2 Non-Linear Models

A non-linear model studies the non-linear relationship between dependent variables y_i and independent variables x_i. These models are called non-linear regressions, which

describe the dependent variable as $y_i = f(x_i, \theta_i) + \epsilon$, where θ are other unknown parameters (Wikipedia, 2020). Additionally, the independent variables x_i can be of any dimension, for example in images we would have the dimension $r \times c \times 3$, where $r \times c$ is the size of the image and 3 is the color scale.

Similarly to what was explained for linear models, the way to evaluate non-linear models is through the error calculated from the cost function defined by $E(\theta_i) = L(y_i, f(x_i, \theta_i))$. In the case where our x_i are images, the approximation function can be a deep convolutional network of the form $f(x_i, \theta_i) = D(A_1(P_1(C_1(...A_n(P_r(C_s(...)))))))$, where $C_j()$ are convolution operations (also called filters), $P_j()$ are pooling functions, $D_j()$ is a linear combination (called dense layers), and $A_j()$ are activation functions. Below we describe each one:

- **Convolution**: $h_{ij}(f, g) = (fg)(i, j) = \sum_{m=-\infty}^{\infty} \sum_{n=-\infty}^{\infty} f(m, n) g(i + m, j + n)$, where $f(m, n)$ corresponds to the pixel at position (m, n) of the input image and $g(i + m, j + n)$ are the corresponding pixels of the convolution matrix. This function is also called a "filter".

- **Pooling**: is a technique for resizing a matrix $I \in \mathbb{R}^{W \times H}$ through a *Pooling* sub-matrix $P \in \mathbb{R}^{w_p \times h_p}$. The variable *stride* is defined as the distance between each pixel. Let $I_{i,j}$ be the position of pixel (i, j) of the image I and also the pooling matrix. To simplify the notations, we define $k = 0, \ldots, W$ and $l = 0, \ldots, H$ as the positions of pixels in rows and columns. The following pooling operations are defined:

  - **Max Pooling**: Returns a matrix of dimension $(\lfloor \frac{W - w_p}{stride} \rfloor + 1, \lfloor \frac{H - h_p}{stride} \rfloor + 1)$. Each value of the matrix will be of the form: $\max_{k \le i \le k + w_p, l \le j \le l + h_p} I_{i,j}$.

  - **Global Max Pooling**: Returns a value associated with a matrix of the form: $\max_{i,j} I_{i,j}$.

  - **Average Pooling**: Returns a matrix of dimension $(\lfloor \frac{W - w_p}{stride} \rfloor + 1, \lfloor \frac{H - h_p}{stride} \rfloor + 1)$. Each value of the matrix will be of the form: $\frac{1}{w_p * h_p} \sum_{i=k}^{k + w_p} \sum_{j=l}^{l + h_p} I_{i,j}$.

  - **Global Average Pooling**: Returns a value associated with a matrix of the form: $\frac{1}{WH} \sum_{i=0}^{W} \sum_{j=0}^{H} I_{i,j}$ (Lin et al., 2013).

- **Activation Functions**: are functions used to confine a set of values to a specific domain, the most well-known are:

  - Hyperbolic Tangent (**Tanh**): $f(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$.
  - **Sigmoid**: $f(x) = \frac{1}{1 + e^{-x}}$.
  - **ReLU**: $f(x) = \max(0, x)$.
  - Linear Function (**Linear**): $f(x) = x$.

(a) (b) (c)

Figure 2.2: Prediction: Dog; explanations presented by the methods (a) CAM, (b) GBP, and (c) guided-CAM. Source: grad-CAM (Selvaraju et al., 2017).

2.1.4 Explanation Methods

Explanation methods allow us to understand the behavior and decision-making process of a model. They can be divided into two main classes: *white-box* methods, which are easier to analyze, such as linear models, and *black-box* methods, which are more challenging to analyze due to their complexity and large number of parameters, such as deep neural networks (Molnar, 2022). Some explanation methods of this type include **Convolution Visualization** (Zeiler y Fergus, 2013b), *smooth* (Ancona et al., 2017), *saliency maps* (Simonyan et al., 2013), and *class activation maps* (CAM) (Zhou et al., 2014), among others.

To understand the calculation, let's define $x \in \mathbb{R}^d$ as a feature vector of an image, a model $S : \mathbb{R}^d \to \mathbb{R}^C$ where C is the number of classes to evaluate, and an explanation method $E : \mathbb{R}^d \to \mathbb{R}^d$ that defines an explanation map. Then, a gradient-based explanation for a variable x is of the form $E_{grad}(x) = \frac{\partial S}{\partial x}$, where the gradient quantifies how much each dimension of x changes the prediction $S(x)$ in a small neighborhood around x (Adebayo et al., 2018). Some gradient-based methods include:

- **Gradient Input**: The gradient calculation is of the form: $x \odot \frac{\partial S}{\partial x}$, which reduces visual diffusion and gradient saturation (Shrikumar et al., 2016).
- **Integrated Gradients**: The gradient calculation is of the form: $(x-\bar{x})x \int_0^1 \frac{\partial S(x+\alpha(x-\bar{x}))}{\partial x}$, also reducing gradient saturation through scale integral, $\bar{x}$ represents the absence of features in x (Sundararajan et al., 2017).
- **CAM (Class Activation Map)**: The gradient calculation is of the form: $M_{cam} = \sum_k w_k^c F^k$, where M_c is a *class activation map* of class c, k represents a convolutional layer, and F^k is the result of applying GAP ($F^k = \frac{1}{Z}\sum_i \sum_j A_{ij}^k$) to each convolution, where A^k is the filter corresponding to the k-th convolution, (i, j) represent the pixels of matrix A^k, and Z is the product of the dimensions of A^k (Zhou et al., 2016a).

- **GradCAM**: Taking into account the result obtained in CAM, $\alpha_k^c = \frac{1}{Z}\sum_i \sum_j \frac{\partial M_c}{\partial A_{ij}^k}$ is defined as the GAP of the gradients of CAM called *partial linearization*, finally $M_{gradcam} = ReLU(\sum_k \alpha_k^c A^k)$ (Selvaraju et al., 2017).
- **GBP (Guided Backpropagation)**: The calculation is as follows: $B_i^t = (f_i^t > 0) \times (B_i^{t+1} > 0) \times B_i^{t+1}$, where $B_i^{t+1} = \frac{\partial f^{out}}{\partial f_i^{t+1}}$, $f_i^{t+1} = relu(f_i^t)$ is the activation of the current layer of the network, and B_i^t is called the GBP of layer t and i is a sample (Springenberg et al., 2014).
- **Guided GradCAM**: It is the element-wise product between GradCAM and GBP (Selvaraju et al., 2017).
- **SmoothGrad (*SmoothGrad* (SG))**: Reduces noise and visual diffusion in saliency maps with a weighted sum of explanations of noisy copies of x. For an explanation E, we have: $E_{sg} = \frac{1}{N}\sum_{i=1}^{N} E(x + g_i)$ where $g_i \sim \mathcal{N}(0, \sigma^2)$ is the noise (Smilkov et al., 2017).

In Figure 2.2, an example of some methods such as CAM, Grad-CAM, and GBP is shown. These methods are widely used to interpret models that use images as training data.

In this section, we have presented concepts such as learning techniques, linear and nonlinear models, as well as the operations involved in each. These terms and definitions will be used later in this document, especially from the next sections where we address related works. As a reminder, the following sections are divided into (a) urban perception analysis and (b) feature extraction and visual components; and (c) interpretation and visualization of extracted features.

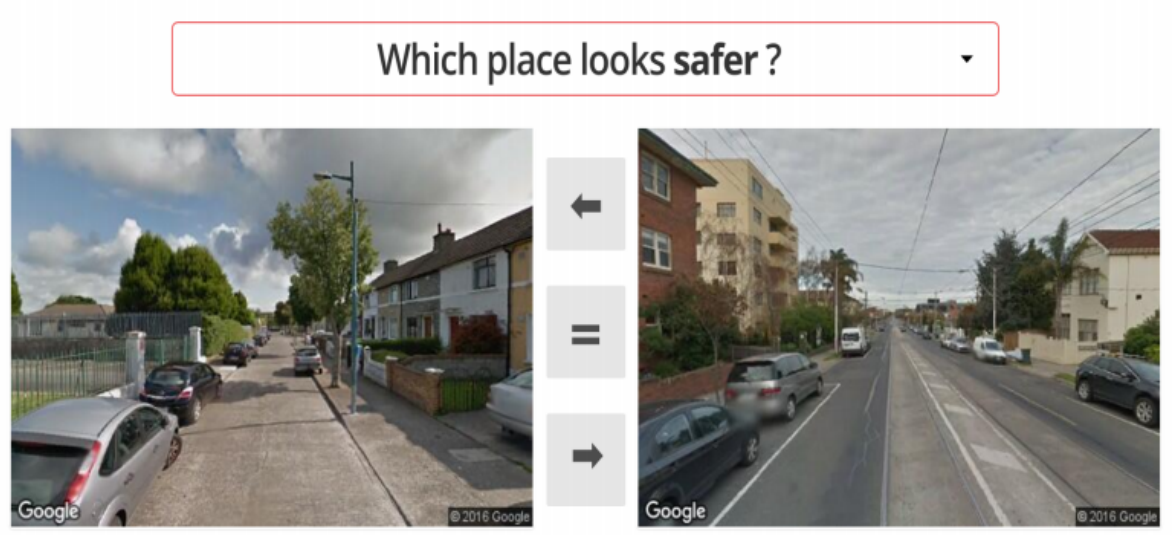


Figure 2.3: *Place Pulse* website, where information about street perception is collected by choosing between two street images. Source: *Place Pulse* (Salesses et al., 2013).

2.2 Urban Perception Analysis

In this section, we will discuss related works on urban perception analysis using various methods to relate the visual appearance of streets and other non-visual data, such as crime rates. In 2011, the MIT Media Lab initiated the project called Place Pulse (Salesses, 2012), which collected information about images from various volunteers with the aim of answering the following question: "Which Place Looks Safer/Unique/Wealthy?". Each volunteer had to choose between two random images downloaded from the cities of Boston, New York, Linz, and Salzburg. The work conducted a study on the perception differences of each of the evaluated cities based on their visual aspects, creating a quantitative measure of the contrasts within a city. As important results, we have the creation of the Place Pulse dataset version 1.0 and the finding that the visual aspects between Boston and New York were more pronounced than between Linz and Salzburg. Furthermore, when compared with crime data within these cities, it was found that places with similar visual appearance had similar crime rates.

In 2014, the Place Pulse 1.0 dataset spurred an increase in the number of studies and analyses on urban perception, such as learning specific characteristics to predict the safety level of a street. One such study (Ordonez y Berg, 2014) used comparisons regarding the categories of safety, uniqueness, and wealth. In addition to the data provided by Place Pulse 1.0, images were collected from New York (8863), Boston (9596), and 2 additional cities: Chicago (12,502) and Baltimore (11,772). Training was performed with the original cities of New York and Boston from Place Pulse 1.0. This work presented two models, a *Support Vector Machine* (SVM) (Boser et al., 1992) classifier and a *Support Vector Regressor* (SVR) (Smola y Schölkopf, 2004) model for predicting scores (numbers between 0 and 10) and perception, respectively. Both models were trained with l_2 regularization and used feature extractors such as GIST (Oliva y Torralba, 2001), SIFT + Fisher Vectors (Perronnin et al., 2010), *Deep Convolutional*

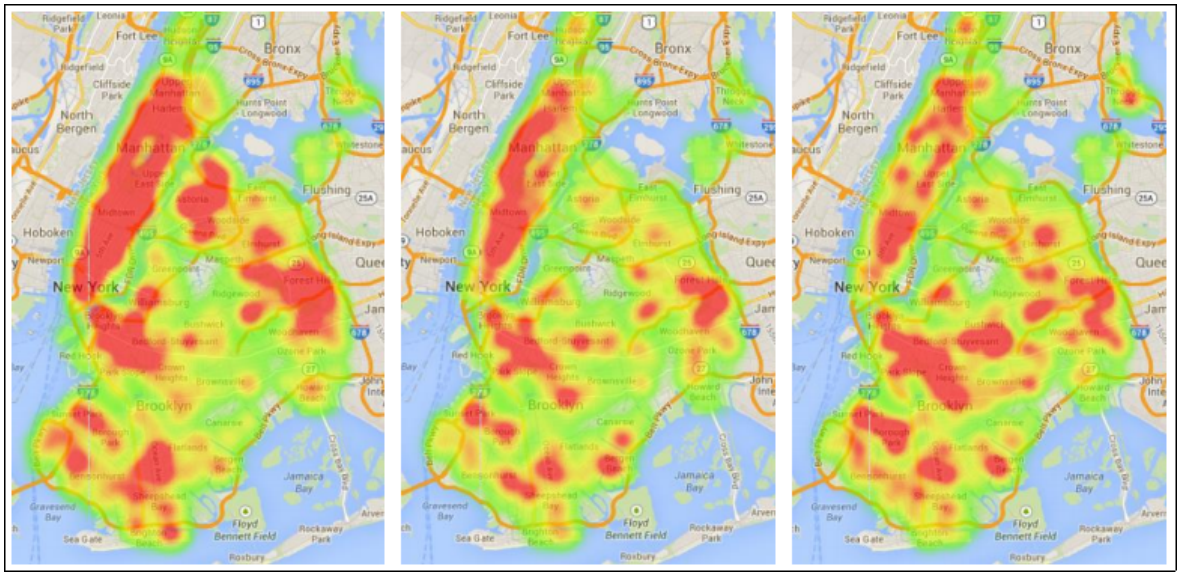


Figure 2.4: Results of regression evaluation on Place Pulse 1.0: (left) scores for the city of New York; (middle) regression predictions in New York using a model trained on New York; and (right) regression predictions in New York using a model trained on Boston. Source: (Ordonez y Berg, 2014).

Activation Feature (DeCAF) (Donahue et al., 2014).

For training, they used 5-fold cross-validation trained on each city and performed comparisons by evaluating on others. For labeling, they assigned scores below 5.0 as -1, otherwise 1. The main results showed that feature extraction with a deep network like DeCAF outperformed other methods such as GIST or SIFT + Fisher Vectors. Another finding was that the regressor had better accuracy for images with scores between 4-7. Additionally, they analyzed the collective perception prediction using the *K-Nearest Neighbors* (KNN) method (see Figure 2.4), showing that nearby regions have similar predictions. This work marks the beginning of using Place Pulse 1.0 with Deep Learning.

Later, Li et al. (2015a), using data from the cities of Boston and New York from Place Pulse 1.0, propose an analysis and exploration of aesthetics, environment, and psychological benefits in urban residences, prioritizing how green areas can help increase the perception of safety on streets in places such as residences, industrial areas, public places, institutions, etc. To process the images and filter green areas, they normalized the values of the *Red-Green-Blue* (RGB) channels and calculated the parameter Green Index defined as $G_I = 2G - R - B$. Then, through the Otsu (Otsu, 1975) filter and other pixel operations to remove brightness and contrast (shadows). Also, using the respective latitude and longitude of each image, using MassGIS Data-Land Use 2005 (Massachusets-Office-Goverment, 2005), they downloaded other *Field of View* (FOV), thus increasing the specially selected images from places such as: residential areas, public places (hospitals, universities, schools, parking lots, museums, prisons, etc.), industrial areas, cemeteries, open spaces, and recreation areas. Finally, with a linear regression, they related the presence of green areas and perception scores,

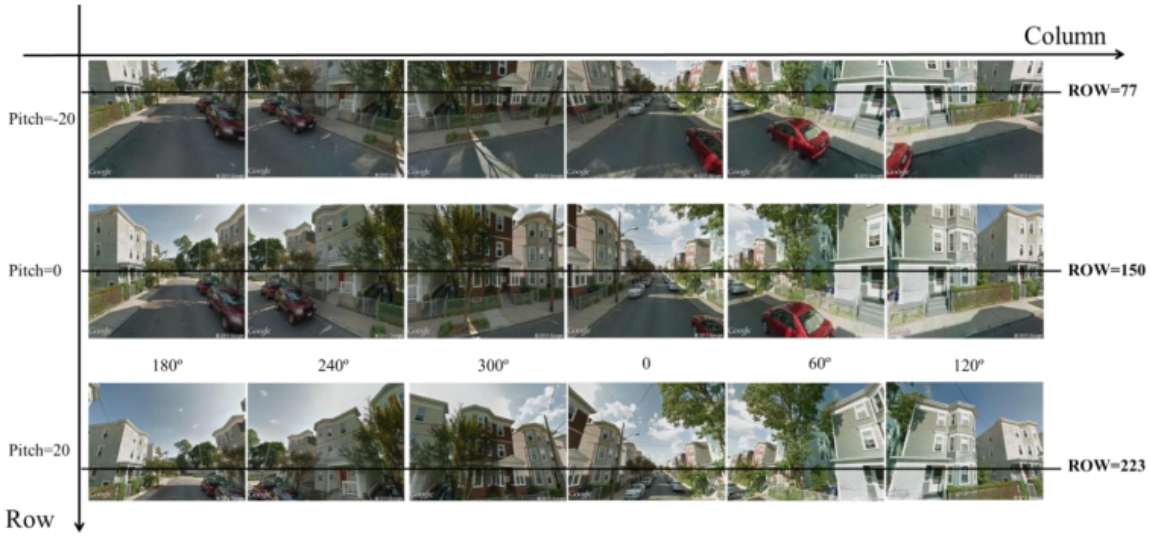


Figure 2.5: Different directions and angles of a street in Boston. The rows represent variations in height, and the columns represent the viewpoint. Source: (Li et al., 2015a).

thus demonstrating the importance and positive influence of green areas in places with high scores. As conclusions, it was presented that green areas impact positively on the perception of safety in most cases, however, they can be perceived as unsafe because they obstruct the view of a place contrasting with the theories presented in Fisher (1992); Nasar et al. (1993).

Another study strongly based on Place Pulse 1.0 was the development of the Wmodi platform (Acosta y Camargo, 2018b). This platform, similar to Place Pulse 1.0, collects information through surveys (see Figure 2.6 (a)), with the only difference that Wmodi uses images from the city of Bogotá, Colombia. As a pre-selection of images, they used the SIFT extractor (Lowe, 2004) to determine if there were minimal features or if they were only walls or black backgrounds, resulting in 5505 images. Likewise, the website obtained around 17,703 comparisons, where each image was compared on average 6 times. Additionally, they collected 5657 ties, 5946 safe, and 6100 unsafe perceptions. For the processing of scores, they used the TrueSkill algorithm (Herbrich et al., 2007), which enabled online score updating, generating a map of security perception scores (see Figure 2.6 (b)). They used the VGG19 network (Simonyan y Zisserman, 2014) and the GIST and HOG methods as feature extractors to then be trained in an SVR. The main contributions are: (i) the perception of high insecurity is related to places with few green areas, places with high traffic density, main avenues, roads under bridges, dirt roads; (ii) the Wmodi platform.

Similarly, we have studies focused on understanding and exploring the correlation between urban perception and crime statistics such as StreetNet (Fu et al., 2018). For this study, a dataset was constructed using indices of theft, aggravated robbery, petty theft, armed robbery, unauthorized entries into homes, etc. from the cities of New York and Washington DC. To rank the severity of crimes, the Preference Learning

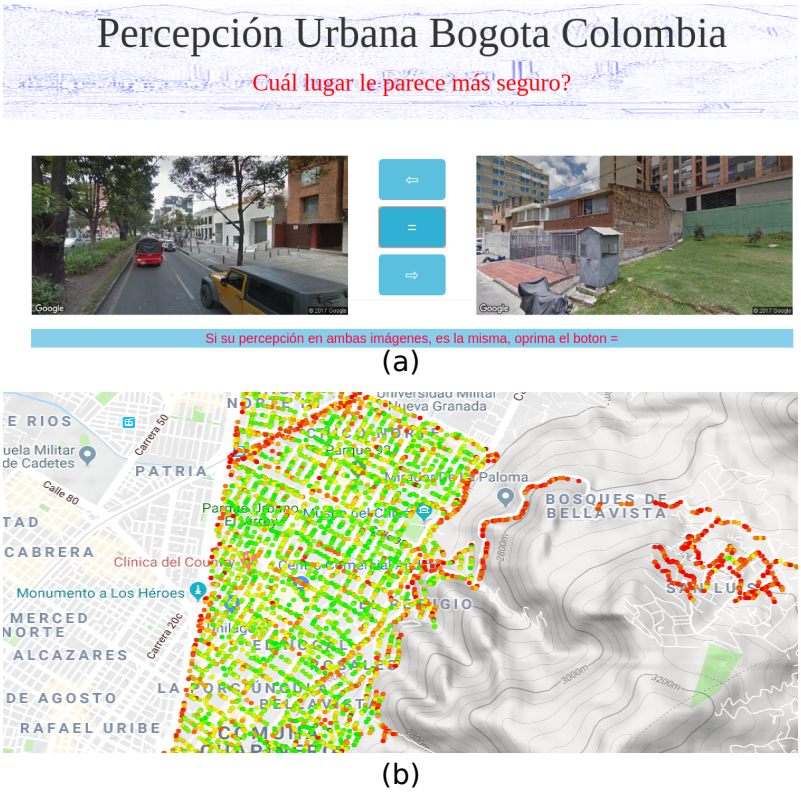


Figure 2.6: (a) Wmodi website for collecting information about street perception. (b) Map of safety perception scores for the Chapinero district (Bogotá, Colombia). Source: (Acosta y Camargo, 2018a).

method (Har-Peled et al., 2003) was used in each location, in addition to geographical references of the surrounding streets using the CycloMedia GlobalSpotter API (CycloMedia, 1980), where the central point is called the "sample point". In Figure 2.7 (b), geographical references can be observed from each sample point. From these sample points, the *Direction based Street View Retrieval* (DSVR) algorithm is used to group crimes committed in a certain established radial zone. In Figure 2.7 (c), it can be observed how the points where crimes occurred are grouped around a reference point. Notable results include: (i) creation of datasets for Washington DC and New York City (NYC) called DC-1k, DC-2k, NYC-1k, and NYC-2k generated using a radius of 1000 and 2000 feet respectively; (ii) the StreetNet model, which allows predicting what type of possible crime could occur based on crime data that occurred around and the characteristics of the location (see Figure 2.7 (a)).

In this section, related works on the study and analysis of urban perception were presented, some of which are based on the Place Pulse 1.0 dataset. Most of these works aim to relate some feature of the city to perception scores. Additionally, other datasets related to crime rates and visual aspects present in certain cities were utilized. In this paper, we focus on the study of the Place Pulse 2.0 dataset, which we will describe later, presenting an analysis of urban perception using the proposed methodology that we will describe in the following chapter.

Figure 2.7: Results: (a) Prediction of potential crimes based on a particular street view. (b) Geographical reference points taken topologically from a committed crime. (c) Geographical reference points obtained after applying the DSVR algorithm. Source: (Fu et al., 2018).

2.3 Feature Extraction and Visual Components

In this section, we present works related to urban perception but more focused on feature extraction. We also mention that we can divide them into two main categories: those of low/mid-level, which would be conventional methods, and high-level ones, which use deep networks for extraction. Some of the low/mid-level methods are GIST (Oliva y Torralba, 2001), SIFT + Fisher Vectors (Perronnin et al., 2010), *Histogram of Oriented Gradients* (HOG)+*Color descriptor* Dalal y Triggs (2005), *Geometric Probability Map* (Hoiem et al., 2007) y *Color Histograms* (Novak et al., 1992; Chakravarti y Meng, 2009) y métodos de alto nivel son *AlexNet* (Krizhevsky et al., 2012), *VGGNet* (Simonyan y Zisserman, 2014), *ResNet* (He et al., 2015), *PlacesNet* (Zhou et al., 2014).

