UNIVERSITÀ POLITECNICA DELLE MARCHE
CORSO DI DOTTORATO DI RICERCA IN INGEGNERIA DELL'INFORMAZIONE
CURRICULUM IN INGEGNERIA INFORMATICA, GESTIONALE E DELL'AUTOMAZIONE

Deep Learning based models for Space Understanding

Ph.D. Dissertation of:
Massimo Martini

Advisor:
Prof. Primo Zingaretti

Curriculum Supervisor:
Prof. Franco Chiaraluce

XX edition - new series

UNIVERSITÀ POLITECNICA DELLE MARCHE
CORSO DI DOTTORATO DI RICERCA IN INGEGNERIA DELL'INFORMAZIONE
CURRICULUM IN INGEGNERIA INFORMATICA, GESTIONALE E DELL'AUTOMAZIONE

Deep Learning based models for Space Understanding

Ph.D. Dissertation of:
Massimo Martini

Advisor:
Prof. Primo Zingaretti

Curriculum Supervisor:
Prof. Franco Chiaraluce

XX edition - new series

UNIVERSITÀ POLITECNICA DELLE MARCHE
CORSO DI DOTTORATO DI RICERCA IN INGEGNERIA DELL'INFORMAZIONE
FACOLTÀ DI INGEGNERIA
Via Brecce Bianche – 60131 Ancona (AN), Italy

Ringraziamenti

Ho sempre pensato a questa sezione come a un'opportunità per fare mente locale e ricapitolare tutto ciò che mi è successo durante questi 3 anni di Dottorato. Così in futuro, rileggendo questi commenti, potrò rivivere tutte le esperienze passate.

Il primo doveroso ringraziamento và alla mia famiglia, che mi ha sempre supportato e sopportato fin dall'inizio di questo percorso, nonostante sia partito affrontando un lutto in famiglia, seguito da infortuni e problematiche varie (giusto per non farci mancare nulla).

Un secondo ringraziamento và ai primi compagni che ho incontrato in questo percorso. Andrea, la prima persona con cui ho collaborato, sempre disponibile ed in gamba. Luca e Marco, compagni di sventura incontrati, per caso e per fortuna, nel cuore del centro di Ancona (tra due salite che mi rimarranno per sempre impresse nella mente...). Indimenticabili le storie e le risate che ci hanno accompagnati nel corso di questi anni. Come non ricordare le nottate passate a programmare insieme a Marco fino alle 3 di notte, con annessa pausa panino e birra a mezzanotte, nel pub sottocasa. Successivamente, coinquilini per un anno e con mille idee per la testa ma, purtroppo, ostacolati da una pandemia globale... (o per fortuna, altrimenti saremmo morti di reumatismi dovuti all'umidità di quella casa).

Un ringramento speciale và ovviamente al mio intero gruppo di ricerca, il VRAI, con cui ho condiviso questi anni di Dottorato, progetti, hackathon (ricordo quello di Rimini, in cui fummo gli ultimi ad arrivare ma i primi a mangiare), momenti felici e