2.3.1 Low-level feature extraction

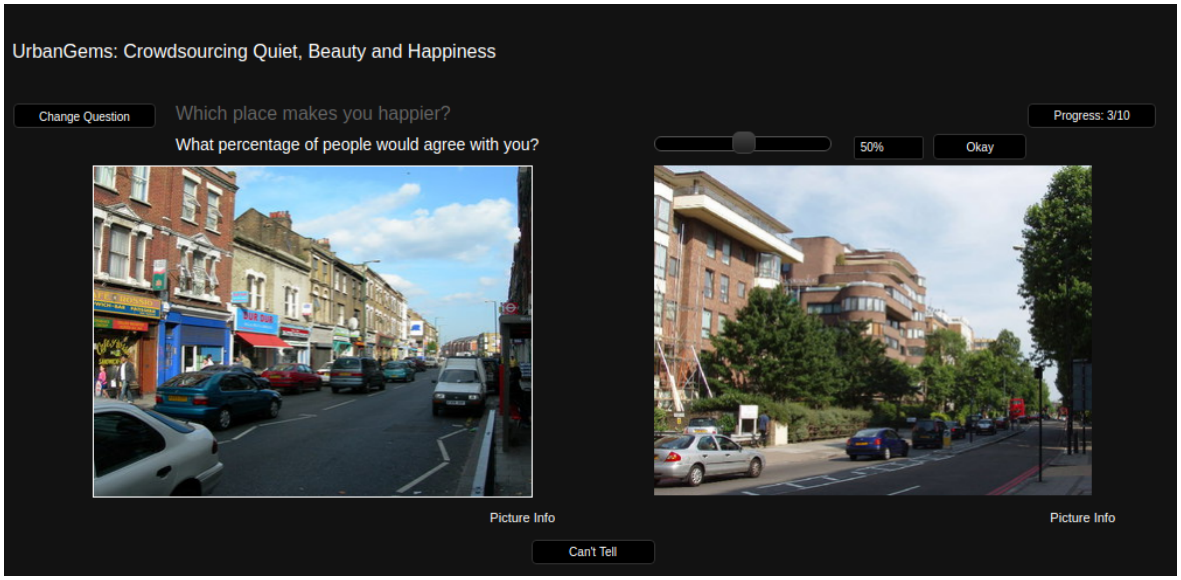
One of the most noteworthy works is "What makes Paris look like Paris?" (Doersch et al., 2012). In this study, the aim is to understand and identify the differences between various cityscapes in Europe. As an experiment, a questionnaire was conducted with 11 participants using 100 randomly selected images downloaded from *Google Street View* (GSV), with 50 % from Paris and the rest from different cities. The questionnaire involved determining whether a particular street belonged to Paris or not (ignoring

Figure 2.8: (a) Paris, France. (b) Prague, Czech Republic. (c) London, England. Visual correspondence between each element present in each city. Source: (Doersch et al., 2012).

any text present in the image). On average, they correctly identified around 79 % of the images. However, when text present in the images was taken into account, the average accuracy increased to 90 %. This demonstrates that people are sensitive to information within an image (e.g., signs, posters), aiding in quicker identification of the city. For the final experiment, 10,000 images per city were collected from 12 cities: Paris, London, Prague, Barcelona, Milan, New York, Boston, Philadelphia, San Francisco, Sao Paulo, Mexico City, and Tokyo. For the study, they divided the information into two categories: (i) images of Paris, and (ii) images of the other cities. They also assumed that visual patterns such as trees, cars, sky, etc., would exist in different cities and thus were not taken into account.

Using HOG+Color descriptor, features were extracted, and then grouped using Locality-Sensitive Hashing (Gong et al., 2012). For the creation of the groups, 25,000 images from different cities were randomly selected and subjected to KNN, resulting in 20 groups, of which only those with the highest proportion of nearest neighbors from the Paris set were kept. From the total 25,000, it was reduced to 1000 images as centers. For training the features, SVM with 3-fold cross-validation was used. Notable results include: (i) it was evident that in many cities in Europe, very similar visual appearances were observed. In Figure 2.8, the visual attributes corresponding to Paris, Prague, and London are observed, showing slight but marked differences between the elements; (ii) a robust method to differentiate the characteristics of all the studied cities, despite being quite similar.

Another work on the exploration of visual components was "What Makes London Look Beautiful, Quiet, and Happy?" (Quercia et al., 2014); aiming to gather information about the perception of cities, and also to analyze a factor of collective perception with the question "What percentage of people would agree with you?" and colors (e.g. of the streets) associated with that perception. In Figure 2.9 (a), the Urbangems web-



(a)

(b)

Figure 2.9: (a) Urbangems website; (b) The results of the visual words associated with the beauty category, where the red points represent an image. Source: (a) (UrbanGems, 2014), (b) (Quercia et al., 2014)

site (UrbanGems, 2014) can be seen where people had to choose between two images out of a total of 700,000 and give a percentage of people who would agree in the same way. Through the responses of 3301 users where each user performs a round composed of 10 comparisons, obtaining on average a preference about beauty (171), quiet (12), and happy (16). The images were collected through GSV from places near metro stations within a radius of 300 meters.

Once the information equivalent to 17,261 comparisons was obtained, we proceeded to analyze which colors have the greatest correlation with the images of *beauty* (beauty), *quiet* (silent), and *happiness* (happiness) using the RGB channels of the image, as well as the textures with *Global Edge Histogram* (GEH) Park et al. (2000) with the *region-based MPEG-7 Edge Histogram descriptor* Manjunath et al. (2001) tech-

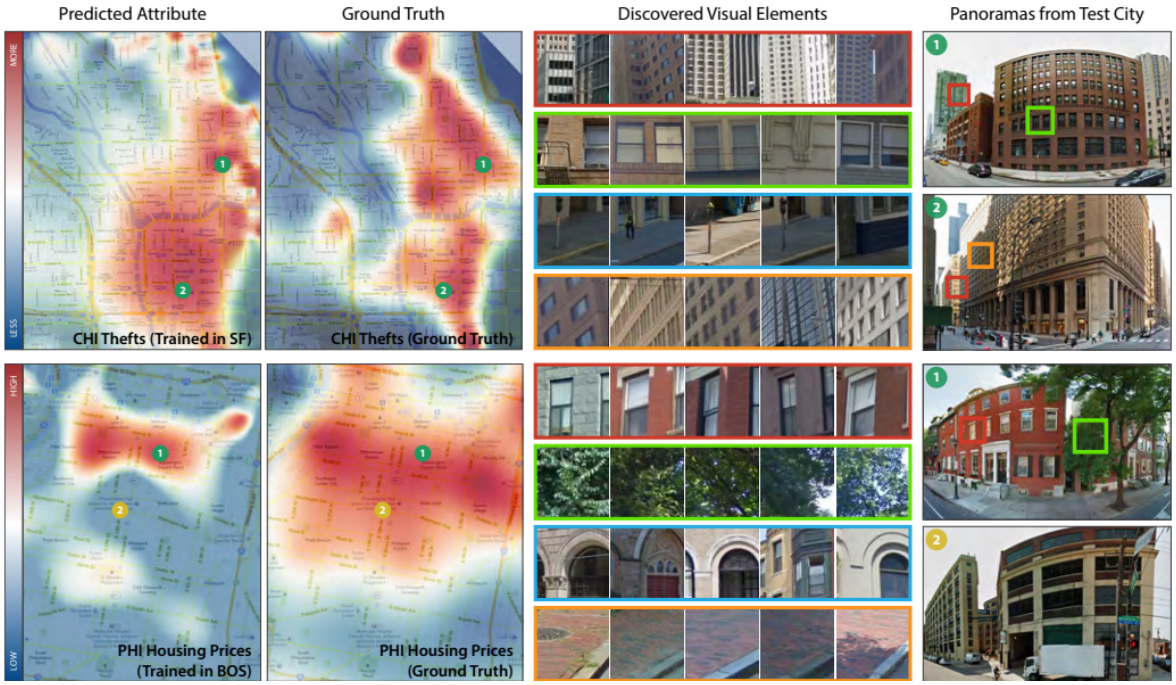


Figure 2.10: Results of attribute prediction. The first row shows the results of the model trained on San Francisco and tested on Philadelphia, the second row shows a case of error when the house price prediction model trained on Boston is tested on Philadelphia. The prediction error is due to some visual elements present in Boston that influence the "high prices" category not being present in Philadelphia. Source: (Arietta et al., 2014).

nique. Taking into account the points where users clicked within the image (called points of interest), they divided the images into 4 contours: horizontal, vertical, diagonal, and non-directional. To analyze the points of interest in the images, *Speeded Up Robust Features* (SURF) Bay et al. (2006) was used, grouping them with the *K-means* algorithm into 500 groups of *visual word* Jurie y Triggs (2005). In Figure 2.9 (b) you can see the *visual words* or points of interest of an image, these points can describe an image. As results they presented: (i) the selected points associated with the beauty category are Victorian houses, public gardens, residential and the least associated are government buildings, bridges, and roads. (ii) the points associated with the silent category are trees, hedges, forests, and residential windows, on the contrary, the least associated are construction sites and buses; (iii) the points associated with happiness are trees, buses, and people, on the contrary, the least associated are bridges, streets, and wire fences.

Another work focused on visual appearance is *City Forensics* (Arietta et al., 2014), which proposes a method to predict, identify, and corroborate a correlation between the visual appearance of a city and its non-visual attributes. The data for "non-visual attributes" include violent crime rates (CrimeMapping, 2012), robbery rates (CrimeReports, 2013), house prices, population density, presence of trees (UrbanForest, 2014), presence of graffiti (obtained from reports), and perception of danger. To obtain more data, panoramic images with a 360-degree FOV and a 20-degree tilt angle

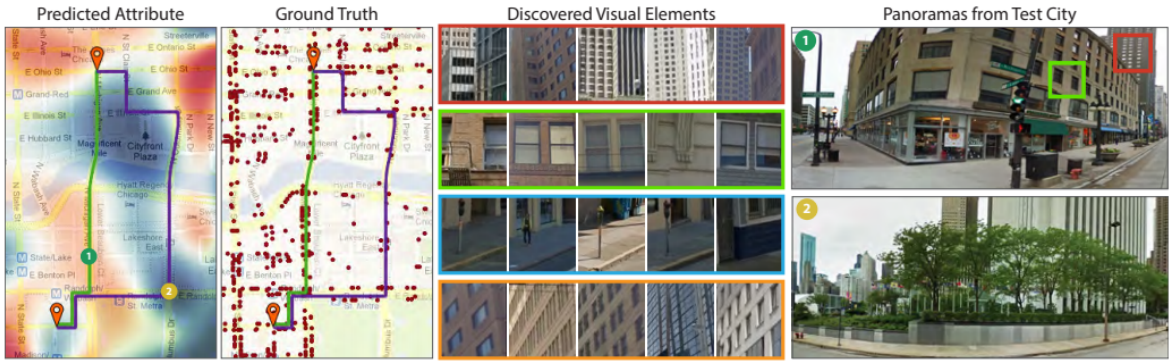


Figure 2.11: Another result is the prediction of safe routes (purple), avoiding the majority of robbery events (red circles) in Chicago, differentiating from the direct route (green). To calculate this route, the robbery rate in Chicago was predicted using the model trained in San Francisco, which contains visual elements related to traffic where robberies are common. In contrast, the predicted route contains green areas and trees, which have a low robbery rate. Source: (Arietta et al., 2014).

were downloaded using GSV from the cities of San Francisco, Chicago, and Boston. Between 30,000 and 170,000 panoramas were obtained per city, of which 10,000 were used for training. Visual attributes were identified using HOG+*Color descriptor*, and labels were annotated for each non-visual set (e.g., for house prices, values above the mean were annotated as positive and below as negative). Visual attributes were interpolated using the latitudes and longitudes of the images through the *Radial Basis Function* (RBF) method (Broomhead y Lowe, 1988), which was also the model used to train 10,000 samples with a ratio of 2000 positive and 8000 negative samples. Finally, an SVM was trained, refined with 3 iterations of the *hard negative mining technique* (Felzenszwalb et al., 2008) to relate each panorama to a non-visual attribute. As a notable result, there is a correlation between the visual appearances of each city and their respective crime rates, robbery rates, house prices, population density, presence of trees, presence of graffiti, and perception of danger. Additionally, three applications were identified: safe routes (see Figure 2.11 first column), city limits or divisions (see Figure 2.10 third column), and the ability of some visual components to describe a city (e.g., graffiti, bricks, windows with styles, and light poles describe Chicago).

In 2014, using data from the cities of New York and Boston from the *Place Pulse 1.0* dataset, *StreetScore* (Naik et al., 2014) presented a study comparing which feature extractor would be suitable for the images in that dataset. The compared extractors were GIST, *Geometric Probability Map*, *Texton Histograms* (Martin et al., 2001), *Color Histograms* (Novak et al., 1992; Chakravarti y Meng, 2009), *Geometric Color Histograms* (Rao et al., 1999), HOG (Dalal y Triggs, 2005), Dense SIFT (Lazebnik et al., 2006), LBP (Ojala et al., 2002), *Sparse SIFT histograms* (Sivic y Zisserman, 2004), and SSIM (Matas et al., 2004), which were trained in an SVR. The *TrueSkill* algorithm was used to process the scores, highlighting that on average each image was compared only 6 times, which was far from the optimal convergence between 12 and 36 comparisons. For the evaluations, approximately 200 images were downloaded for every 1.6 km^2, thus achieving the broadest coverage of a city, resulting in approximately

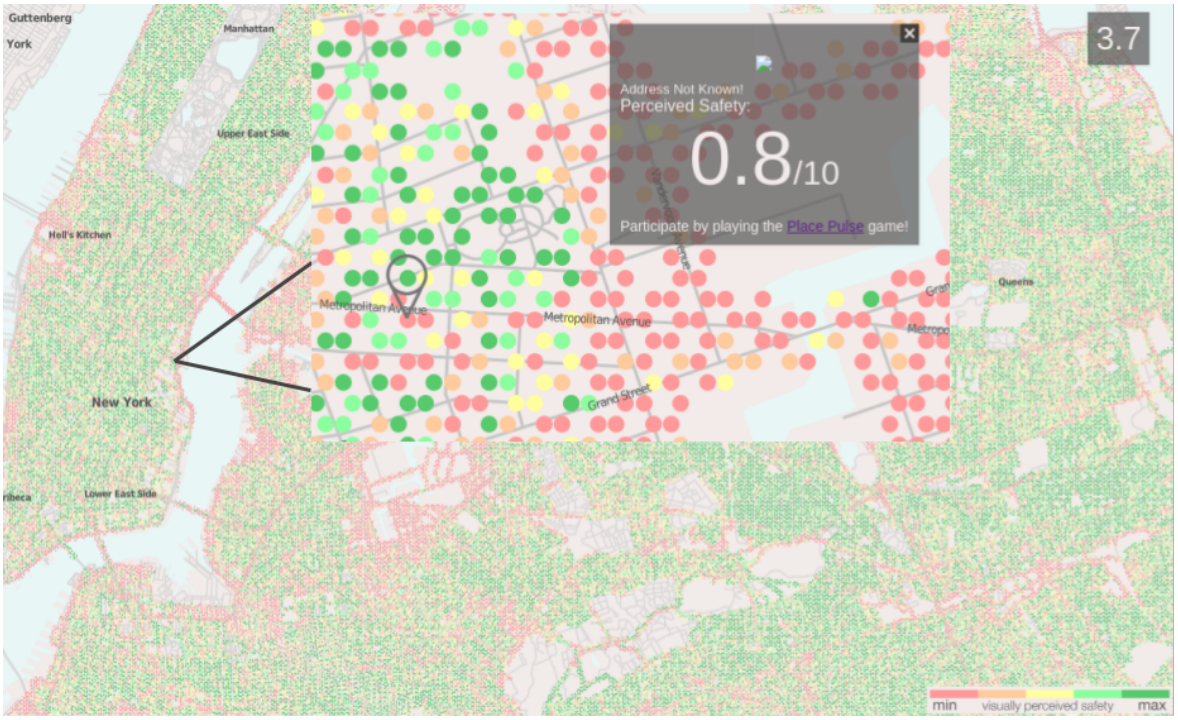
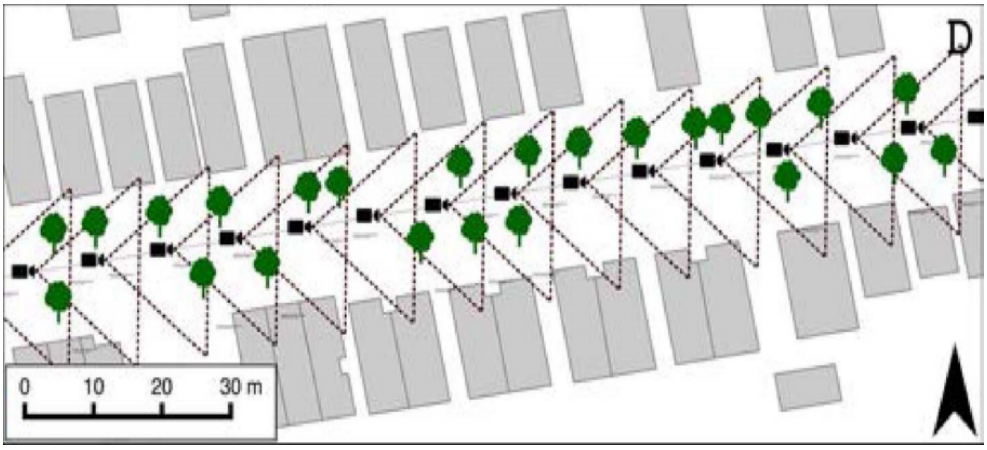


Figure 2.12: Results of perception score predictions in the city of New York: generation of a *HPM* where scores range from 0 to 10, with a red dot indicating the perception of an unsafe street and a green dot indicating the perception of a safe street. Source: (MIT-Media-Lab, 2014).

1 million images from 27 different cities in the USA. The results showed that *Color Histograms*, GIST, and *Geometric Color Histograms* performed the best. Based on this result, *StreetScore* was defined, which is a concatenation of the features extracted by these three methods. In Figure 2.12, the result of score prediction using *StreetScore* trained in Boston and New York, and evaluated in New York, can be observed.

Continuing the study presented in Li et al. (2015a) (described in the previous section), Li et al. (2015b) extended the analysis of green areas in panoramic images, aiming to identify the presence of green areas on streets, as well as their influence on the perception of safety. They modified the FOV by changing the angles to 0, 60, 120, 180, 240, and 300 degrees. Using images from the cities of East Village, Manhattan District, and New York, they located 300 randomly generated locations with ArcGIS 10.2 (ArcGis, 1999), with a separation of 300 meters between each location, resulting in approximately 28,448 m^2 where every 100 meters provide a global view of a location (see Figure 2.13 (a)). Image processing was performed using the *Green View Index* (GVI) method (Yang et al., 2009) and operations through the RGB channels of each image at the pixel level as follows: $G_I = (G - R) * (G - B)$, where if G_I is positive, that pixel is considered vegetation. Finally, an average is obtained between the number of pixels considered as vegetation and the total number of pixels found in all evaluated images. As a result, the following was presented: (i) a *Human Perception Mapping* (HPM) of the mentioned cities highlighting the association between green areas and robberies, where a higher concentration of vegetation indeed corresponds to a lower number of robberies;



(a)

(b)

Figure 2.13: (a) Distance map taken between each image, (b) Treepedia website with green area indexes of trees in the city of Boston. Source: (a) (Li et al., 2015b), (b) (MIT-Media-Lab, 2015)

(ii) the Treepedia website (MIT-Media-Lab, 2015) (see Figure2.13 (b)), where this analysis was extended to 30 cities.

2.3.2 High-level feature extraction

Following with high-level feature extraction methods, here we address the concepts of deep networks, whether as feature extractors or for training from scratch. One of these works was carried out by Porzi et al. (2015), which proposes to identify visual elements and assign a respective perception "ranking" (e.g., safety) to street images provided by the *Place Pulse 1.0* dataset and other datasets such as: (i) *ImageNet* (Deng et al.,

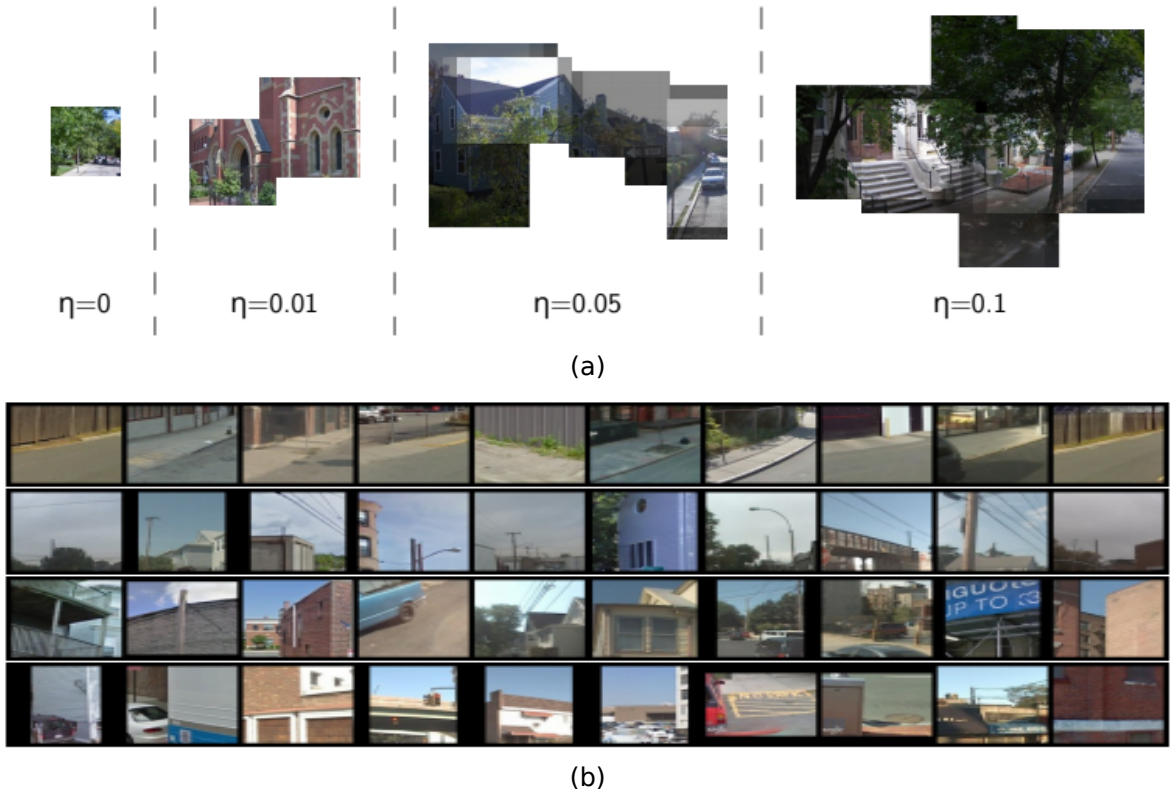


Figure 2.14: (a) Visual representation of the pooling method showing the results for different "n" (pooling factor), it is observed that if $\eta = 0$ it is the classical max pooling and if $\eta = 1$ is is average pooling; (b) the most relevant patterns found with the pooling technique and highly related to the perception of safety found by the network *AlexNet-Places205* + $rCNN_2$. Source: (Porzi et al., 2015)

2009) with 1000 classes of animals and objects; (ii) *Places205* (Zhou et al., 2014) with 205 categories of scenes or environments (e.g., restaurant, forest, cafes, etc.); and (iii) SUN (Xiao et al., 2010) with objects and scene categories. Using *TrueSkill* in the scores and SSIM, GIST, HOG, and *AlexNet* (Krizhevsky et al., 2012) and its rCNN variant (proposed by the authors) with weights previously trained on SUN, *Places205*, and *ImageNet*. The authors proposed a pooling layer defined by $\prod_{\eta_i}(M) = 1 + \lceil \eta_i(wz - 1) \rceil$ where $0 \leq \eta_i \leq 1$ and $M \in \mathbb{R}^{(w \times z)}$ resulting from the convolutions. In Figure 2.14 (a), the results of applying this pooling with different values of η_i can be observed. Finally, for training, they used an SVM to train the features GIST, HOG, SSIM, *AlexNet-ImageNet*, *AlexNet-Places205*, *AlexNet*-SUN, rCNN-*ImageNet*, rCNN-*Places205*, and rCNN-SUN. Then, they added a regularized *RankingSVM* (Joachims, 2002) with l_2 regularization (Tikhonov, 1943). Notable results include: (i) the implementation of a generic pooling function, allowing for obtaining different regions of an image; (ii) the rCNN model that performed best with the configurations: *AlexNet-Places205* + $rCNN_2[m = 24, \eta$=(0, 0.01, 0.05, 0.1)], which means features from the second layer, 24 linear filters, and the specified η values.

In 2016, Dubey et al. (2016) extended the *Place Pulse 1.0* dataset from 4 cities,

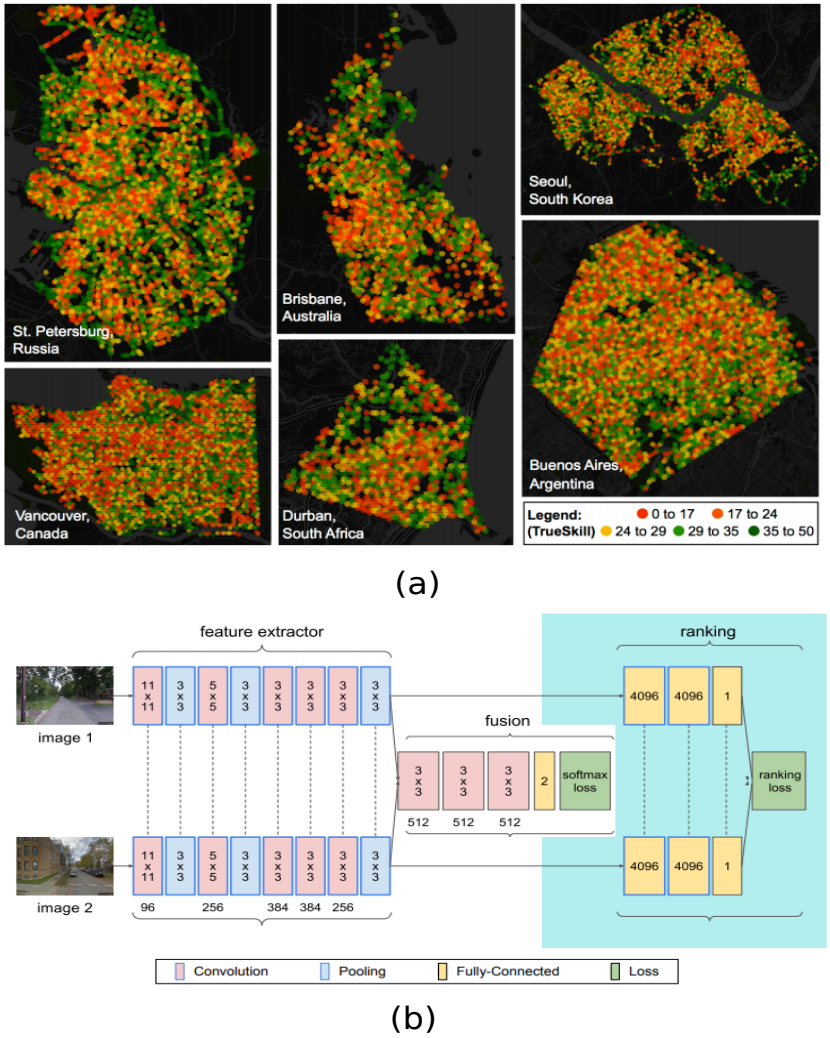


Figure 2.15: (a) Map of security scores predicted by the RSS-CNN network (*VGGNet*) and processed with *TrueSkill*, showing the level of security in a particular city; (b) architecture of the proposed networks (model *AlexNet*). Source: (Dubey et al., 2016).

73,806 comparisons, and 4,136 images evaluated in 4 categories to *Place Pulse 2.0* with 56 cities, 1,223,649 comparisons, and 111,390 images evaluated in 6 categories; where the extended categories are *safe*, *lively*, *boring*, *wealthy*, *depressing*, and *beautiful*. Additionally, based on the results presented in *StreetScore* (Naik et al., 2014), the authors indicated that the analysis conducted in New York and Boston was not entirely correct, due to the low number of comparisons available at that time (an average of 6), which did not meet the minimum required for the *TrueSkill* algorithm to converge. Therefore, two models were proposed: (i) *StreetScore-CNN* (SS-CNN) and (ii) *Ranking SS-CNN* (RSS-CNN). The architecture of SS-CNN consists of fusing the pre-trained networks *AlexNet*, *PlacesNet* (Zhou et al., 2014), and *VGGNet* (Simonyan y Zisserman, 2014); generating the new *SiamesesNet* (Koch et al., 2015). The RSS-CNN is used as a *RankSVM* (Joachims, 2002) as an output function to determine the winner between the comparison of two images. In Figure 2.15 (a), the result of the HPM obtained from the predictions of the SS-CNN and RSS-CNN networks is observed. Key results

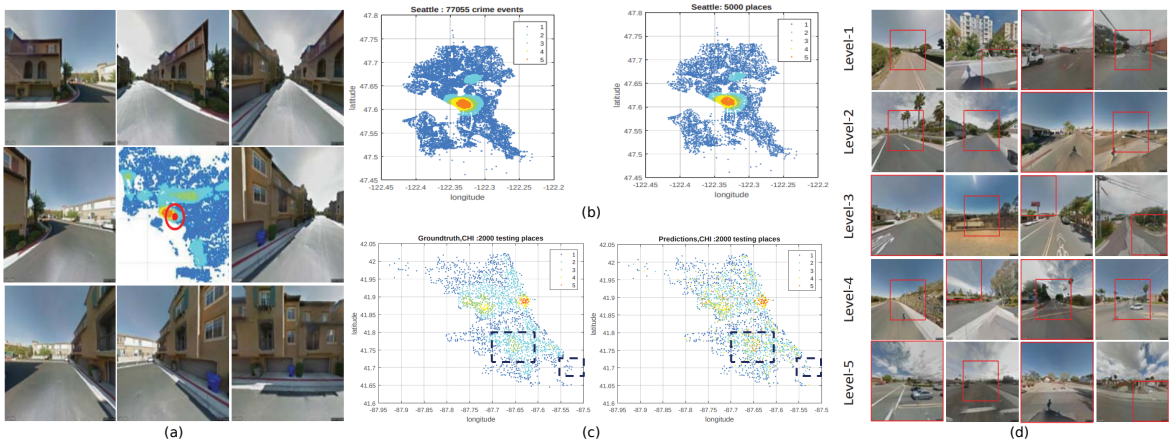


Figure 2.16: (a) The 8 viewpoints taken from a position where a crime occurred. (b) Application of clustering over the city of Seattle with crime events (latitude and longitude) grouped into 5000 clusters represented by a point with latitude and longitude. (c) Comparison of real data and predictions for the city of Chicago, highlighting the places where prediction errors occurred. (d) Examples of regions associated with a level of criminality and with the highest perception score represented by red rectangles in the images. Source: (Liu et al., 2017).

include: (a) the creation and release of the ***Place Pulse 2.0*** dataset. (b) It is not possible to use the *TrueSkill* algorithm because a minimum of around 24 or 36 votes per image is needed, which translates to between 1.2 and 1.9 million comparisons in total. (c) The SS-CNN and RSS-CNN networks for predicting a winner between two images.

Additionally, Liu et al. (2017) propose a method for identifying regions within a given image that have a relationship with urban perception in cities and urban environments (e.g., residences, streets). The experiments were carried out using images from 5000 different locations in the cities of Chicago, San Francisco, Seattle and New York; where in each location 8 different FOV were chosen (see Figure 2.16 (a)). Additionally, each location was selected based on crime data collected by government agencies over a 15-year period (e.g. robberies, fights, assaults, among others), as well as perception scores found by *StreetScore* and *Place Pulse 1.0*; obtaining a total of 1,434,558 crime events from the 4 cities, generating an associated perception score. For data processing, the *Parzen Window* (Parzen, 1962) method was used to estimate the density of each place and quantify the density in 5 levels, places with low density are interpreted as safe. To avoid redundancies in the data, *K-means* was applied to the locations (latitude and longitude) to create 5000 groups (see Figure 2.16 (b)).

For data processing, *bag of street view images* is defined at each location (containing the 8 FOV); Likewise, each image at a different angle is composed of a set of regions (*bag of image regions*), likewise, an image region is in turn recursively decomposed into a set of sub-regions (*bag of sub-regions*). This data set was renamed *Place-Centric*, which is made up of the 5000 images for each of the 4 cities where each image has 8 regions and 40 sub-regions. To train this data, a variation *Multi-Instance regressor*

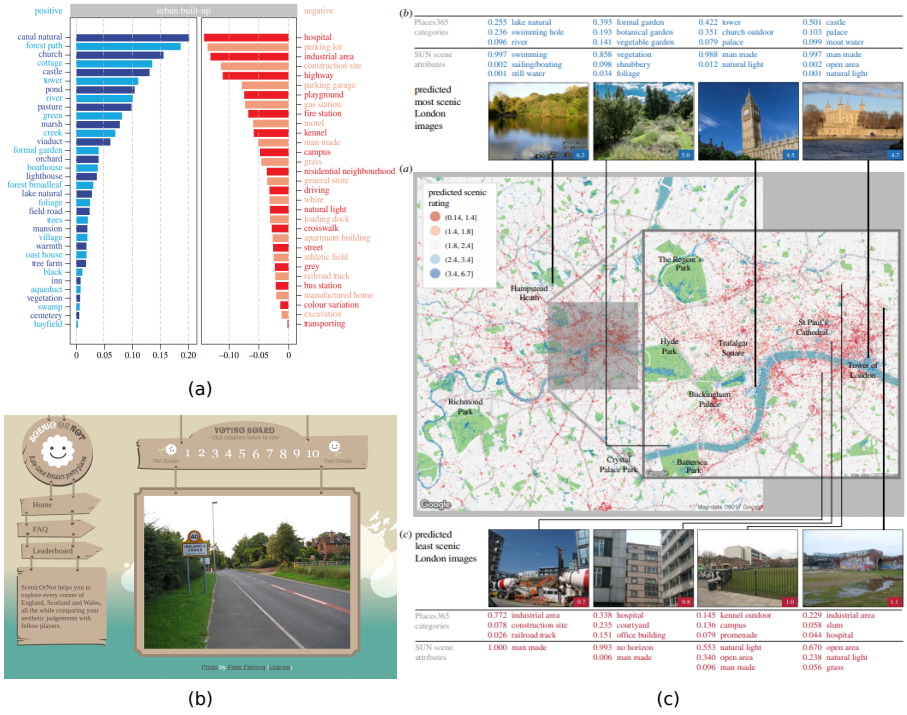


Figure 2.17: (a) Characteristics and attributes that describe an image as scenic, showing the characteristics considered positive or negative. (b) Website *Scenic-or-Not*. (c) General map over the scenes and their respective attributes. Source: (a) (Seresinhe et al., 2017), (b) (UK-gov, 2017; Seresinhe et al., 2017), (c) (Seresinhe et al., 2017)

(MiR) (Ray y Page, 2001) called HDMiR is used; where the inputs are the features extracted by the *VGGNet* network. Figure 2.16 (c) shows the results of the prediction through a HPM of crime levels. As notable results: (i) creation of the *Place-Centric* data set composed of images and criminal records; (ii) a method called HDMiR for the prediction of a perception score based on location density by crime rate.

Another work focused on visual appearance is "*What makes an outdoor space beautiful?*" (Seresinhe et al., 2017), which presents a study on protected spaces in the United Kingdom (e.g. natural areas, landscapes, fields, among others); called scenic. This study was carried out through the game *Scenic-Or-Not* (UK-gov, 2017), which contains more than 217,000 images where one represents 1 km^2 of Great Britain and comes from the website *Geograph* (UK-gov, 2015). To evaluate and identify the attributes contained in each image, the *AlexNet-Places205* network was used with weights previously trained on the *Scene UNderstanding* (SUN) (Xiao et al., 2010) data set to obtain which of the 102 attributes (e.g. trees, flowers, vegetation, shops, among others) were present in each image. Likewise, they contrasted the characteristics of each image using *ResNet-152* previously trained on *Places365* (Zhou et al., 2017), which predicts between 365 categories of type of scenery (e.g. mountains, lake natural, residential, train station, among others). They called the composition of the colors black, blue, brown, lead, green, orange, pink, purple, red, white and yellow present in the images as *ElasticNet*. Which was used to concatenate it with the extractions of *Places205*+SUN and *Places365*, being trained with an SVR to predict the scene level

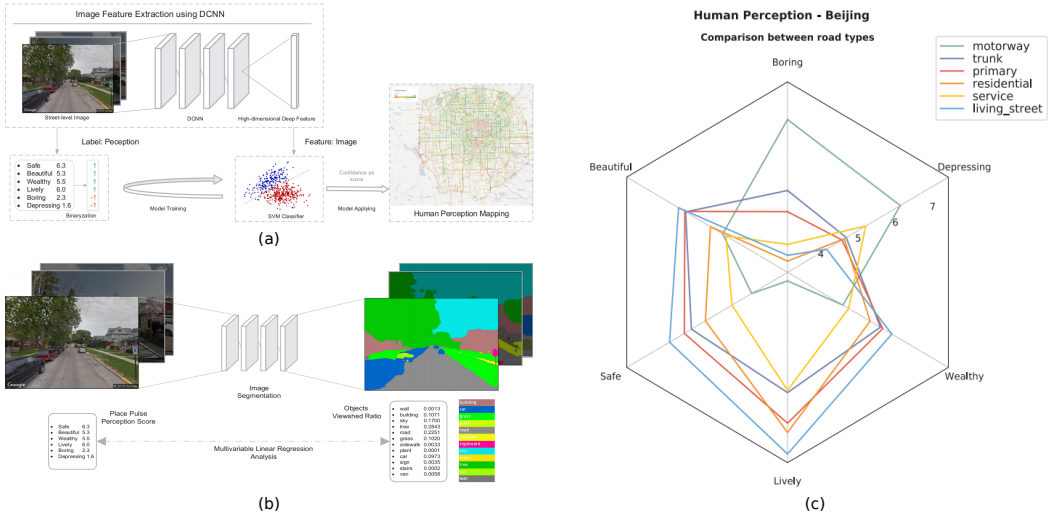


Figure 2.18: (a) Training model on *Place Pulse 2.0* data and generating perception scores about Beijing. (b) Training model on Beijing images, their perceptual scores and visual components (segments) extracted from the streets. (c) Results of the type of perception present in the different types of environment present in the city of Beijing-China. Source: (Zhang et al., 2018).

scores. As relevant results we have: the identification of the most important categories and attributes of each image (see Figure 2.17 (a)); showing that natural features such as coastlines, mountains, natural rivers and man-made structures (e.g. towers, castles and viaducts) lead to places considered more scenic. In contrast, scenes with trees, places with green areas such as grass or fields are considered less scenic (see Figure 2.17 (c)).

In 2018, Zhang et al. (2018) used data from *Place Pulse 2.0* and the object segmenter *PSPNet* (Zhao et al., 2017) to perform an analysis on which features or objects have the greatest influence on the prediction of the perception of security in images of the cities of Shanghai and Beijing obtained through *Tencent Street View Service* (Tencent-Street-View-service, 2016). It should be noted that some of the images of Beijing were obtained from *Place Pulse 2.0*. To obtain an image map, images were chosen at an interval of 50 meters away from each other, with a size of 400×600 pixels and camera angles 0, 90, 180 and 270. With this process, around of 245,388 images of Shanghai and 135,175 images of Beijing. The scores of Beijing-*Place Pulse 2.0* in the 6 categories were extracted. To generate the HPM of Beijing (see Figure 2.18 (a)), a SVR was used with the features extracted from *ResNet*. Then, the *PSPNet* network was used to obtain the objects present in the images to train it using a *Multi Linear Regressor* (MLR) (Tranmer) (see Figure 2.18 (b)). This allowed us to understand the perception of each type of environment, such as roads, streets, parks, residential areas. Figure 2.18 (c) shows the 6 mentioned categories and the level of perception according to each type of environment (residential, roads, etc.). As notable results we obtain: (i) the analysis of the presence of predominant objects in various parts of the city; (ii) the generation of a HPM and a relationship between the presence of objects and types of places.

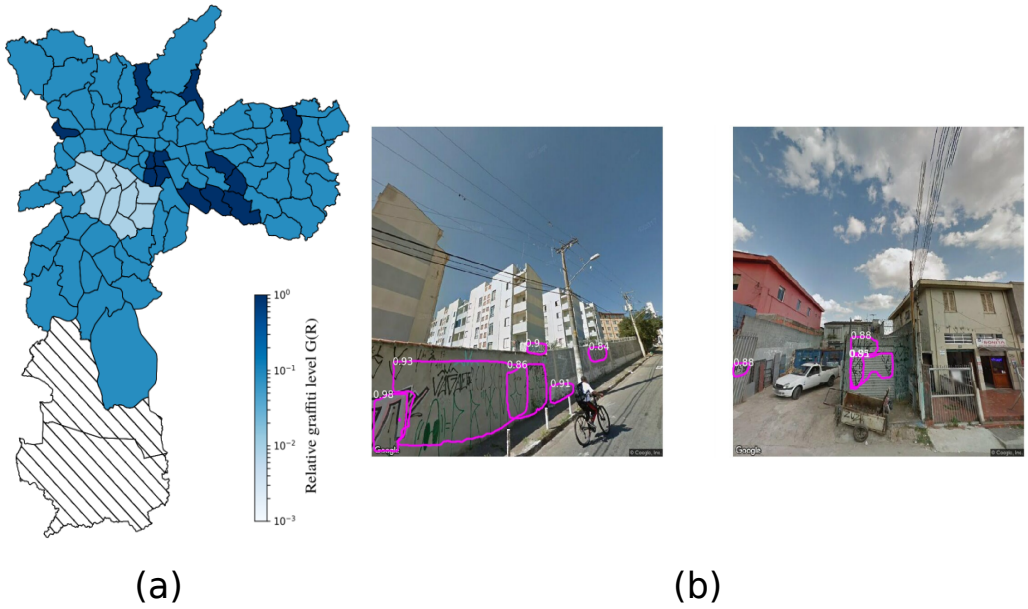


Figure 2.19: (a) Relative level of graffiti in the city of Sao Paulo. (b) Results of graffiti detection in the city of Sao Paulo, the probability of detection is also shown. Source: (Tokuda et al., 2019)

On the other hand, many psychological studies mentioned above concluded that graffiti has a high influence on the presence of crimes in a certain area, as well as influencing the perception of low security on the streets. Due to this, studies have been carried out related to the presence of graffiti and its influence in the city such as Sao Paulo - Brazil (Tokuda et al., 2019), Belo Horizonte - Brazil Diniz y Stafford (2021) and Medellín - Colombia (Alzate et al., 2021) . (I) The first work studied in Sao Paulo used the Mask R-CNN detector (He et al., 2017) with ResNet-101 feature extractor (He et al., 2015) and weights previously trained on MS-COCO (Lin et al., 2014); training was performed on about 10,000 images obtained through GSV. Once trained, the level of presence of graffiti in the city was evaluated (see Figure 2.19 (b)) obtaining an average presence using the GVI method, comparing it spatially with the *Human Development Index* (HDI) (Human Development Index, which measures the rate of growth, birth rate and educational improvement) (see Figure 2.19 (a)); showing as a result that the greater the presence of graffiti in certain places coincide with the places with low HDI. (II) The second work studied in Belo Horizonte, makes a geographical comparison of the presence of graffiti and the crime rate such as: attacks on people, home invasions, sexual violence, drug and weapons trafficking; All these data were obtained from the local police in the years 2011, 2015 and 2017. This study was carried out in the central city of Belo Horizonte, using the *zero-inflated negative binomial regressor* (Garay et al., 2011) method to find a correlation between the presence of graffiti and serious crimes, showing that in Belo Horizonte there is little or no relationship between the two. (III) The third work studied in Medellín explores the types of graffiti present, such as artistic and vandalism. They used the *Faster R-CNN* model (Ren et al., 2017) and the STORM dataset (Charalampos et al., 2019). The results (i) showed that graffiti has a greater presence in commercial places corresponding to the city center, industrial areas and

little presence in residential areas; (ii) they also presented an extended version of the STORM dataset, adding only 373 images.

In this section, different works aimed at predicting street perception (e.g. safety) were presented using different feature extraction methods based on deep convolutional networks, and then training the extracted features using various models such as SVM or derivatives such as *RankSVM* and DCNN. In addition, they showed studies on the impact of the visual appearance of the streets, such as the presence of trees, buildings, green areas or graffiti. These studies were carried out with real crime history data or other data such as house prices, robbery rates, among others; with which they carried out the evaluation of the predictions of their models.

2.4 Interpretation and visualization of extracted features

This section presents the work on the explanation and visualization of the characteristics of a certain correlation between an image and urban perception, taking as a study basis the influence of visual elements present in the analyzed street images. Visualization methods for understanding deep networks are through methods that highlight the features learned in the network, such as *Guided Back-Propagation* (GBP) (Springenberg et al., 2014), *Class Activation Map* (CAM) (Zhou et al., 2016a) and *Grad-CAM* (Selvaraju et al., 2017).

In 2019, continuing the idea presented at Doersch et al. (2012), where *Place Pulse 2.0* and the predictions of possible winners were studied through a *SiameseNet*, Min et al. (2019) addressed the idea to find the most relevant visual characteristics of the prediction of a comparison at the time of making it. That is, if we compare two images in the safe category, what characteristic indicates "this image is more secure?". To do this, they proposed the method called MTDRALN which learns all perceptual attributes simultaneously. This method is composed of two *multi-task siamese networks* (von Platen et al., 2020) with two types of subnetworks, one for classification with weights previously trained on the *Places205* network and another for ranking using a *RankSVM*. In these networks, each *SiameseNet* will learn a relative attribute of each pair of images to be compared; Through a sparse matrix of attributes, it allows easy and fast exchange between characteristics of a set of attributes $A = \{a_m\}_{m=1}^{M}$ for each attribute "m" (p. e.g. insurance) related and unrelated.

The objective of the method is to obtain the aforementioned sparse matrix defined as the attribute values represented by $W = [w_1, w_2, ..., w_M] \in \mathbb{R}^{D \times M}$ which It is divided into groups of relative attributes, for example with *Place Pulse 2.0* it is called a positive group: *safe*, *lively*, *beautiful*, *wealthy* and a negative group: *depressing*, *boring*. To train the model, they reduced the data set by filtering the 161,882 ties within the total of 1,208,808 comparisons. Likewise, they only analyzed the cities of New York, Berlin, Tokyo and Moscow. Finally, to determine which objects are the most influential in each category, they use the *PSPNet* network to obtain object segmentation with the objective of calculating an intersection between the perception scores obtained by the

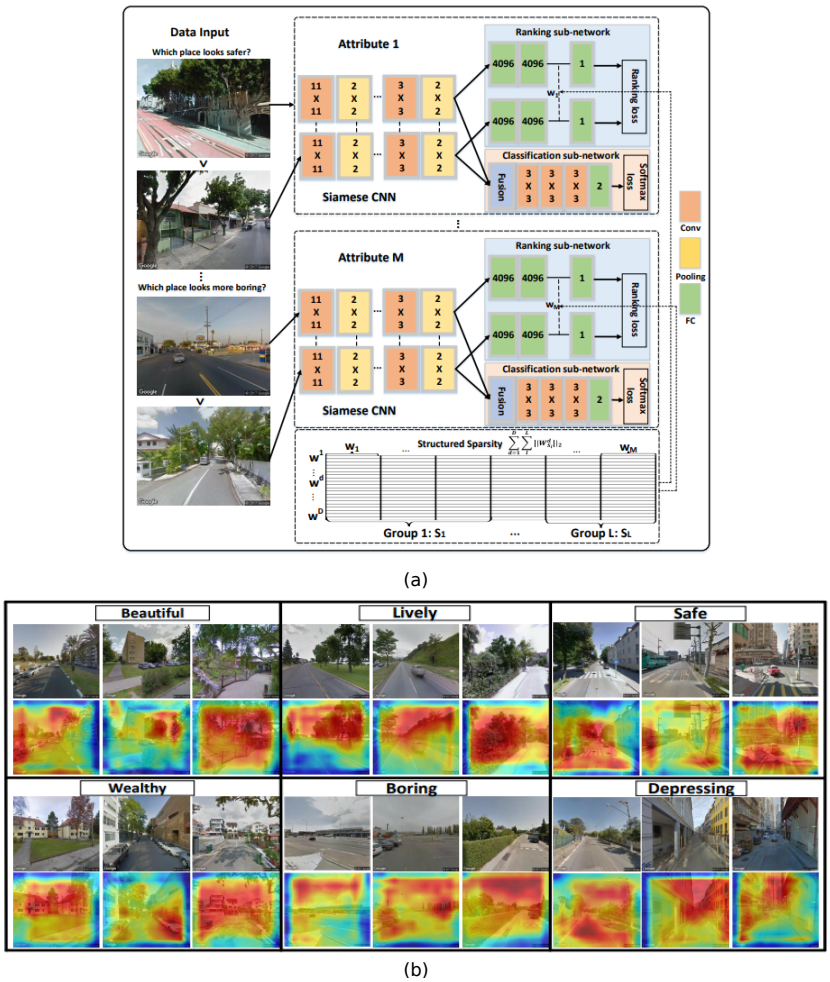


Figure 2.20: (a) Architecture of the *MultiTask Deep Relative Attribute Learning Network* (MTDRALN). (b) Results of applying *Grad-CAM* to the images of each category, showing that the relative attributes encompass almost the entire image, showing that the images classified in the same category have a similar visual appearance. Source: (Min et al., 2019).

RankSVM and the areas generated by the *Grad-CAM* interpretation method as can be seen in Figure 2.20 (b). As a notable result we highlight: (i) implementation of a double *SiamesNet* together with the interpretation using *Grad-CAM* to understand what characteristics of the two input images determine the predominant category in both.

In that same year, Xu et al. (2019) proposed another method also based on the analysis of image comparisons and their characteristic attributes. The objective is to identify which factors within an image directly influence the perception of streets, highlighting the importance of the semantic information provided by the objects through the feature maps of each attribute (e.g. safe); These feature maps are obtained by the *Grad-CAM* technique. Therefore, 2 models are proposed for image comparison, the first predicts the winner by generating a perception score from that comparison; and the second predicts a perceptual score for both images. These models are made up of

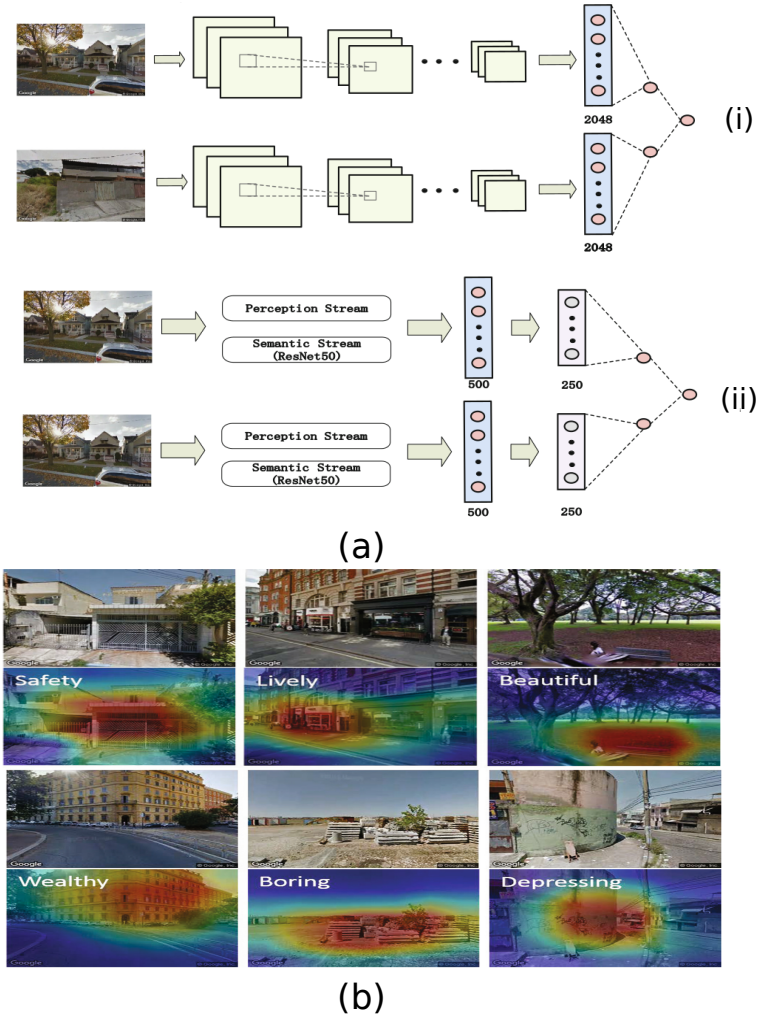


Figure 2.21: (a)-(i) Architecture of *Perception Rank Network* (PRN) and (a)-(ii) *Semantic-Aware Perception Network* (SAPN). (b) Results of applying *Grad-CAM* to some images from the *Place Pulse 2.0* data set, choosing one for each of the categories: *safe, wealth, depression, boring, lively, beautiful* . Source: (Xu et al., 2019).

2 sub-networks called *Perception Stream* and *Semantic Stream*.

Both sub-networks are a *fine-tunned* of *ResNet-50* previously trained on *ImageNet* modifying them from blocks 4 and 5 of the network, changing each *max pooling* by a *Global Average Pooling* (GAP) whose architectures of both networks are shown in Figure 2.21 (a). The first sub-network called PRN evaluates the characteristics between 2 images compared to each other, obtaining as output the regression value of both images. The second sub-network called SAPN evaluates the score of an image; and also has two sub-components that use the *ResNet-50* model. The first *Semantic Stream* (S-Stream) component uses the output of size 1000 from *ImageNet*; This network outputs the score between the two. Instead, the second component uses the GAP extracted from the last convolution ($1 \times 1 \times 2048$) called *Perception Stream* (P-Stream). As notable results we have: the SAPN has a better result than the PRN in predicting perception scores. Furthermore, it shows that the models trained on *safety, lively,*

beautiful, wealthy have better results evaluated among themselves than when evaluating one of these with the categories *boring, depressing* and vice versa.

In this section we saw works related to the interpretation of models such as *Siamese Network* and PRN and the prediction of perception, both models looked for different characteristics between each pair of images obtained from the data to later use a method of interpretation (generally the *Grad-CAM* observed in both methods. Figure 2.20 (b) and Figure 2.21 (b)) shows the results of both works, highlighting the characteristics that the models consider relevant within each image evaluated in a certain perception category.

2.5 Final considerations

In this chapter we have made an exhaustive presentation of the works related to our work, which has two main themes: Analysis and prediction of the perception of security and the use of the data set called *Place Pulse 2.0*. For the most part, we found that a large part of the works studied have as their main focus how to find a method to predict the perception of urban safety using the *Place Pulse* data set or others. The main purpose is to find what aspects of the visual appearance of the streets can influence this perception. Over the years, each work has proposed an increasingly complex model to extract features and highlight them, or on the other hand, they seek to complement information using other data sets and try to describe a more general picture. In our work, our main objective is to analyze, explore and understand the composition of data in *Place Pulse 2.0*; so that once you understand how the data behaves. This approach differs from those mentioned above, since none of them performed an exploratory analysis of the data before proposing some type of solution. This step will allow us to understand what type of model or technique would be most appropriate for the data studied.

A review has been carried out on the works related to: (a) analysis of urban perception and (b) extraction of features and visual components from images and (c) interpretation and visualization of extracted features. We have also described some studies carried out on urban perception and how an attempt is made to quantify the level of perception by explaining through the relationships found in the characteristics learned by models trained on street image data with perception scores. However, it is worth mentioning that in none of the works described and related to the *Place Pulse 2.0* data set an analysis of the data was carried out. In the next chapter we will describe in detail the analysis carried out on the *Place Pulse 2.0* data set.

Chapter 3

Exploratory Analysis of the Data Set *Place Pulse 2.0*

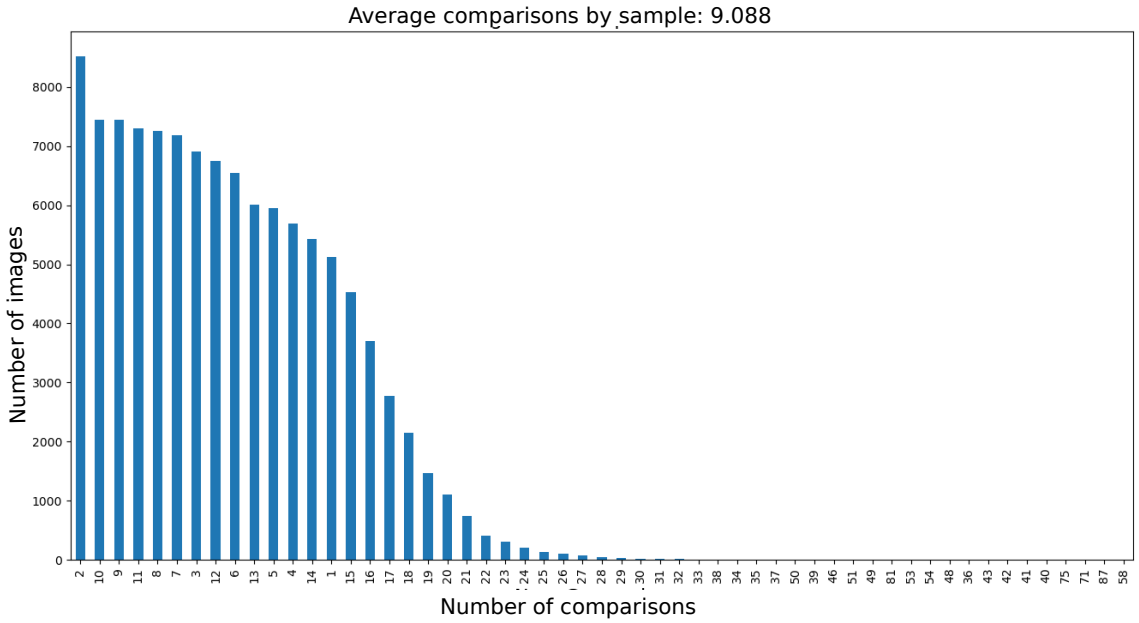


Figure 3.1: We show the number of comparisons in the category *safety* (safe), in which it is observed that the number of comparisons does not exceed on average 10 per image, in addition , most images were compared only 2 times. Source: The author.

As we wrote in Chapter 1, the motivation of this work is to study the perception of urban security through the study and analysis of the *Place Pulse* data set; with which an exploratory study and analysis is proposed to understand the behavior of the data. In general terms, it is known that *Place Pulse 2.0* is a set of comparisons between two images of the same or different cities, evaluated in 6 different categories: safe, depressive, boring, opulent, beautiful and good to live respectively (a From now on, we will refer to them as *safety, depressing, boring, wealthy, beauty* and *lively*) and not necessarily the same number of times. We also know that the average number of comparisons does not exceed 10 comparisons per image (see Figure 3.1), which is why

some algorithms like *TrueSkill* do not work very well (Dubey et al., 2016; Naik et al., 2014) .

It is important to mention that in this work, we will be focusing exclusively on the *safety* perception category, because this category has the largest number of image comparisons. The analysis carried out was divided into small sections that we will describe below: (i) description of the data; (ii) calculation of perception scores; (iii) analysis of the possible "levels of geographic generalization" of the data; and (iv) data disparity analysis.

3.1 Data Description

The data set that we will use in our work is obtained from the website *Place Pulse* (MIT-Media-Lab, 2013) having 2 versions, the first is *Place Pulse 1.0* from 2013 and the second version is *Place Pulse 2.0* from 2016, on which this work is focused. In both versions of *Place Pulse* it is made up of 8 fields: for each comparison there is the positions of the images (latitude and longitude), the image identifiers (right and left), the result of the comparison and the respective category evaluated. Figure 3.2 shows the raw data, that is, how the unprocessed data is. It is observed that they are comparisons between two images emphasizing the winner.

Id img lef	Id img right	winner	left latitude	left longitude	right latitude	right longitude	category
513d7e23fdc9f	513d7ac3fdc9f	draw	40.744156	-73.93557	-33.52638	-70.591309	depressing
513f320cfdc9f	513cc3acfdc9f	left	52.551685	13.416548	29.76381	-95.394621	safe
513e5dc3fdc9f	5140d960fdc9f	right	48.878382	2.403116	53.32932	-6.231007	boring

Figure 3.2: We show the composition of the data set *Place Pulse*, the comparison between the two images and the winner in each category is observed. Source: The author.

Place Pulse 1.0: At the end of 2013, *Place Pulse 1.0* contains 73,806 comparisons, 4,136 images and 4 cities from two countries (US and Austria): New York City, Boston, Linz and Salzburg and three types of comparisons: *safe*, *wealth*, and *unique*.

Place Pulse 2.0: In 2016, *Place Pulse 2.0*, which already contained around

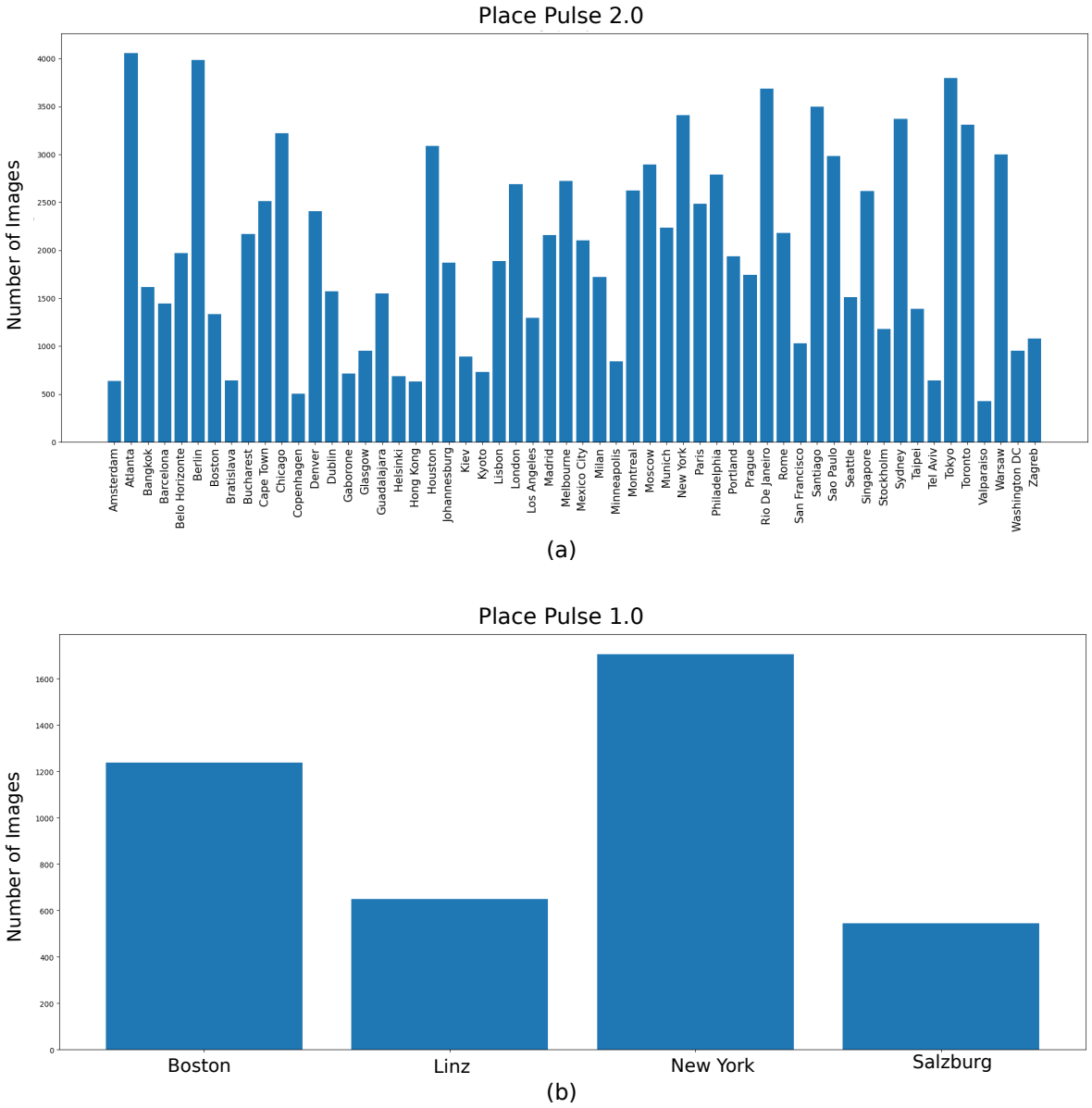


Figure 3.3: (a) Relationship between cities and the number of images within the *Place Pulse 2.0* dataset; (b) Relationship between cities and number of images within the *Place Pulse 1.0* data set. Source: The author.

1.22 million comparisons of 111,390 images of 56 cities from 32 countries across the 5 continents, such as seen in Figure 3.4, from which we can notice that there are more images of cities in Europe and North America than in other places; Likewise, there are six types of categories already mentioned above.

Figure 3.4: Map of the 56 cities with street images contained in the *Place Pulse 2.0* data set, it is observed that Europe and North America have the greatest number of cities evaluated in the website *Place Pulse* (MIT-Media-Lab, 2013) than elsewhere. It is worth mentioning that points with the same color belong to the same country. Source: The author.

3.2 Calculation of Perception Scores

In this section we will describe the equations used to calculate the weighted scores in each category, it is worth mentioning that these are strongly related to the number of times an image won or lost; based on their per-image comparisons. As an example, in Figure 3.1 we show the number of comparisons in the safe category. Next, we will describe and present the equations we used to calculate these perception scores.

By having an image i compared to other images many times in different categories, the percentage of times that i was chosen indicates the intensity of the perception of the image, since of all the images evaluated, it will be obtained that both % was considered to have greater perception (e.g. security) compared to the rest of the images. Furthermore, let $i^{'}$ be an image compared to i, the intensity of $i^{'}$ also affects the intensity of the image i, therefore, the positive rate $isdefinedW_i = \frac{w_i}{w_i+d_i+l_i}$ and the negative rate $L_i = \frac{l_i}{w_i+d_i+l_i}$ of an image i of a certain category. Where w_i indicates how many times he won, l_i how many times he lost and d_i tied; Then from this, the *Q-score* called $q_{i,k}$ is calculated for each image i of a certain category k:

$$q_{i,k} = \frac{10}{3}(W_i + \frac{1}{w_i}(\sum_{k_1=1}^{w_i} V_w(k_1)) - \frac{1}{l_i}(\sum_{k_2=1}^{l_i} V_l(k_2)) + 1) \tag{3.1}$$

Equation 3.1 is interpreted as the positive rate of image i as a weighted approx-

imation over all images k_1 it won and a penalty over all images k_2 it lost ; where V_w is the vector of positive rates of the images with which he won and V_l is the vector of negative rates of the images with which he lost, to finally obtain a score between 0 and 10 obtaining a scale used in studies previous (Nasar et al., 1993; Nasar, 1998).

Once this step is completed, we can extract information from the described data set. Tables 3.1 and 3.2 show the respective statistics for each version. For example, in *Place Pulse 1.0* the number of images per city and the average perception score for each evaluated category are obtained. It is observed that *Place Pulse 2.0* has more information, especially at the continent, country and city level. As well as the number of images associated with each category and their respective average. Figure 3.3 shows the relationship between the number of images per city, in which the increase in the number of cities and amount of images per city is observed. It is also observed that the number of images per city is very uneven, especially in *Place Pulse 2.0*.

Place Pulse 1.0: In Table 3.1 you can see the number of images per city and the average scores for each category obtained from an information cleaning. It is observed that the *safety* category has the highest average score, as well as the largest number of images evaluated. Which is verified by observing that the images and their positions exist and correspond to a certain street.

Place Pulse 2.0: In Table 3.2 you can see the number of images per city, country and continent is larger than the previous version. Furthermore, the average scores for each category are higher. We also highlight that the *safety* category has the highest number of comparisons and the average score. Furthermore, we see that the country USA has the largest number of images and cities evaluated in total. That is why we divide the continent of America into South and North; In addition to the images being very different in visual appearance,

Place Pulse 1.0				
Cities	# images	average *safe*	average *wealth*	average *unique*
New York	1705	4.47	4.31	4.46
Boston	1237	4.93	4.97	4.76
Linz	650	4.85	5.01	4.83
Salzburg	544	4.75	4.89	5.04
Total	4136			

Table 3.1: Statistical data about cities and average scores for each perception category within the *Place Pulse 1.0* dataset obtained from the JSON file downloaded from the website MIT-Media-Lab (2013).

Place Pulse 2.0			
Continent	#countries	# cities	# images
Europa	19	22	38 747
América del Norte	3	17	37 504
América del Sur	2	5	12 524
Asia	5	7	11 417
Oceanía	1	2	6097
África	2	3	5101
Total	32	56	111 390

(a)

Place Pulse 2.0		
Category	# comparisons	average scores
Safety	368 926	5.18
Lively	267 292	5.08
Beautiful	175 361	4.92
Wealthy	152 241	4.89
Depressing	132 467	4.82
Boring	127 362	4.81
Total	1 223 649	

(b)

Table 3.2: Perception Score Statistics: (a) Continents and Images; note that we divided North America and South America, (b) Number of comparisons; Note that *safety* was the most compared category and with the average highest score.

3.3 Analysis of Levels of Geographic Generalization

Up to this point, we have already calculated the respective perception scores for each city, however, once we know which city an image belongs to, we could extend this information to know which country it belongs to and which continent. As we mentioned in the previous chapter, we define "geographic generalization levels" as those regions that we can use to segment the data by city, country, continent or at a global level. Following this idea, the perception scores were calculated at each of these levels; That is, through latitude and longitude data, we filter comparisons between two images whose locations are at the same "geographic generalization level". To calculate the scores, it was preferred to divide the continent of America into two: North America and South America. Once the comparison scores were calculated by filtering images compared in the same city, same country, same continent and at a global level (not to say "same world"), we proceeded to observe the distributions of the scores found.

In Table 3.3 you can see the impact on the scores after being calculated through these levels, likewise, we noticed a reduction in the number of images for each category evaluated being that the security category is the one that maintains the largest number of images in all cases and the highest average perception score (see Table 3.2 (b)). It is

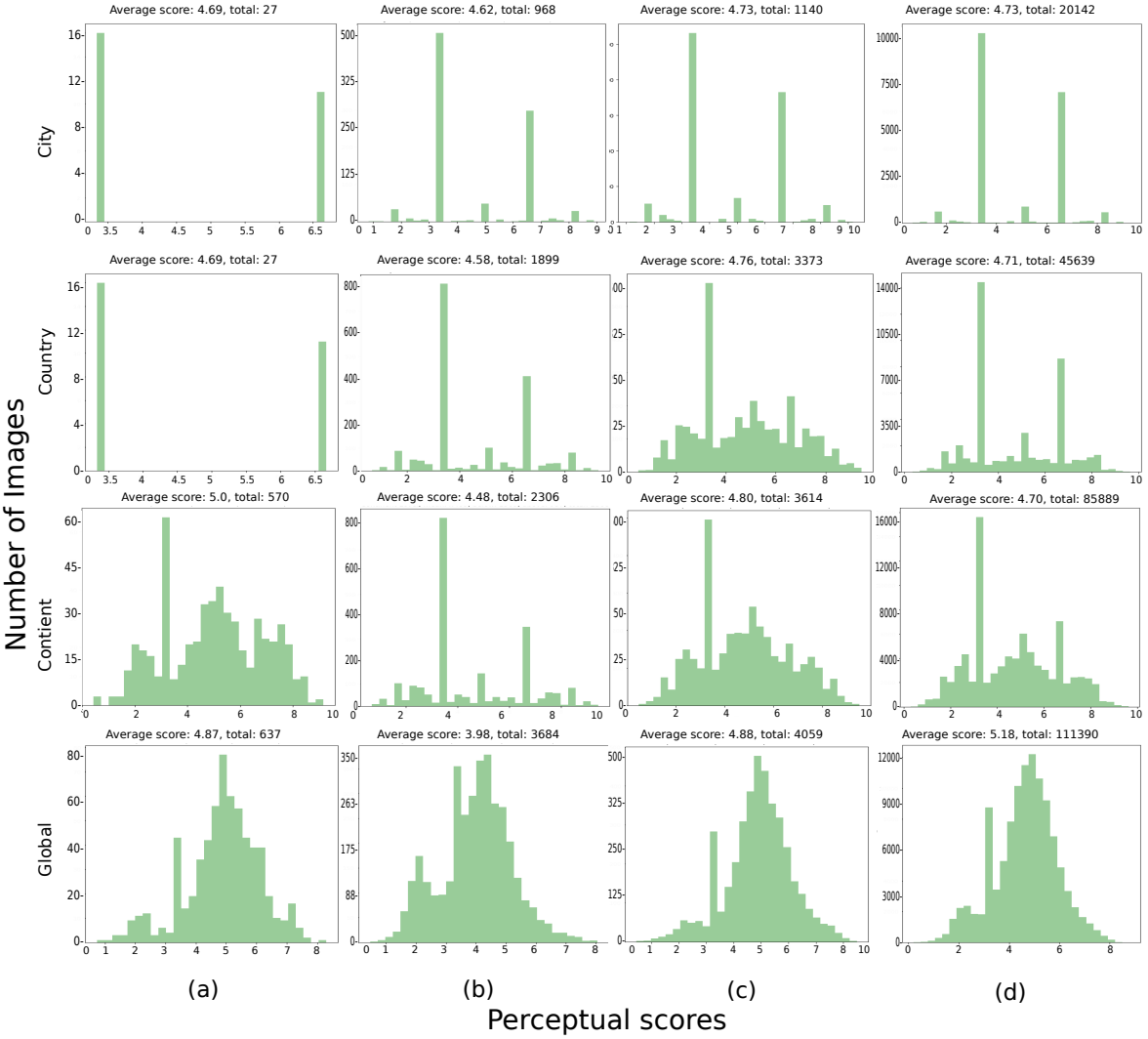


Figure 3.5: Distribution of perception scores at different "geographic generalization levels" in 3 different cities: (a) Amsterdam-Netherlands, single city analyzed; (b) Rio de Janeiro-Brazil with 3 cities; (c) Atlanta-USA, which has 17 cities; (d) all cities (global). Source: The author.

observed that the number of comparisons of an image with another from the same city, country or continent is much lower than with others globally. This only corroborates the idea that the images evaluated in pairs were randomly selected and not filtered by the same location. Likewise, a drastic reduction is observed in the images evaluated in the same city and globally (close to 82 % of the total). It is important to mention that in general, all countries have a maximum of 3 cities, some have only one and the only case with more than 3 cities is the USA. Knowing this, we proceeded to observe the distribution of the scores obtained from each level, we noticed that in those where we only have one city (most countries in Europe) or 2 or three cities (such as Brazil, Chile, Mexico and Japan) the calculation of scores at the city and country levels had no impact. On the other hand, the case of a city in the USA (e.g. Atlanta) did show a considerable change between city and country, this is because having 17 cities, it has a greater number of comparisons at the country level. At the continent level, we have

Place Pulse 2.0						
Geographic level	*safety*	*lively*	*Beautiful*	*Wealthy*	*Depressing*	*Boring*
City	20 143	14 803	9410	7642	6556	6148
Country	45 640	38 216	28 811	24 326	21 171	20 931
Continent	85 890	79 788	66 792	57 780	52 504	52 031
Global	111 390	111 349	110 767	107 796	105 496	106 364

Table 3.3: Number of images obtained per category after performing the calculation at each "geographic generalization level". We see that the *safety* category presents perception scores for all *Place Pulse 2.0* images at a Global level (See Table 3.2).

that Africa and South America present a terrible distribution, this is due not only to the small number of cities (3 in Africa and 5 in South America), but also to the number of comparisons obtained. At a global level, a more equitable distribution is observed with respect to the calculated scores.

As seen in Figure 3.5, we compare the distribution of each case found: countries with one/two cities (Netherlands-Amsterdam), countries with 3 cities (Brazil-Rio de Janeiro) and the USA with 15 cities (only country with more than 3 cities). Due to this, making a specification between the different proposed levels is not possible because the number of comparisons decreases, directly affecting the scores and number of images. Furthermore, it is observed that even after calculating at a global level, a large number of images are observed that have a score of 3.33; This is because most images were compared at most 2 times (see Figure 3.1) of which, it did not win even once. For example, at the city level we noticed that the number of comparisons between 2 images of Rio de Janeiro, of the total of 3684 images, we only obtained 968 with scores 3.33 and 6.66 corresponding to the vast majority of images. From this, we ruled out the possible approach of analyzing cities locally, whose images were compared in greater quantity with other images of different cities.

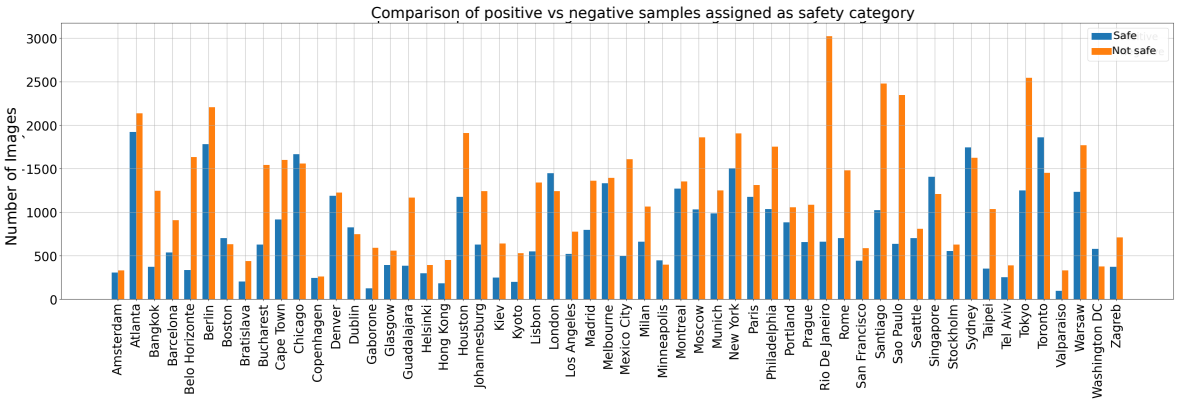


Figure 3.6: Using a threshold of 5.0 to designate safe or unsafe, the disparity in the number of images between safe and unsafe perception by each city. It is observed that in most cities, the unsafe perception is much higher (e.g. Rio de Janeiro and Sao Paulo). Source: The author.

3.4 Data Disparity Analysis

As shown in the previous section, when observing the distributions at the city, country and continent levels of each city, we perceived that having a greater number of images with a score of 3.33 generated a disparity in the data for the city, country and continent. However, at a global level we observed a mitigation in the number of disparate images but still maintaining a high number of images compared with scores of 3.33; Furthermore, it is observed that at a global level, the distribution of scores has a better variety than at other levels. This is because in general the global average is 5.188 (see Table 3.2 (b)). Therefore, it was decided to use the value 5.0 as a threshold for the division of the secure and non-secure classes. This is also because at the global "geographic generalization level", the largest number of images is found with a value of 5.0. $\pm$ 0.1 (see Figure 3.5).

In Figure 3.6, we show the disparity of the scores using the number 5.0 as a threshold to assign the labels to each image (safe and not safe). It can be seen that the vast majority of cities present a high disparity, especially in cities like Rio de Janeiro, Belo Horizont and Sao Paulo, which curiously belong to the same country. While in cities such as Washington DC, Toronto, Sydney, Singapore, London, Boston and Chicago they present (to a lesser extent) a disparity favoring the perception of safety. Additionally, we have cities like Atlanta, Amsterdam, Denver, Dublin, Montreal, Melbourne and Minneapolis with a very close ratio between safe and unsafe. As a final comment, due to the previously obtained results of "geographic generalization levels" and the disparity found, we decided to continue our analysis and experiments focused on perception scores calculated at a global level, not by city, not by country. , nor by continent.

3.5 Final Considerations

In this Chapter, the exploratory analysis carried out on the *Place Pulse 2.0* data set was presented, the respective perception scores in the security category were calculated. These values were calculated across different "geographic generalization levels" such as city, country, continent and global. When analyzing the resulting distribution of scores, it is observed that the best possible score is obtained when we use all the data. Likewise, taking the value of 5.0 as a threshold, a disparity of approximately 11 thousand images is obtained in the "not safe" category. This analysis allowed us to understand the limitations of the data set, such as: (i) the data set is biased by the individual perception of each volunteer who participated in the creation of this data set; (ii) it is necessary to analyze all the data together, it is not possible to carry out regional analysis; (iii) the data present a great disparity from the chosen threshold (following the distributions of the calculated scores). In the next Chapter we present the models and metrics that we will use in the experiments.

Chapter 4

Prediction of Urban Safety Perception

This chapter presents the models, techniques and metrics that we will use to classify security perception. To do this, we define a guideline of experiments to be carried out, we will divide them into 3 groups according to the type of technique used and type of learning. At a general level, we will use two types of learning: (a) Supervised Learning and (b) Semi-Supervised Learning. In the Supervised Learning group, we will use two techniques called *transfer-learning* and *fine-tuning*. In the semi-supervised learning group we will use a GAN model, which is composed of two sub-models called discriminator and generator. As we mentioned before, the main task will be the classification of images between secure and non-secure classes. Said training will be carried out in all 56 cities and at a global level, in addition, the metrics that we will use for these experiments are *Accuracy*, *F1 score* and *Area Under Curve* (AUC) calculated from the values obtained from the textitPrecision-Recall. However, to compare the performance of the models, we mainly rely on the values reported by *AUC* and *F1 score*. These metrics are calculated as:

$$Accuracy = \frac{T_P + T_N}{T_P + T_N + F_P + F_N} \tag{4.1}$$

$$Precision = \frac{T_P}{T_P + F_P} \tag{4.2}$$

$$Recall = \frac{T_P}{T_P + F_N} \tag{4.3}$$

$$F1_{score} = 2\frac{Precision * Recall}{Precision + Recall} \tag{4.4}$$

Where T_P means *True Positive*, T_N means *True Negative*, F_P means *False Positive* and F_N means *True Negative*.

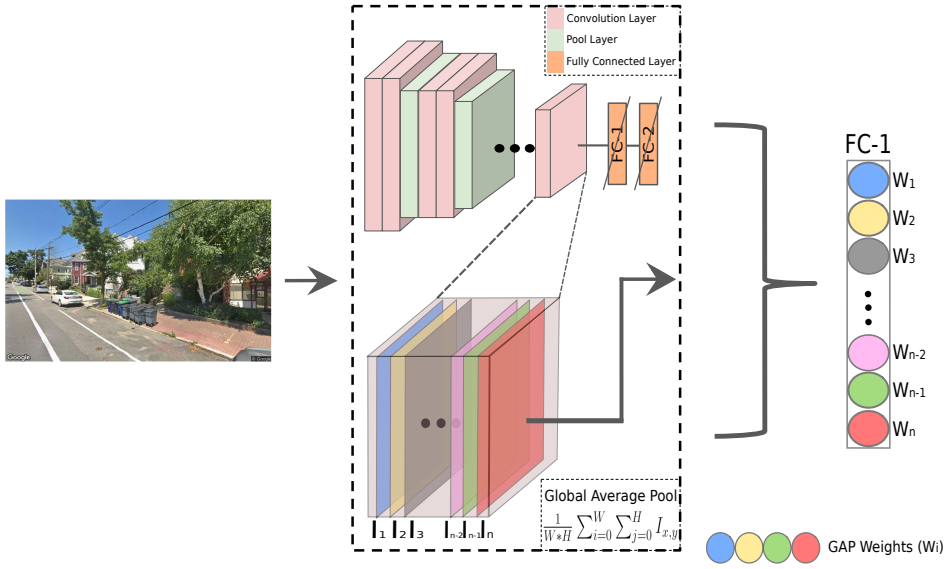


Figure 4.1: The modification of the VGG16 model, called "VGG-GAP" from now on, is presented. For both cases, *baseline* and *fine-tuning* we will use this architecture, in the first case as a feature extractor; for the second, it will be a replacement of the *max pooling* and *flatten* layers of the original model. Source: The author.

4.1 Group Models *Transfer-Learning (Baseline)*

We will define our group of *baseline* models as the set of models based on *transfer learning*. For the experiments, we mainly use the VGG16, ResNet and Exception networks with weights previously trained on *ImageNet* (Russakovsky et al., 2015) and *Places365* (Zhou et al., 2016b)), which we will use as feature extractors. We decided to choose the weights previously trained on two different databases, these are: (i) *ImageNet* presents excellent performance in training and predicting images such as animals or objects; (ii) *Places365* presents a relationship to the prediction of places/scenes, such as residential places, streets, parks, etc. Which is directly related to our studied data set (street images). We decided to give a deeper study to the *VGG16* network over others, this is because the *VGG16* network presented better performance, performance and accuracy in data sets such as *Places365*, SUN and *Scene15* explained and shown in detail in the study of Ali y Zafar (2018); which are data sets composed of images of streets or environments (results reported in Zhou et al. (2017, 2016c)).

Likewise, we make a small change in our *VGGNet baseline* model in which we will remove the last two dense layers and add a GAP, supporting that the use of this technique, with respect to a *max pool* performs better when extracting features (Lin et al., 2013). In Figure 4.1 it can be seen that from the last convolution block, the GAP calculation is carried out, which we use as characteristics; We will call this model "VGG_GAP" to differentiate it from the original "VGG". To make it easier to recognize our models from the group *baseline* we will call them: "TL_VGG16", "TL_VGG16_GAP", "TL_VGG16_Places" and "TL_VGG16_GAP_Places"; where those that have "_Places" are those that use the weights pre-trained on *Places365*. Therefore,

the architectures would be the following: (i) "TL_VGG16" and "TL_VGG16_Places" are the original architecture of VGG16, so the extracted features would be a vector of size 4096 of the last dense layer; (ii) "TL_VGG16_GAP" and "TL_VGG16_GAP_Places" by having the GAP method a vector of size 512 will be extracted, which is a general weighting of each textitfeature map obtained in the last convolution. Finally, once the characteristics have been extracted with these 4 models, we will proceed to perform our safety perception classification training using the linear and non-linear models: (i) *Logistic Regression*: $L(y, f(x)) = \sum_{i=1}^{n} log(e^{(-y_i f(x_i))} + 1)$; (ii) *Ridge Classifier*: $L(y, f(x)) = sgn(||y - f(x)||_2^2 + ||w||_2^2)$; (iii) *Linear SVC*: $L(y, f(x)) = \sum_{i=1}^{n} max(0, 1 - y_i f(x_i))$ and (iv) *RBF SVC*.

4.2 Group Models *Fine-Tuning*

As we describe in sub-Section 4.1, we employ the *VGG16* network due to the good performance reported on the *Places365* dataset. We will use the same architectures described in the previous section *baseline* (see Figure 4.1), with the difference that this group of models which we will call "FT_VGG", ' 'FT_VGG_GAP", "FT_VGG-Places" and "FT_VGG_GAP-Places" will be trained by freezing some convolution layers (limited by our memory and computational power). Similarly, we will assign prefix "Places" or nothing respectively to the pretrained weights of *Places365* and *ImageNet*. For the experiments, we freeze all the layers of each architecture until convolution block 4, so only the last block and the dense ones will be trained (in the case of "FT_VGG16" and "FT_VGG16_Places "). Finally, in all cases we add a last dense layer with only two outputs (corresponding to the safe and non-safe classes) with activation function *Softmax*: $f(x_i) = \frac{e^{x_i}}{\sum_j e^{x_j}}$ and loss function *Categorical Cross-Entropy*.

4.3 Model GAN Semi-Supervised

As we mentioned earlier in Chapter 3 about the possible limitations of *Place Pulse 2.0*, we proposed a method based on semi-supervised learning; which could mitigate and perform well against the characteristics of *Place Pulse 2.0* such as: few images, data with disparity and little data generalization. However, why use a Semi-Supervised model? This set of techniques that use labeled and unlabeled data, present better performance and a considerable improvement during learning in cases where there is data disparity. Because we have little data, the use of a Semi-Supervised GAN (Salimans et al., 2016) is proposed.

We chose to use this method due to the aforementioned limitations (few data and unbalanced data) which a GAN can easily address. The first limitation: **unbalanced data set**, results have been seen from the application of GANs on unbalanced data (Sampath et al., 2021; Zhou et al., 2018), demonstrating that this type of architectures allow learning characteristics of images and classify which class a given image belongs

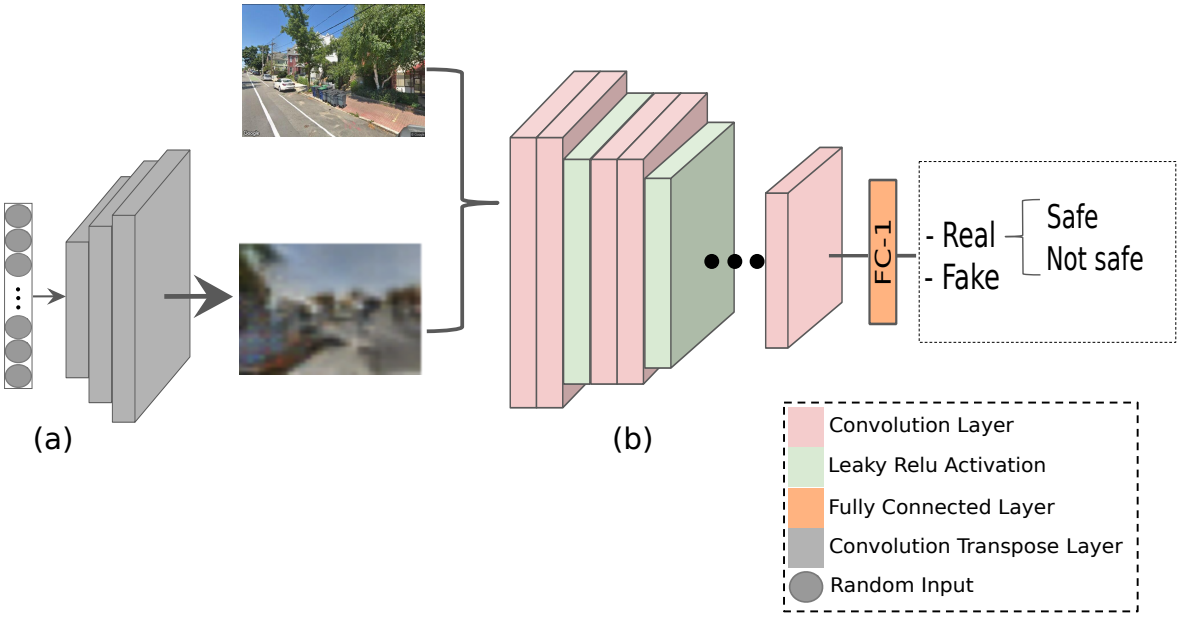


Figure 4.2: The implemented "SSL-GAN" model is presented in general terms, composed of two main components: (a) Generator model, this model is responsible for generating images based on the learned characteristics; (b) Discriminator model, which is responsible for two sub-tasks, the first is the classification of an image between safe or not safe (supervised learning) and the second is the classification of an image between real or fake (unsupervised learning). Source: The author.

to. The second limitation: **data set with few samples**, the use of a GAN with limited data like *Place Pulse 2.0*, having only about 110 thousand images without any *Data Augmentation* is ideal (Karras et al., 2020; Cenggoro et al., 2018). Likewise, a semi-supervised GAN allows us not only to generate data, but also allows us to classify the data. Unlike a GAN *vanilla* which is focused on generating and differentiating between a generated data distribution and the input data distribution. A semi-supervised GAN (from now on we will call it "SSL_GAN") in addition to performing said generation and discrimination task, it also performs the data classification task.

Discriminator					
Layer	Input	Channels	*Kernel size*	*Stride*	Activation
Conv	$32 \times 32 \times 3$	32	3×3	1	LeakyReLU
Conv	$32 \times 32 \times 32$	32	3×3	2	LeakyReLU
DropOut (0.2)	$16 \times 16 \times 32$	-	-	-	-
Conv	$16 \times 16 \times 32$	64	3×3	1	LeakyReLU
Conv	$16 \times 16 \times 64$	64	3×3	2	LeakyReLU
DropOut (0.2)	$8 \times 8 \times 64$	-	-	-	-
Conv	$8 \times 8 \times 64$	128	3×3	1	LeakyReLU
Conv	$8 \times 8 \times 128$	128	3×3	2	LeakyReLU
DropOut (0.2)	$4 \times 4 \times 128$	-	-	-	-
Conv	$4 \times 4 \times 128$	256	3×3	1	LeakyReLU
Flatten	$4 \times 4 \times 256$	-	-	-	-
Dense	128	-	-	-	-
DropOut (0.4)	128	-	-	-	-
Dense	3	-	-	-	Softmax
Total	1 107 882				

Generator					
Layer	Input	Channels	*Kernel size*	*Stride*	Activation
Espacio Latente	100	-	-	-	-
Dense	4096	-	-	-	LeakyReLU
Re-dimensionar	$4 \times 4 \times 256$	-	-	-	-
Deconv	$4 \times 4 \times 256$	256	4×4	2	LeakyReLU
Deconv	$8 \times 8 \times 256$	128	4×4	2	LeakyReLU
Deconv	$16 \times 16 \times 128$	64	4×4	2	LeakyReLU
Conv	$32 \times 32 \times 64$	3	3×3	1	Tanh
Total	2 119 811				

Table 4.1: Configuration of the discriminator and generator models of our semi-supervised GAN called "SSL_GAN" that will be used to train the data set *Place Pulse 2.0* . It is also worth highlighting the number of parameters to train for each model, giving the detail of each layer used in the construction of each model. Likewise, we mention that the value of the parameter α of the function *LeakyReLU* is 0.2.

In summary, a semi-supervised GAN combines an Unsupervised model (classification between data generated with a real or false label) and another Supervised (classification of the data discriminated as real among the proposed classes). Figure 4.2 shows the structure of our GAN; in which from a vector of random values between 0 and 1 (also called "noise") the model can learn to generate artificial images almost as good as those of the original data set. The configuration of our GAN is shown in Table 4.1, in which we can observe the architectures for the discriminator and generator in detail and the operations associated with each one.

4.4 Final Considerations

This Chapter has presented the approaches, methods and models to be used in our training experiments on the previously described *Place Pulse 2.0* data set. In this study we cover 3 types of techniques: (i) *transfer -learning*; (ii) *fine-tuning*; and (iii) a semi-supervised GAN. Likewise, we have described the metrics that we will use to evaluate the performance of the models. These evaluations are intended to verify what type of technique could be the most appropriate given the limitations of *Place Pulse 2.0* presented in Chapter 3. In the next chapter we will see in detail the methods, hyperparameters and experiments carried out. , which are part of our proposed methodology. The behavior of each of the models of the *transfer-learning*, *fine-tuning* groups will be seen in detail with the respective modifications using the GAP and the GAN semi-supervised.

Chapter 5

Results

This Chapter presents the evaluations and reports of results corresponding to the models presented in Chapter 4. To carry out the experiments, an environment composed of an NVIDIA GeForce GTX 1070 GPU, driver 460.91.03, CUDA version 11.2 and 8.11 Gb was used. of VRAM; 12-core Intel(R) Core(TM) i7-8700 CPU @ 3.20GHz each and a total of 31.1 Gb of RAM. In all experiments, except for "SSL_GAN", we performed training per city and training at a global level (using all cities).

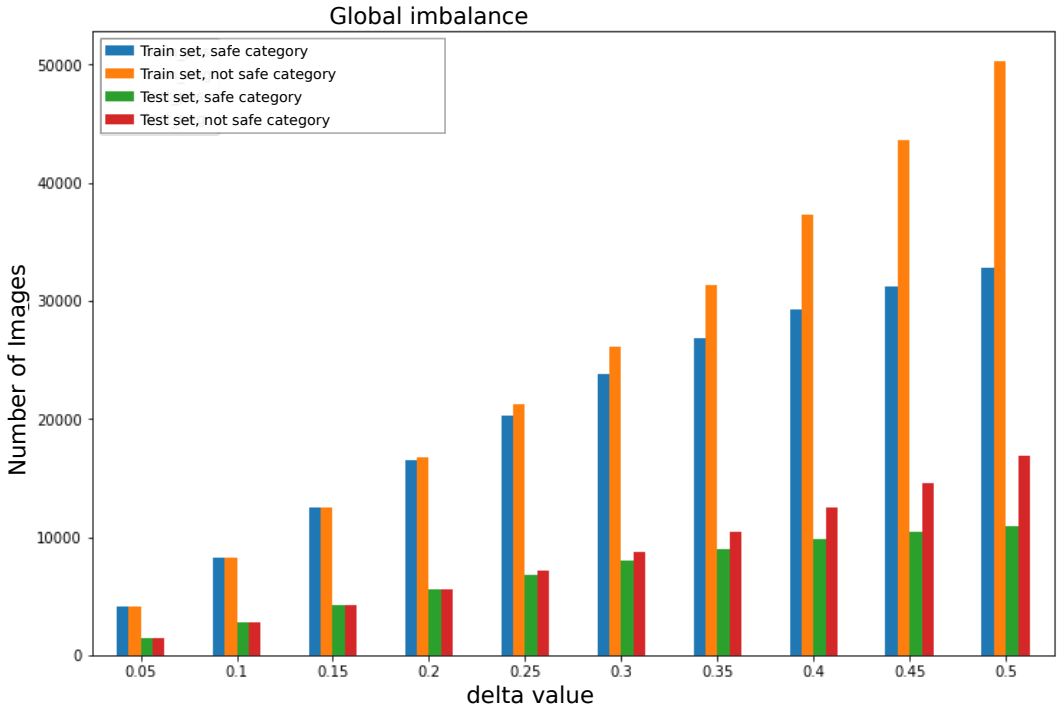


Figure 5.1: Distribution of the disparity in the number of images at a global level (joining all cities) corresponding to each class. It is evident that the higher the value of the parameter δ, the greater the disparity of the data. Source: The author.

As studied in previous works previously mentioned, an additional parameter called δ (delta) was defined. This parameter δ allows us to choose a subset of the

Summary of hyper-parameters							
Method	Input	*Batch*	*Opt*	*LR*	*Ep/It*	*CV*	Data
TL_VGG	4096	-	lbfgs	-	1000	5	Global/City
TL_VGG_GAP	512	-	lbfgs	-	1000	5	Global/City
FT_VGG	$224 \times 224 \times 3$	128	Adam	$1e^{-3}$	100	5	Global/City
FT_VGG_GAP	$224 \times 224 \times 3$	128	Adam	$1e^{-3}$	100	5	Global/City
SSL_GAN_Dis	$32 \times 32 \times 3$	128	Adam	$1e^{-3}$	100	5	Global
SSL_GAN_Gen	100	128	Adam	$1e^{-3}$	100	5	Global

Table 5.1: List of hyper-parameters used in each model during training: (i) *Batch*: size of data to be trained; (ii) LR: learning rate; (iii) *Opt*: Optimizer; (iv) *Ep/It* means *epochs* or *iteration* (only in TL models are iterations used); and (v) *CV* cross-validation. The libraries used for the implementation of the experiments were *Sklearn* (Pedregosa et al., 2011) and *TensorFlow-Keras* v2.3 (Abadi et al., 2015) for the "TL model groups " and "FT"/"SSLGAN" respectively.

total set of images from the perception scores in the following way: The range of values of δ varies from 0.05 to 0.45 which means the percentage of data that We will choose from each class, when $\delta = 0.5$ is understood to be all the data. For example, for a value of $\delta = 0.45$ we will choose the images associated with the 45% of the highest perception scores and the images associated with the 45% of the lowest perception scores.

Figure 5.1 shows the variation of the value δ and its impact on our data set (training and testing). The main idea of this value δ is to observe the behavior of the model when selecting a data set with a similar amount of both classes. Likewise, we observe that as the value of δ increases with a step of 0.05, the class disparity also increases at the city level and at the global level. This shows us that for low values of δ it is possible to have data parity, it is observed that from δ=0.2 there is already a small disparity in the training set. For δ =0.5, we have that the disparity is around 11 thousand images in favor of the unsafe category.

5.1 Experiments performed

First, we need to generally describe how we perform our experiments. For each type of model we use different techniques to find the best parameters that present the best results. To carry out the experiments we use the functions *GridSearchCV* and *KerasClassifier* to create our search mesh of the best parameters and hyper-parameters for our models, likewise, we also use a **5-step cross-validation** dividing the data set into **80 % for training and 20 % for testing**. For the models of the *Transfer-Learning* group, we use the methods *LinearSVC*, *RBF SVC*, *Logistic Regression* and *Ridge Classifier*, in all of them we modify the regularization l_2 using values of α from 10^{-4} to 10^2 and the parameter "solver" varying from *liblinear* and *lbfgs*.

For the models of the group *Fine-Tuning* we add other parameters, these are the *batch size* varying from size 8 to 128 (in powers of 2), we also vary the model optimizer between the set *SGD*, *RMSprop*, *Adagrad*, *Adadelta*, *Adam* and *Adamax* with *learning rate* varying between 10^{-6} up to 10^{-1}. The number of *epochs* was kept constant at 100, because we used the *earlyStopping* method, which controlled the behavior of the model. Likewise, in Table 5.1 we summarize all hyper-parameters that generate the best configuration and best results for each case. Finally, it is important to mention that for each cross-validation the *stratified Kfold* method was used to guarantee that in each validation there is a number of samples with a similar proportion of each secure and non-secure class.

5.2 Group Models *Transfer-Learning* (*Baseline*)

In Figure 5.2 we show the results of the *baseline* "TL_VGG_GAP" model in which we observe that as the value of δ increases from 0.05 to 0.5, The reported *accuracy* is decreasing. However, in Figure 5.2 column (b), the *accuracy* reported for the city of Rio de Janeiro increases as δ increases; We see this similar behavior in other cities with a large amount of uneven data (see Figure 3.6). But we cannot say that these results are correct, because when analyzing the AUC metric, we observe that there is a considerable drop. This tells us that, based on the idea of data disparity, the model can correctly classify the images corresponding to the class with the greatest amount of data. On the other hand, having small values of δ and high values in *accuracy* and AUC indicates that the model is learning to classify correctly; This may be because the lower the value of δ, there is greater proportionality in the number of images of each class (see Figure 5.1). In addition, we must mention that unlike other cities, Rio de Janeiro has much more variety in its streets, from green areas (southern zone) to alleys (northern zone); The opposite is the case in most cities where there is only the presence of green areas (e.g. Atlanta, Berlin, Amsterdam, among others) or skyscrapers (e.g. Boston, New York, Vancouver, among others). As we mentioned, monotonous cities at the image level, such as Atlanta, show a behavior similar to that seen at a global level; the higher the value of δ there is a tendency to decrease the value of the metrics AUC and *accuracy* (see Figure 5.2 (c)).

In Table 5.2 we show the respective averages of the 5 cross-validations reporting the metrics obtained after evaluating them with the training and test data in each model at a global level. We should mention that we included models such as *ResNet50* and *Xception* with weights previously trained in *ImageNet* to do the comparative analysis with respect to the architectures we proposed. However, they did not show any notable results so they were discarded for the *fine-tuning* models. It is observed that models based on *VGGNet* (either pre-trained on *ImageNet* or *Places*) perform better than *Xception* and slightly better than *ResNet*. Furthermore, the "TL_VGG16_GAP" model trained using a *Linear SVC* obtained the best results at the level of *accuracy* and AUC (despite a relatively low value). Likewise, it is observed that in all cases the *rbf SVC* model had a terrible performance, it was not able to learn or differentiate the images

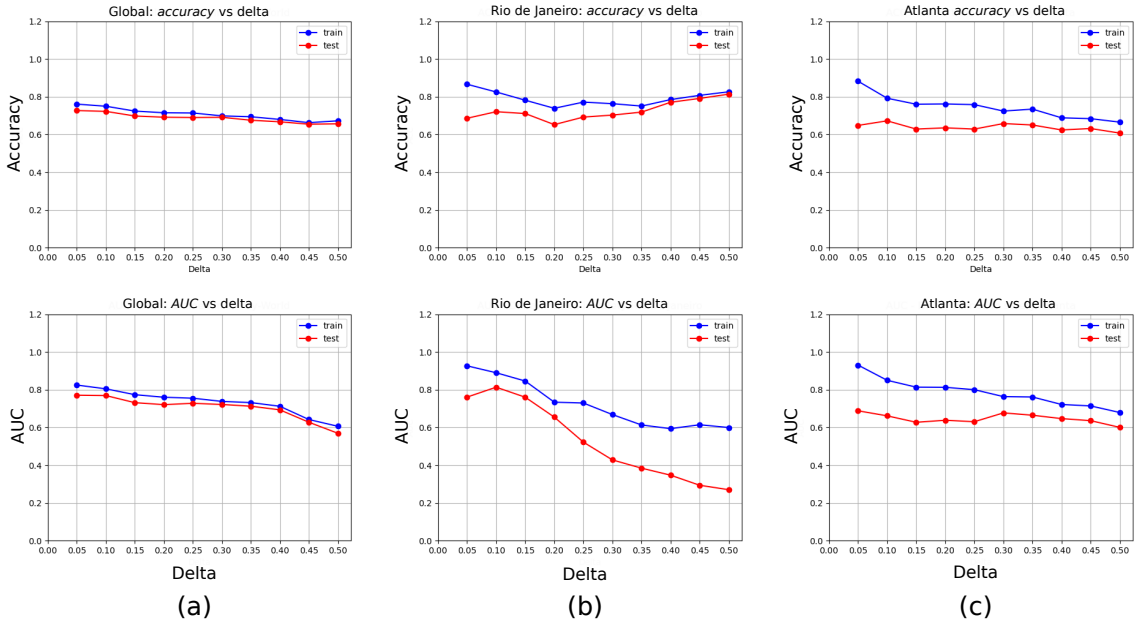


Figure 5.2: Results from "TL_VGG_GAP" (best model). We observe 3 specific cases, in column (a) results are shown at a global level where the decrease in precision is presented as δ increases; column (b) corresponding to Rio de Janeiro, greater precision is presented at higher values of δ; and column (c) Atlanta, a behavior similar to the majority and the global is presented.

between both classes.

		auc		*accuracy*		*f1 score*	
Model	Method	train	test	train	test	train	test
VGG	*LinearSVC*	63.62	56.50	68.85	65.22	54.78	**49.41**
	Logistic	60.63	**57.52**	67.25	**65.72**	51.42	49.07
	Ridge Classifier	64.72	54.75	69.44	64.38	56.50	49.34
	RBF SVC	45.14	42.42	52.13	52.37	46.93	46.59
VGG_GAP	*LinearSVC*	59.01	**57.93**	66.51	**66.09**	49.52	49.06
	Logistic	58.07	57.57	65.95	65.59	46.06	45.61
	Ridge Classifier	59.20	57.93	66.59	65.89	50.27	**49.76**
	RBF SVC	42.93	41.70	50.25	50.35	47.16	46.75
VGG Places	*LinearSVC*	64.44	57.14	69.48	65.79	56.39	51.20
	Logistic	61.74	**58.35**	68.16	**66.44**	53.77	**51.28**
	Ridge Classifier	65.20	55.76	69.84	64.86	57.56	50.67
	RBF SVC	47.32	45.25	56.56	55.69	44.78	44.21
VGG_GAP Places	*LinearSVC*	60.26	**59.76**	67.38	**66.96**	51.65	51.04
	Logistic	59.40	58.97	66.81	66.62	49.16	48.90
	Ridge Classifier	60.45	59.15	67.45	66.94	52.23	**51.53**
	RBF SVC	44.40	42.47	52.59	52.54	43.39	45.05
ResNet50	*Linear SVC*	61.62	59.10	68.10	**66.42**	53.63	50.80
	Logistic	60.04	**59.15**	67.25	66.37	51.47	49.70
	Ridge Classifier	62.11	58.38	68.36	66.08	54.59	**51.00**
	RBF SVC	45.36	44.07	53.46	53.57	44.99	44.98
Xception	*LinearSVC*	55.29	**53.25**	64.43	**63.33**	41.66	39.69
	Logistic	53.48	52.75	63.56	63.14	36.72	35.87
	Ridge Classifier	57.23	52.22	65.22	63.04	45.63	42.11
	RBF SVC	45.57	44.99	49.12	49.12	55.01	**55.05**

Table 5.2: Each train and test column reports the average value of evaluating the models obtained from the 5 cross-validations in each data set. The ResNet and Xception models were pre-trained on *ImageNet*.

5.3 Group Models *Fine-Tuning*

As we describe in Chapter 3 and describe in Table 5.1 we train the *fine-tuning* models in all cities and globally. Therefore, in Figures 5.4 and 5.3 a color map is observed on the *accuracy* obtained by each model trained in each city and at a global level, evaluated in the same city, the other cities and in all of them (global level). We see the need to discard *Xception* due to its terrible performance and reported metrics. It is also observed that there are cities that maintain a high *accuracy* (compared to the average, for example, Rio de Janeiro and Belo Horizonte); Likewise, we can observe that between both types of "FT_VGG" and "FT_VGG_GAP" models, the cities Taipei, Singapore, Philadelphia, London, kyev, Dublin and Cape Town maintain a *accuracy* low in all evaluations.

Table 5.3 reports the average values of the 5 cross-validations carried out on all models, from which we exclude *Xception* due to the low values reported in Table 5.2, only *ResNet50* showed a similarity with the other models described. However, the best model obtained up to this point is "FT_VGG_GAP_Places" which narrowly surpassed "FT_VGG_Places". We also observed that despite maintaining a *accuracy* similar to the "TL" models, the *auc* and *F1 score* increased their values, thus demonstrating that the *fine-tuning* showed better performance compared to the *transfer-Learning* models, which are observed in the reported metrics, in other words, they have a close *accuracy* value but these models are more promising, since that manage to predict the disparity correctly.

	auc		*accuracy*		*f1 score*	
Model "FT"	train	test	train	test	train	test
VGG	77.83	77.42	74.01	64.71	74.01	64.69
VGG_GAP	76.14	75.59	69.40	66.88	69.41	66.87
VGG_Places	74.95	74.75	68.71	67.26	68.71	67.27
VGG_GAP_Places	77.98	**77.5**	70.52	**67.28**	70.52	**67.28**
ResNet50	76.36	72.71	70.36	65.64	67.35	64.98

Table 5.3: Average values of each metric obtained after evaluating each model on the training and test data. Despite the very close results between all the models, we observed a drastic increase in the metrics, as well as understanding that the model manages to distinguish both classes more effectively.

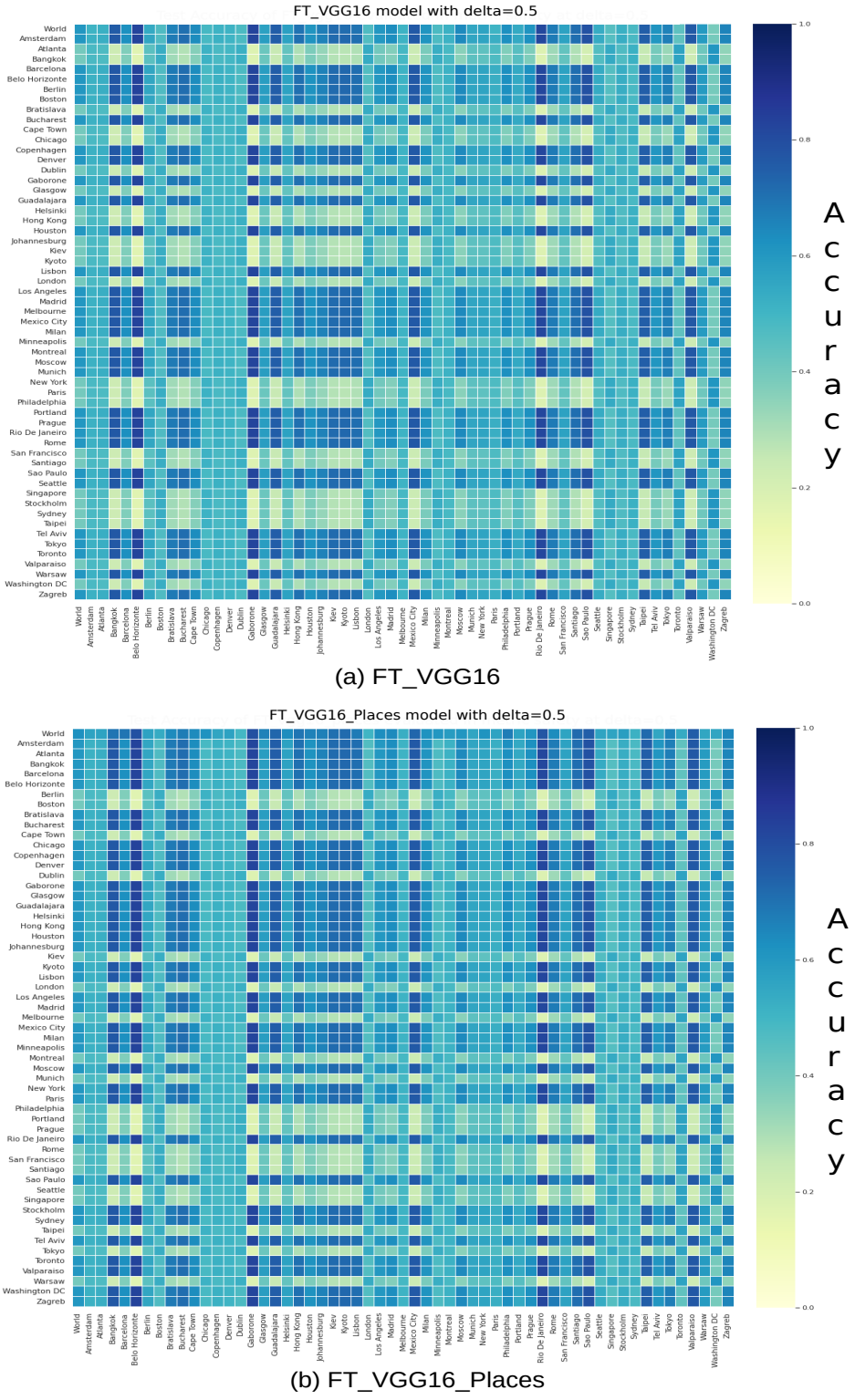
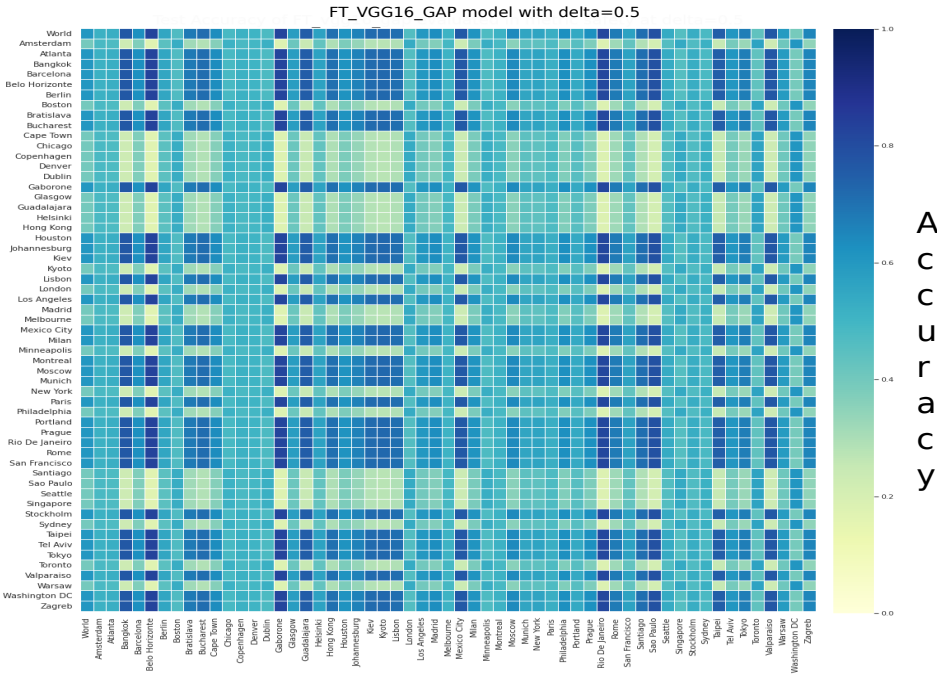
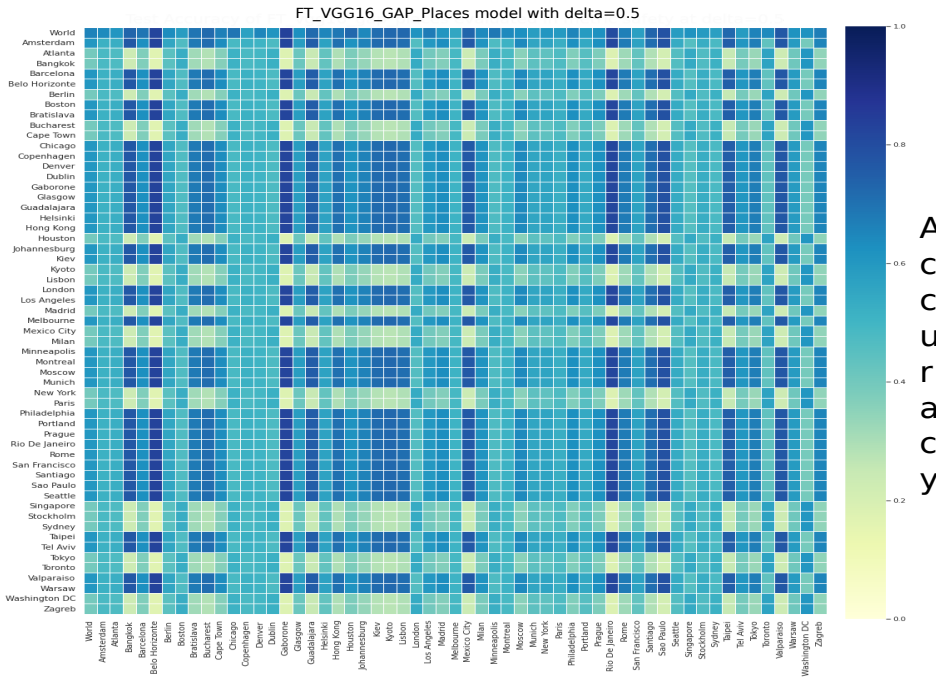


Figure 5.3: We observe in the color map composed of the *accuracy* values of each trained model (row) evaluated in each city (columns) and also at the global "*World*". (a) Results of the "FT_VGG" model evaluations in each city. (b) Results of the "FT_VGG_Places" model evaluations in each city. Source: The author.



(a) FT_VGG16_GAP



(b) FT_VGG16_GAP_Places

Figure 5.4: We observe in the color map composed of the *accuracy* values of each trained model (row) evaluated in each city (columns) and also at the global level "*World*". (a) Results of the "FT_VGG_GAP" model evaluations in each city. (b) Results of the evaluations of the "FT_VGG_GAP_Places" model in each city. Source: The author.

5.4 Model GAN Semi-Supervised

Due to the time it takes to train this model (approximately between 2∼4 days for each cross-validation) it was decided to train using all the data, in order to compare the general results with the other models already shown. As we previously showed in Tables 4.1 and 5.1, we describe the configuration of our discriminator and generator models, as well as the hyper-parameters used. Once the GAN is trained, the metrics reported by the final discriminator model are present in Table 5.4, where it is observed that in the last epoch the model performs over-training (*overfitting*) of the data, which is why our results are low compared to previous models, however, the AUC metric is higher than the previously reported results.

		auc		*accuracy*		*f1 score*	
Model	CV	train	test	train	test	train	test
SSL_GAN 32x32x3	0	80.95	80.97	90.26	59.06	90.26	59.04
	1	81.43	81.45	89.42	61.50	89.42	61.48
	2	81.43	**81.45**	89.56	**62.58**	89.56	**62.57**
	3	80.59	80.66	90.01	61.52	90.01	61.54
	4	80.61	80.63	89.38	61.14	89.38	61.13

Table 5.4: Metrics obtained from the 5 cross-validations evaluated on the training and test data set, it is observed that the AUC reported is much greater than the previous ones with a value above 80 %.

Knowing that the results in Table 5.4 are the evaluations in the last epoch, we proceeded to look for the moment in which the *overfitting* began, noting that it was reached in various iterations in each validation. In Figure 5.5 we highlight the iterations where the best result was achieved for the *accuracy* and *loss* histories. On average, the iteration with the best reported metrics is around iteration 25k. These iterations with the best metrics are shown in Table 5.5.

			auc		*accuracy*		*f1 score*	
Model	CV	iteration	train	test	train	test	train	test
SSL_GAN 32x32x3	0	23 788	73.89	73.89	78.90	78.12	78.90	78.12
	1	58 550	80.21	**80.22**	92.18	**81.25**	92.18	**81.25**
	2	21 951	73.60	73.60	81.25	79.68	81.25	79.68
	3	23 180	73.53	73.53	76.56	78.90	76.56	78.90
	4	8602	69.84	69.84	74.21	78.90	74.21	78.90

Table 5.5: Metrics reported after evaluating each model with the highest *accuracy* reported during training. Some stability is observed in the reported values.

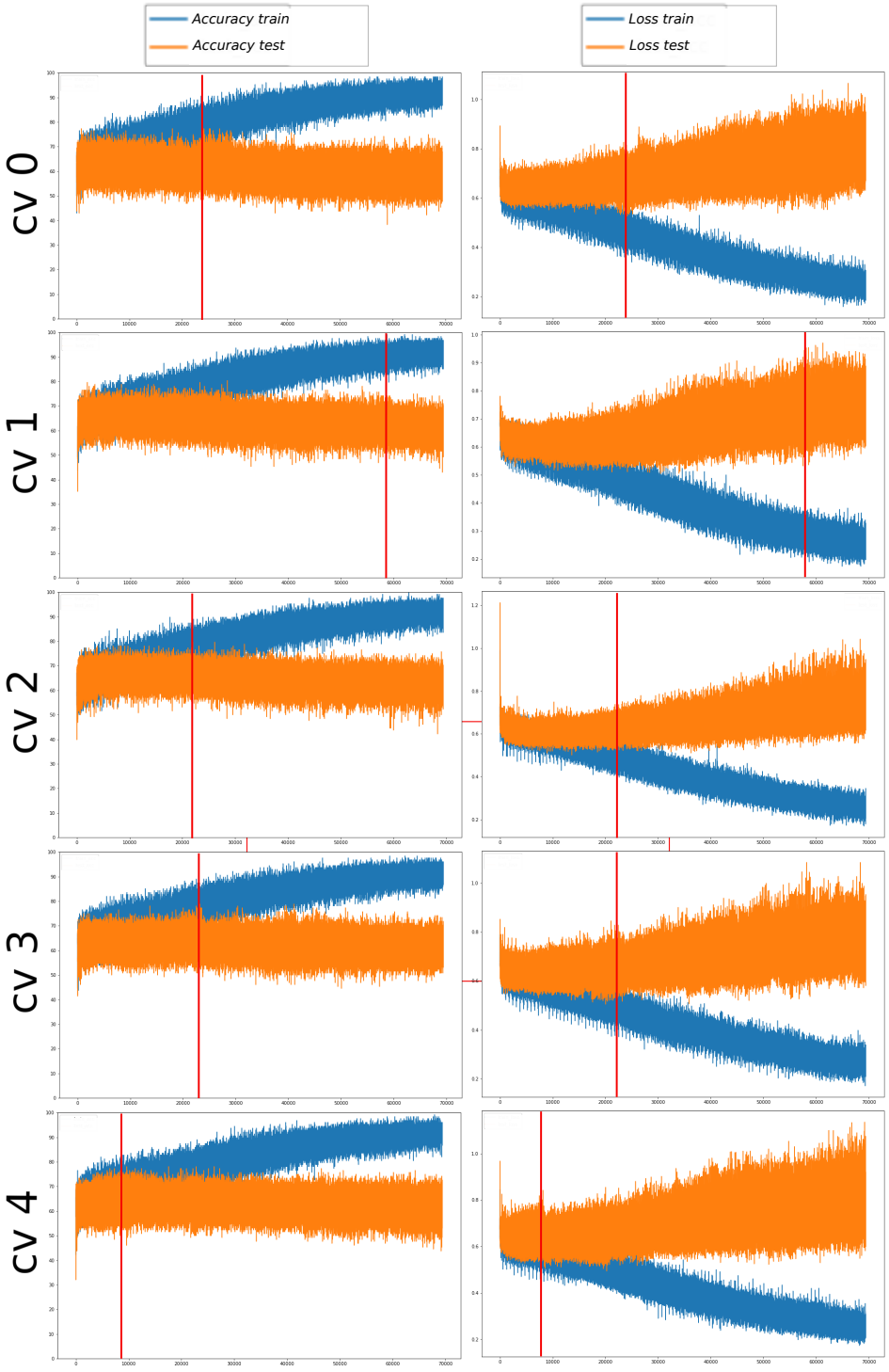


Figure 5.5: History of *accuracy* and *loss* in each cross-validation (CV), we highlight the iteration or training step with highest value (left column) across a red line. In the right column, we highlight the same iteration or step for the *loss*. As reported in Table 5.5. Source: The author.

Figure 5.6: (a) Real images of the data set; (b) Images generated in the last training step. The high quality of the generated images of size $32 \times 32 \times 3$ is observed, making it difficult to distinguish them visually. Source: The author.

Finally, in Figure 5.5 you can see the history of the *accuracy* and *loss* of the training, we frame in each figure a red line corresponding to the iteration where the highest *w*as reached accuracy. An important observation is the fact that *auc* is greater than *F1 score*. This is due to the previously analyzed data disparity, with the data set with the largest number of examples being the unsafe category; then we can affirm that our model is more robust in identifying secure and non-secure examples. We also observe that the metrics *auc* and *F1 score* present a value close to each other, giving us to understand that these models achieve stability when evaluating the data. This stability is strongly related to: (i) how well it identifies the classes, (ii) how many samples it manages to classify correctly, and (iii) the relationship between successes and errors in the prediction. Additionally, in Figure 5.6 we show a set of images generated by our generating model 5.1 (b). The good quality of the generated images of size $32 \times 32 \times 3$ is observed, which can be compared and confused with real images.

5.5 Website

In the present work, the need was seen to have a web system with the purpose of reconciling interaction and visualization of the results of the data training in a quick and simple way. Likewise, it is possible to observe the results and metrics reported for each value of the variable δ. It is also possible to visualize the results of each cross-validation by reporting the previously defined metrics, the AUC graphs calculated from the *Precision-Recall* and the average of said metrics. Also, a summary of the results of each method used. The web system has a simple design, focused mainly on presenting training results, as well as a direct comparison between each method used. The web

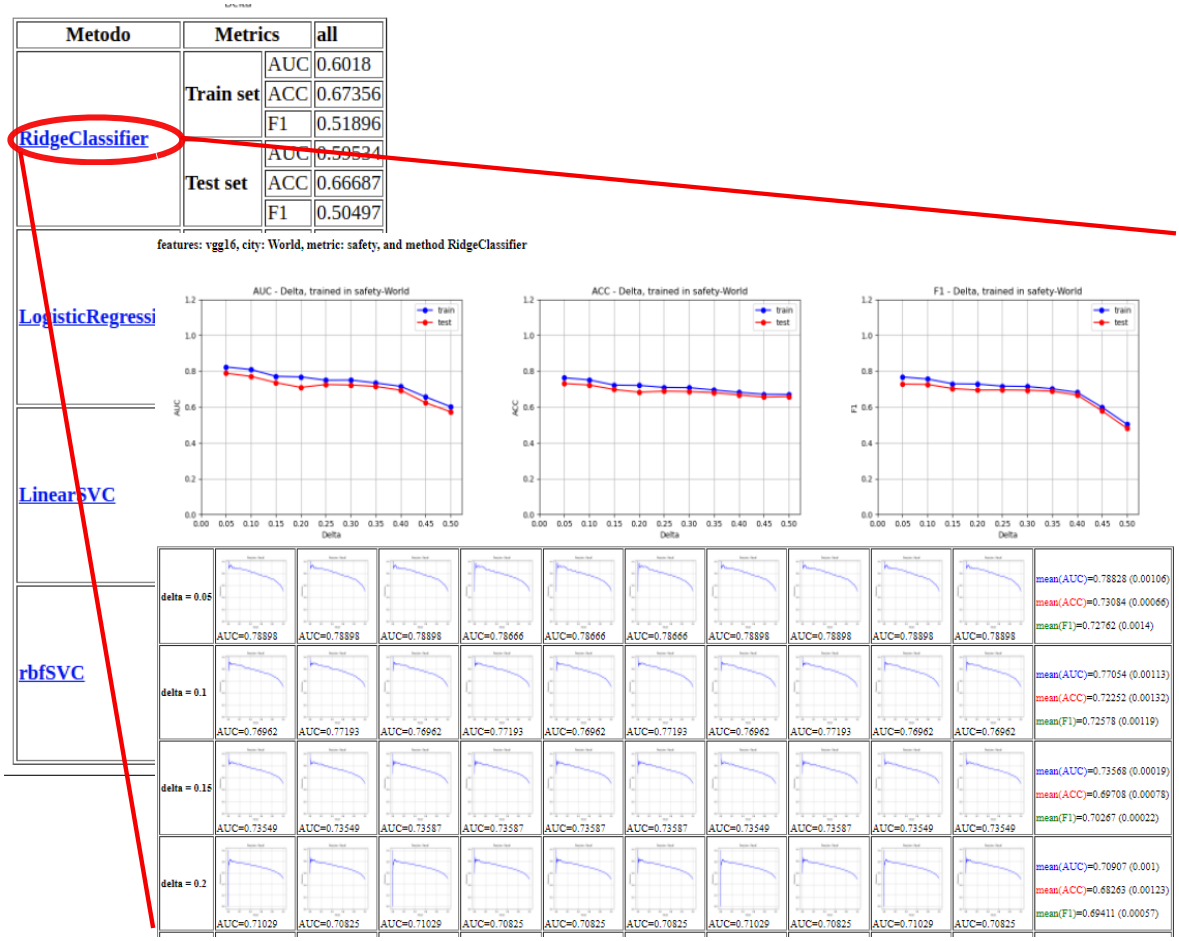


Figure 5.7: This tab corresponds to the results of the cross-validations and a summary of the reported metrics, as well as the graphs associated with the value of each metric for each value of δ. Note: this image corresponds to a model from the *transfer-learning* group. Source: The author.

system is composed of 3 main panels: (i) *Baseline* results; (ii) Results of the *Fine-Tuning* models; and (iii) "SSL_GAN" results. In Figure 5.7 we show the appearance of the website. Starting with a table where the metrics obtained from each method in the training and test data sets are reported. It can be seen that to see the training results in detail, just click on the name of the method and you will be redirected to a tab with details such as the cross-validations, the graphs of each *Precision-Recall* and the general graph of the textitAUC, *accuracy* and *F1 score* reported for each value of δ.

5.6 Final Considerations

The results of the evaluations of the models previously described in Chapter 4 have been presented, as well as the metrics reported in each model using a 5-set cross-validation. For the experiments, the data was divided into 80 % for training and 20 % for testing.

From the results obtained, we observe that the semi-supervised model presents a more stable behavior with respect to the others (when observing the values of the reported metrics). We observed that not only the *AUC*, but also the *accuracy* and *F1 score* resulted in a high and very close value, which was expected. In the following, we present the discussions and limitations.

Chapter 6

Discussions and Limitations

6.1 Discussions

In this work, a methodology has been described that allows studying and analyzing the *Place Pulse 2.0* data set with the aim of finding and highlighting the possible limitations it may present; This motivation is because in the vast majority of works reviewed, they always focus more on the search for a model (increasingly more complex) that has the best performance with *Place Pulse 2.0*. However, none of them performed any prior analysis of the data.

6.1.1 Exploratory analysis of the data set *Place Pulse 2.0*

The analysis of *Place Pulse 2.0* begins by calculating the 111,390 perception scores in the safety category on all streets using the Equation 3.1 described in Chapter 3, the final values of which are in a range from 0 (not very confident) to 10 (very confident). During the calculation process, the idea was raised of analyzing the data by defining geographic regions called "geographic generalization levels" that cover the comparisons made in 2 images at the following levels: (i) same city; (ii) same country (including all cities in that country); (iii) same continent; and (iv) global (all data). Once these calculations were carried out, we were able to observe 2 problems: (a) the loss of information: as shown in Table 3.3, we see that as we use a smaller region, the number of images decreases considerably; (b) the distribution of perception scores is unreliable, as shown in Figure 3.5. We see that in cities with few compared images (e.g. Amsterdam) they present scores with greater concentration at 3.33 and 6.66; this is because the calculation of perception scores also depends on the number of comparisons made (see Figure 3.5 (a)).

This behavior is evident in all cities. At the country level, the distribution will change depending on how many cities are in that country. For example, Brazil has 3 cities and for the data from Rio de Janeiro a slight but insufficient change is observed

(see Figure 3.5 (b)). Likewise, in the case of the USA, which contains 17 cities, it is observed that the number of comparisons is not sufficient (see Figure 3.5 (c)). At the continent level we can already observe a significant change in distribution. However, in continents with few cities such as South America, Africa, and Asia, this lack of comparisons is still observed. The opposite is true for North America (17 cities) and Europe (22 cities). From these results, we conclude that it is not possible to use *Place Pulse 2.0* without considering comparisons at a global level (all cities) due to the few comparisons made between images of the same city.

Another result of observing the distributions is the number of images concentrated in the interval of 4.5 and 5.5, as we describe in Section 3.2 resembles a Gaussian distribution with a center close to the value of 5.0, this being value close to the average of all images (5.18). For this reason, to label the images in the safe and unsafe categories, a threshold of 5.0 was established. From this threshold, it is observed that we have a data disparity, that is, the number of images of the safe class was different from that of the non-safe class. As Figure 3.6 shows, it occurs in almost all cities. This threshold is not possible to change since by decreasing the threshold, we would be giving priority to the images that were compared in greater quantity, and on the contrary, by increasing the threshold we would be increasing the disparity of images.

6.1.2 Prediction of urban safety perception

Following the analysis of the data, we focus on evaluating different types of approaches based on the convolutional network model, which can present good performance against the nature of *Place Pulse 2.0*, likewise, good performance in the urban safety perception classification task (or simply safety perception). For this, a *pipeline* of experiments was proposed based on two types of learning already mentioned: (i) Supervised Learning and (ii) Semi-Supervised Learning. The metrics used to evaluate our models are AUC, *F1 Score* generated by *Presicion-Recall* and *Accuracy*. It was decided to use these metrics because the main task is classification of two categories (secure and non-secure). For supervised learning, it was decided to use two techniques: **transfer-learning** and **fine-tuning**, which used networks such as *VGGNet*, *ResNet50* and *Xception*. The three networks pre-trained on *ImageNet*, in addition, for the *VGGNet* model the pre-trained weights from *Places365* were also used due to the nature of the data; which are images of exterior and interior places such as residential areas, streets or restaurants.

The first result corresponds to the **transfer-learning** models, called "TL", which consist of using models based on DCNN as feature extractors (outputs of the last layer). As described in Section 5.2, we trained these extractors using 4 linear and non-linear methods, the results of which are in Table 5.2. We observe that despite having an *accuracy* of 65 % the metrics *F1 score* and *AUC* do not show good performance and this is because the model is predicting with better accuracy the class with the greatest number of examples. Additionally, it was decided to experiment with other models such as *ResNet50* and *Xception* previously trained on *ImageNet* taking into account two points: the number of parameters (less than 23 million each) and the

higher performance than *VGGNet* on said database. However, only *ResNet50* showed similar results to the 4 main models, and *Xception* was discarded. The main results of the "TL" models helped us understand the behavior of the data when trained. Using the variation of the parameter δ defined in Chapter 5 and shown in Figure 5.1, we observe the values of the reported metrics from δ=0.05 to δ=0.5, showing us that despite having a high *accuracy* at δ=0.5, it does not mean that it is a robust model with good performance (see Figure 5.2). Likewise, we verified that for the "geographic level" such as city, countries and continent it was not sustainable to use variations in the value of δ, since in some cities there were a low number of images that made training impossible.

The second result corresponds to the **fine-tuning** models, called "FT", they used the same architectures defined for *transfer-learning* except for *Xception*. For the experiments, we freeze and train from the fifth convolution block for the models based on *VGGNet* and in the case of *ResNet50* from the residual block 14. The results shown in Table 5.3 show an *accuracy* close to those reported in the "TL" group models. However, there was a notable improvement in *AUC* and *F1 score* thus showing that the learning and prediction of the model with respect to both classes improved, because these metrics reflect how good a model is at differentiating the evaluated classes. Figures 5.4 and 5.3 show a *colormap* of the *accuracy* of each trained model evaluated in each city. The main motivation was to observe the performance of models trained and evaluated in different cities. It is observed that the Global model for all the 4 trained models maintains good performance, and therefore, it is because Global includes the data from all cities. In the other cities it is observed that there are cities in which they have high *accuracy* and in others they do not, and that depending on the model, this varies. Furthermore, it is observed that all models maintain the same performance evaluated in cities such as Chicago, Copenhagen, Denver, Dublin, Minneapolis, Montreal, Seattle, New York and Portland.

The third result corresponds to the **semi-supervised GAN** model, which consists of using a set of data with the respective labels between 1 for safe and 0 for not safe that will be released in the supervised model, We will also use another set of real data without any associated label, which will be used to train the generator and the unsupervised model. Just like a GAN *vanilla*, our "SSL_GAN" works in a similar way, except for the difference in the discriminator. The discriminator is responsible not only for discerning between real or fake, it will also identify which class an evaluated image belongs to, whether it is a generated one or one from the data set. As we explained in Section 4.3, the choice of a semi-supervised model was due to the limitations previously found in the data set. Due to the difference in training time between the different models, the "SSL_GAN" was only trained using all the data. Table 6.1 gives an approximate idea of the training time used by each model in each data set (global or city), it is observed that the "SSL_GAN" time is much higher compared to the other models. We mention that the reported time corresponding to "56 city" is an average training time for all cities separately.

	Training time for each model	
Method	Data type	Average time
SSL_GAN (32×32)	Global	∼1 week and a half
FT_VGG	Global	∼8 hours
FT_VGG	56 Cities	∼6 hours
FT_VGG_GAP	Global	∼7 hours
FT_VGG_GAP	56 Cities	∼5 hours
TL_VGG	Global	∼15 minutes
TL_VGG	56 Cities	∼10 minutes
TL_VGG_GAP	Global	∼9 minutes
TL_VGG_GAP	56 Cities	∼6 minutes

Table 6.1: Table of average training times carried out for each model performing the 5 cross-validations. For the case of "TL" we are reporting the total average of training the 4 models for each case.

6.2 Limitations

The study of urban perception is a very complex field since it is not possible to describe a general perception (Wilson y Kelling, 1982) and the perception for each person varies depending on the environment where a person lives (Keizer et al., 2008), that is, that perception is very relative and differentiated for each person. Thus, we present the limitations found in the present study, which are strongly linked to the data set. As mentioned above, the *Place Pulse 2.0* dataset analyzed allowed us to understand and establish a methodology focused directly on the data, rather than the conventional method of thinking of some complex model to fit.

6.2.1 Individual perception of participants

The construction of the *Place Pulse* data set was from comparisons between two images through a website. For this, various volunteers carried out the vote on a set of completely random images. That said, some images may have been compared and voted on by one or a few specific volunteers. This generates a difficulty because a person's perception of security is influenced by the environment where they live, generating a biased individual criterion. This is a problem when trying to carry out a specialized study by city, country, continent because the images were compared and voted on by various users from different places.

This limitation was not possible to solve, since it is a problem inherent to the data set, starting from its idealization and how it was constructed (they did not take into account individual perceptions, individuals from a similar region, etc.).

6.2.2 Little amount of data/images

At a general level, *Place Pulse 2.0* is made up of 1.22 million comparisons in total, in Table 3.2 (b) we show the respective statistics of each category, seeing that the category corresponding to security presents 368,926 comparisons, this being around 30.14 % of the total comparisons. Likewise, although a total of 111,390 images in said category were compared, the data set only has 110,988 images. Comparing with other image datasets with millions of data, 111 thousand images does not compare, furthermore, datasets such as CIFAR10 or MNIST with 60,000 and 50,000 respectively, have a similar proportion number of images per class. In our case, the number of images is not homogeneous per city, that is, we have the case of the city of Atlanta with 4,034 and cases like Amsterdam with 637 images (see Figure 3.3 (a)). Which, by not having a homogeneous set for each case, does not allow for a specialized analysis in each case.

This limitation was possible to mitigate through the use of the semi-supervised model "SSL_GAN", the small number of images and the disproportion of images per city was cushioned through the synthetic images generated in each iteration, these being used in training.

6.2.3 Generalization across city characteristics

Due to the number of cities and the great variety of the number of images of each city present in *Place Pulse 2.0*, it is not possible to find a model that manages to generalize a prediction with high precision. This occurs in cases such as Atlanta or Berlin, whose images are composed (mostly) of a road and trees or grass on the sides; On the other hand, we have fully urbanized cities like Boston. Likewise, unique (or differentiated) characteristics are found in Tokyo and Kyoto, where the streets present a clear difference from other cities, whether in Europe or North America where there are skyscrapers and very tall buildings.

Furthermore, as we describe in Table 3.3 as we studied the possible "geographic levels of generality" we understood that we had a greater loss of information the more specific the level. For example, at the global level we have at the "geographic level" global we have 111,390 images and at the "geographic level" city we have 20,143 images, meaning a reduction of almost 82 % in the number of images of the "security" category.

This limitation was not possible to mitigate because it is strongly linked to the construction of the data set, evaluating two images randomly without taking into account which images they are compared or how many times they are compared, makes the set of data has to be analyzed in total. Thus leaving no possibility of carrying out specific studies along different "levels of geographical generalization".

6.2.4 Dataset disparity

From the limitations set out above, from the calculation of the perception scores carried out, we observe from Figure 3.5 that for the global level, cities have a distribution similar to a Gaussian distribution with a close center in 5.0, whether in countries with 2 or more cities or one city. The median of the scores is the value 4.72 which is very close to 5.0. Using our threshold of 5.0, we have a disparity of 11,233 images from the median. Using the median as a threshold, that is, dividing 50 % as safe and 50 % as not safe does not make any relevant difference when training the data, in addition, manually observing a subset of the 11,233 images corresponding to such scores do not have a visual appearance that would deserve to be labeled as safe. This limitation was possible to mitigate through the use of the semi-supervised model "SSL_GAN", the disparity of data present was mitigated through the generation of synthetic images by the generator during, which were added to the training in each iteration.

6.3 Final Considerations

In this Chapter we have described in detail the discussions of the experiments carried out and also the limitations found in the *Place Pulse 2.0* data set that were exposed in Chapter 3. We also discuss how it was possible to solve these limitations found in our analysis. Unfortunately, it was not possible to solve all of them, since the nature of the data and its construction prevent a more in-depth analysis based on individual perception.

Chapter 7

Conclusions

In this work an exploratory analysis of the *Place Pulse 2.0* data set was presented; through the different studies and approaches that we explain in detail in Chapter 3. As a result of the exploratory analysis, limitations were found in the data set studied. Of the four limitations found and explained, we were only able to computationally resolve three: Few image samples in general, disproportion in the number of images per city, and class disparity. The only unresolved limitation is strongly related to how the data set was constructed, based on each person's perception and choice of which image is the safest. That is, the data set is limited to the individual perception of each person. It was observed that during the construction of the data, random images were compared and evaluated by a random user. Likewise, it was found that the number of images used by each city is not proportional, having cities with approximately 4 thousand images (Sao Paulo) and others with less than 700 (Amsterdam), which prevents an individual analysis for each city. This disproportion does not allow generalization, which is forcing a dependence on each other (through the calculation of perception scores). However, it was possible to combat this data disproportion through semi-supervised learning, a generative model like GAN allows us to artificially extend the data set through the generation of new data.

Consequently, the results of the evaluations of different classifier models were analyzed and presented using techniques such as *Transfer-Learning*, *Fine-Tuning* and GANs. The evaluations were carried out reporting 3 main metrics: *F1 score* which is an average between *Precision-Recall*; *Accuracy* which reports how many images were correctly predicted in each category; and AUC to determine the proportion in which both categories were correctly classified. The main metric used was the AUC because having a data disparity, it is also observed that as a specialized model is trained, a better AUC value is obtained. We notice that an increase is obtained from ∼59 % (*Transfer-Learning*) to ∼81 % (GAN), we also see this behavior in the other two metrics : *accuracy* increases from ∼66 % to ∼81 % and the *F1 score* increases from ∼51 % to ∼81 %. This demonstrates our initial hypothesis, indicating that a model that can resolve the limitations of the data was necessary. For that case, a GAN model against data with disparity mentioned and discussed in detail in the previous chapter

has optimal performance. Finally, we highlight that our GAN is a stable model against the *Place Pulse 2.0* data set whose nature is images of streets from different cities. As a final tool, a web system was presented with which it is possible to carry out easy interaction and concise visualization of the results obtained by each model, such as the results of each evaluation carried out in the cross-validations (reporting the 3 metrics used). As future work, we plan to extend the work by adding more data and increasing the resolution of the GAN. This will serve to identify more specific characteristics of each place.

Bibliography

Abadi, M., Agarwal, A., et al. (2015). TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org.

Abu-Mostafa, Y. S., Magdon-Ismail, M., et al. (2012). *Learning From Data*. AMLBook.

Acosta, S. y Camargo, J. (2018a). wmodi. http://wmodi.com/. [Último acceso: 11-Agosto-2022].

Acosta, S. F. y Camargo, J. E. (2018b). Predicting city safety perception based on visual image content. In *Iberoamerican Congress on Pattern Recognition*, pages 177–185. Springer.

Adebayo, J., Gilmer, J., et al. (2018). Sanity checks for saliency maps.

Ali, N. y Zafar, B. (2018). 15-scene image dataset. figshare. dataset. https://doi.org/10.6084/m9.figshare.7007177.v1.

Alzate, J. R., Tabares, M. S., et al. (2021). Graffiti and government in smart cities: a deep learning approach applied to medellín city, colombia. In *International Conference on Data Science, E-learning and Information Systems 2021*, pages 160–165.

Ancona, M., Ceolini, E., et al. (2017). A unified view of gradient-based attribution methods for deep neural networks. ETH Zurich.

Andersson, V. O., Birck, M. A., et al. (2017). Investigating crime rate prediction using street-level images and siamese convolutional neural networks. In *Latin American Workshop on Computational Neuroscience*, pages 81–93. Springer.

ArcGis (1999). Arcgis. https://www.arcgis.com/index.html. [Último acceso: 11-Agosto-2022].

Arietta, S. M., Efros, A. A., et al. (2014). City forensics: Using visual elements to predict non-visual city attributes. *IEEE transactions on visualization and computer graphics*, 20(12):2624–2633.

Bay, H., Tuytelaars, T., et al. (2006). Surf: Speeded up robust features. In *European conference on computer vision*, pages 404–417. Springer.

Boser, B. E., Guyon, I. M., et al. (1992). A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 144–152. ACM.

Broomhead, D. y Lowe, D. (1988). Multivariable functional interpolation and adaptive networks, complex systems, vol. 2.

Cenggoro, T. W. et al. (2018). Deep learning for imbalance data classification using class expert generative adversarial network. *Procedia Computer Science*, 135:60–67.

Chakravarti, R. y Meng, X. (2009). A study of color histogram based image retrieval. pages 1323 – 1328.

Charalampos, P., Panagiotis, K., et al. (2019). Storm graffiti/tagging detection dataset. https://doi.org/10.5281/zenodo.3238357.

Chen, L.-C., Papandreou, G., et al. (2017). Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848.

Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1251–1258.

CrimeMapping (2012). Crime mapping website. https://www.crimemapping.com/map. [Último acceso: 11-Agosto-2022].

CrimeReports (2013). Crime report website. https://www.crimereports.com/. [Último acceso: 11-Agosto-2022].

CS231n (2022). Stanford cs231n. http://cs231n.stanford.edu/schedule.html. [Último acceso: 11-Agosto-2022].

CycloMedia (1980). Street smart api. https://www.cyclomedia.com/en/urban-road-safety-index. [Último acceso: 11-Agosto-2022].

Dalal, N. y Triggs, B. (2005). Histograms of oriented gradients for human detection. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, volume 1, pages 886–893. IEEE.

Deng, J., Dong, W., et al. (2009). ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*.

Diniz, A. M. A. y Stafford, M. C. (2021). Graffiti and crime in belo horizonte, brazil: The broken promises of broken windows theory. *Applied Geography*, 131:102459.

Doersch, C., Singh, S., et al. (2012). What makes paris look like paris?

Donahue, J., Jia, Y., et al. (2014). Decaf: A deep convolutional activation feature for generic visual recognition. In *International conference on machine learning*, pages 647–655.

Dubey, A., Naik, N., et al. (2016). Deep learning the city : Quantifying urban perception at A global scale. *CoRR*, abs/1608.01769.

EuroStat (2016). Crime statistics. https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Crime_statistics. [Último acceso: 11-Agosto-2022].

Felipe Moreno-Vera, Bahram Lavi, J. P. (2021a). Quantifying urban safety perception on street view images. In *International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*.

Felipe Moreno-Vera, Bahram Lavi, J. P. (2021b). Urban perception: Can we understand why a street is safe? In *Mexican International Conference on Artificial Intelligence (MICAI)*.

Felzenszwalb, P., McAllester, D., et al. (2008). A discriminatively trained, multiscale, deformable part model. In *2008 IEEE conference on computer vision and pattern recognition*, pages 1–8. IEEE.

Fisher, B.S.; Nasar, J. (1992). *Fear of crime in relation to three exterior site features prospect, refuge, and escape.*, volume 24, pages 35–65.

Fu, K., Chen, Z., et al. (2018). StreetNet: preference learning with convolutional neural network on urban crime perception. In *Proceedings of the 26th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 269–278. ACM.

Garay, A. M., Hashimoto, E. M., et al. (2011). On estimation and influence diagnostics for zero-inflated negative binomial regression models. *Computational Statistics & Data Analysis*, 55(3):1304–1318.

Girshick, R. (2015). Fast R-CNN. *2015 IEEE International Conference on Computer Vision (ICCV)*.

Girshick, R., Donahue, J., et al. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. *2014 IEEE Conference on Computer Vision and Pattern Recognition*.

Glaeser, E. L., Kominers, S. D., et al. (2018). Big data and big cities: The promises and limitations of improved measures of urban life. *Economic Inquiry*, 56(1):114–137.

Gong, Y., Lazebnik, S., et al. (2012). Iterative quantization: A procrustean approach to learning binary codes for large-scale image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(12):2916–2929.

Goodfellow, I. J., Bengio, Y., et al. (2016). *Deep Learning*. MIT Press, Cambridge, MA, USA. http://www.deeplearningbook.org.

Google-Developers (2020). Ml practicum. https://developers.google.com/machine-learning/practica/image-classification?hl=es_419. [Último acceso: 11-Agosto-2022].

Google-Motorolla (2019). Crime map. https://www.crimereports.com/home/. [Último acceso: 11-Agosto-2022].

Har-Peled, S., Roth, D., et al. (2003). Constraint classification for multiclass classification and ranking. In *Advances in neural information processing systems*, pages 809–816.

He, K., Gkioxari, G., et al. (2017). Mask R-CNN. *2017 IEEE International Conference on Computer Vision (ICCV)*.

He, K., Zhang, X., et al. (2015). Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

Herbrich, R., Minka, T., et al. (2007). Trueskill(tm): A bayesian skill rating system. In *Advances in Neural Information Processing Systems 20*, pages 569–576. MIT Press.

Hoiem, D., Efros, A. A., et al. (2007). Recovering surface layout from an image. *International Journal of Computer Vision*, 75(1):151–172.

Hu, Y.-T., Huang, J.-B., et al. (2017). Maskrnn: Instance level video object segmentation. In *Advances in Neural Information Processing Systems*, pages 325–334.

Joachims, T. (2002). Optimizing search engines using clickthrough data. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 133–142. ACM.

Jurie, F. y Triggs, B. (2005). Creating efficient codebooks for visual recognition. In *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, volume 1, pages 604–610. IEEE.

Karras, T., Aittala, M., et al. (2020). Training generative adversarial networks with limited data. *arXiv preprint arXiv:2006.06676*.

Keizer, K., Lindenberg, S., et al. (2008). The spreading of disorder. *Science (New York, N.Y.)*, 322:1681–5.

Koch, G., Zemel, R., et al. (2015). Siamese neural networks for one-shot image recognition. In *ICML deep learning workshop*, volume 2.

Krizhevsky, A., Sutskever, I., et al. (2012). Imagenet classification with deep convolutional neural networks. In Pereira, F., Burges, C. J. C., et al., editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc.

Lazebnik, S., Schmid, C., et al. (2006). Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 2169–2178. IEEE.

Li, X., Zhang, C., et al. (2015a). Does the visibility of greenery increase perceived safety in urban areas? evidence from the place pulse 1.0 dataset. *ISPRS International Journal of Geo-Information*, 4(3):1166–1183.

Li, X., Zhang, C., et al. (2015b). Assessing street-level urban greenery using google street view and a modified green view index. *Urban Forestry & Urban Greening*, 14(3):675–685.

Lin, M., Chen, Q., et al. (2013). Network in network. *arXiv preprint arXiv:1312.4400*.

Lin, T.-Y., Maire, M., et al. (2014). Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.

Lindal, P. J. y Hartig, T. (2013). Architectural variation, building height, and the restorative quality of urban residential streetscapes. *Journal of Environmental Psychology*, 33:26–36.

Liu, W., Anguelov, D., et al. (2016). SSD: single shot multibox detector. *European Conference on Computer Vision (ECCV)*, abs/1512.02325.

Liu, X., Chen, Q., et al. (2017). Place-centric visual urban perception with deep multi-instance regression. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 19–27. ACM.

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110.

Lynch, K. (1984). Reconsidering the image of the city. In *Cities of the Mind*, pages 151–161. Springer.

Manjunath, B. S., Ohm, J.-R., et al. (2001). Color and texture descriptors. *IEEE Transactions on circuits and systems for video technology*, 11(6):703–715.

Martin, D., Fowlkes, C., et al. (2001). A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001*, volume 2, pages 416–423. IEEE.

Massachusets-Office-Goverment (2005). Massgis data-land use 2005. https://www.mass.gov/. [Último acceso: 11-Agosto-2022].

Matas, J., Chum, O., et al. (2004). Robust wide-baseline stereo from maximally stable extremal regions. *Image and vision computing*, 22(10):761–767.

MatLab-Developers (2020). Mathworks. https://www.mathworks.com/discovery/object-detection.html. [Último acceso: 11-Agosto-2022].

Mawby, R. (2014). *Crime and Disorder, Security and the Tourism Industry*, pages 383–403. Springer.

Min, W., Mei, S., et al. (2019). Multi-task deep relative attribute learning for visual urban perception. *IEEE Transactions on Image Processing*, 29:657–669.

MIT-Media-Lab (2013). Place pulse. http://pulse.media.mit.edu/data/. [Último acceso: 11-Agosto-2022].

MIT-Media-Lab (2014). Streetscore. http://streetscore.media.mit.edu/. [Último acceso: 11-Agosto-2022].

MIT-Media-Lab (2015). Treepedia. http://senseable.mit.edu/treepedia. [Último acceso: 11-Agosto-2022].

Mohammed, A.-M. y Sookram, S. (2015). The impact of crime on tourist arrivals—a comparative analysis of jamaica and trinidad and tobago. *Social and Economic Studies*, 64(2):153–176.

Molnar, C. (2022). *Interpretable Machine Learning*. 2 edition.

Moreno-Vera, F. (2021). Understanding safety based on urban perception. In *International Conference on Intelligent Computing*, pages 54–64. Springer.

Naik, N., Philipoom, J., et al. (2014). StreetScore: predicting the perceived safety of one million streetscapes. *2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops*.

Nasar, J., Fisher, B., et al. (1993). *Proximate physical cues to fear of crime.*, volume 26, pages 161–178.

Nasar, J. L. (1998). The evaluative image of the city.

Noh, H., Hong, S., et al. (2015). Learning deconvolution network for semantic segmentation. In *Proceedings of the IEEE international conference on computer vision*, pages 1520–1528.

Novak, C. L., Shafer, S. A., et al. (1992). Anatomy of a color histogram. In *CVPR*, volume 92, pages 599–605.

Numbeo (2019). World database of crime index. https://www.numbeo.com/crime/rankings_by_country.jsp?title=2019. [Último acceso: 11-Agosto-2022].

Ojala, T., Pietikainen, M., et al. (2002). Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on pattern analysis and machine intelligence*, 24(7):971–987.

Oliva, A. y Torralba, A. (2001). Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42:145–175.

Ordonez, V. y Berg, T. L. (2014). Learning high-level judgments of urban perception. *European Conference on Computer Vision (ECCV)*.

Otsu, N. (1975). *A threshold selection method from gray-level histograms.*, volume 11, pages 23–27.

Park, D. K., Jeon, Y. S., et al. (2000). Efficient use of local edge histogram descriptor. In *Proceedings of the 2000 ACM workshops on Multimedia*, pages 51–54.

Parzen, E. (1962). On estimation of a probability density function and mode. *The annals of mathematical statistics*, 33(3):1065–1076.

Pedregosa, F., Varoquaux, G., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Perronnin, F., Sánchez, J., et al. (2010). Improving the fisher kernel for large-scale image classification. In *European conference on computer vision*, pages 143–156. Springer.

Pinheiro, P. O., Collobert, R., et al. (2015). Learning to segment object candidates. In *Advances in Neural Information Processing Systems*, pages 1990–1998.

Porzi, L., Rota Bulò, S., et al. (2015). Predicting and understanding urban perception with convolutional neural networks.

Quercia, D., O'Hare, N. K., et al. (2014). Aesthetic capital: what makes london look beautiful, quiet, and happy? In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, pages 945–955. ACM.

Rao, A., Srihari, R. K., et al. (1999). Geometric histogram: A distribution of geometric configurations of color subsets. In *Internet Imaging*, volume 3964, pages 91–101. International Society for Optics and Photonics.

Ray, S. y Page, D. (2001). Multiple instance regression. In *ICML*, volume 1, pages 425–432.

Redmon, J., Divvala, S., et al. (2016). You only look once: Unified, real-time object detection. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Ren, S., He, K., et al. (2017). Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6):1137–1149.

Ronneberger, O., Fischer, P., et al. (2015). U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer.

Russakovsky, O., Deng, J., et al. (2015). ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252.

Salesses, M. P. (2012). *Place Pulse: Measuring the collaborative image of the city*. PhD thesis, Massachusetts Institute of Technology.

Salesses, P., Schechtner, K., et al. (2013). The collaborative image of the city: Mapping the inequality of urban perception. *PLOS ONE*.

Salimans, T., Goodfellow, I., et al. (2016). Improved techniques for training gans. *Advances in neural information processing systems*, 29:2234–2242.

Sampath, V., Maurtua, I., et al. (2021). A survey on generative adversarial networks for imbalance problems in computer vision tasks. *Journal of big Data*, 8(1):1–59.

Sampson, R. J., Morenoff, J. D., et al. (2002). Assessing "neighborhood effects": Social processes and new directions in research. *Annual review of sociology*, 28(1):443–478.

Schroeder, H. W. y Anderson, L. M. (1984). Perception of personal safety in urban recreation sites. *Journal of leisure research*, 16(2):178–194.

Selvaraju, R. R., Cogswell, M., et al. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 618–626.

Seresinhe, C. I., Preis, T., et al. (2017). Using deep learning to quantify the beauty of outdoor places. *Royal Society open science*, 4(7):170170.

Shrikumar, A., Greenside, P., et al. (2016). Not just a black box: Learning important features through propagating activation differences.

Simonyan, K., Vedaldi, A., et al. (2013). Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*.

Simonyan, K. y Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *International Conference on Learning Representations (ICLR)*.

Sivic, J. y Zisserman, A. (2004). Video data mining using configurations of viewpoint invariant regions. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, volume 1, pages I–I. IEEE.

Skogan, W. G. (1992). *Disorder and decline: Crime and the spiral of decay in American neighborhoods*. Univ of California Press.

Smilkov, D., Thorat, N., et al. (2017). Smoothgrad: removing noise by adding noise.

Smola, A. y Schölkopf, B. (2004). A tutorial on support vector regression, statist. *Comput*, 14:199–222.

Springenberg, J. T., Dosovitskiy, A., et al. (2014). Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806*.

Stalidis, P., Semertzidis, T., et al. (2018). Examining deep learning architectures for crime classification and prediction. *arXiv*.

Sundararajan, M., Taly, A., et al. (2017). Axiomatic attribution for deep networks.

Szegedy, C., Liu, W., et al. (2014). Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9.

Szegedy, C., Vanhoucke, V., et al. (2015). Rethinking the inception architecture for computer vision. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Tencent-Street-View-service (2016). Map qq. https://map.qq.com/. [Último acceso: 11-Agosto-2022].

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

Tikhonov, A. N. (1943). On the stability of inverse problems. In *Dokl. Akad. Nauk SSSR*, volume 39, pages 195–198.

Tokuda, E. K., Silva, C. T., et al. (2019). Quantifying the presence of graffiti in urban environments. *CoRR*, abs/1904.04336.

Tranmer, M. Multiple linear regression.

UK-gov (2015). Geograph. http://www.geograph.org.uk/. [Último acceso: 11-Agosto-2022].

UK-gov (2017). Scenic-or-not. http://scenicornot.datasciencelab.co.uk/. [Último acceso: 11-Agosto-2022].

Ulrich, R. S. (1979). Visual landscapes and psychological well-being. *Landscape research*, 4(1):17–23.

UrbanForest (2014). Urban forests map website. http://urbanforestmap.org/. [Último acceso: 25-Agosto-2019].

UrbanGems (2014). Urbangems. http://urbangems.org/. [Último acceso: 14-Octubre-2019].

USA, D. o. J. (2012). Mapping crime: Principle and practice. https://www.ncjrs.gov/pdffiles1/nij/178919.pdf. [Último acceso: 11-Agosto-2022].

Viso-AI (2020). Object segmentation. https://viso.ai/deep-learning/image-segmentation-using-deep-learning. [Último acceso: 11-Agosto-2022].

von Platen, P., Tao, F., et al. (2020). Multi-task siamese neural network for improving replay attack detection. *arXiv preprint arXiv:2002.07629*.

Wikipedia. Linear model. https://es.wikipedia.org/wiki/Modelo_lineal. [Último acceso: 11-Agosto-2022].

Wikipedia. Supervised learning. https://en.wikipedia.org/wiki/Supervised_learning. [Último acceso: 11-Agosto-2022].

Wikipedia (2020). Non-linear models. https://es.wikipedia.org/wiki/Regresi%C3%B3n_no_lineal. [Último acceso: 11-Agosto-2022].

Wilson, J. Q. y Kelling, G. L. (1982). Broken windows. *Atlantic monthly*, 249(3):29–38.

Xiao, J., Hays, J., et al. (2010). Sun database: Large-scale scene recognition from abbey to zoo. pages 3485–3492.

Xu, Y., Yang, Q., et al. (2019). Visual urban perception with deep semantic-aware network. In *International Conference on Multimedia Modeling*, pages 28–40. Springer.

Y., L., Boser, B., et al. (1990). Handwritten digit recognition with a back-propagation network. In *Advances in Neural Information Processing Systems*, pages 396–404. Morgan Kaufmann.

Yang, J., Zhao, L., et al. (2009). Can you see green? assessing the visibility of urban forests in cities. *Landscape and Urban Planning*, 91(2):97–104.

Yu, F. y Koltun, V. (2015). Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*.

Zeiler, M. D. y Fergus, R. (2013a). Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer.

Zeiler, M. D. y Fergus, R. (2013b). Visualizing and understanding convolutional networks.

Zhang, F., Zhou, B., et al. (2018). Measuring human perceptions of a large-scale urban region using machine learning. *Landscape and Urban Planning*, 180:148–160.

Zhao, H., Shi, J., et al. (2017). Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890.

Zhou, B., Khosla, A., et al. (2016a). Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929.

Zhou, B., Khosla, A., et al. (2016b). Places: An image database for deep scene understanding. *arXiv preprint arXiv:1610.02055*.

Zhou, B., Lapedriza, A., et al. (2016c). Places365 results. https://github.com/CSAILVision/places365/. [Último acceso: 11-Agosto-2022].

Zhou, B., Lapedriza, A., et al. (2017). Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1452–1464.

Zhou, B., Lapedriza, A., et al. (2014). Learning deep features for scene recognition using places database. In Ghahramani, Z., Welling, M., et al., editors, *Advances in Neural Information Processing Systems 27*, pages 487–495. Curran Associates, Inc.

Zhou, T., Liu, W., et al. (2018). Gan-based semi-supervised for imbalanced data classification. In *2018 4th International Conference on Information Management (ICIM)*, pages 17–21. IEEE.

Zou, H. y Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)*, 67(2):301–320.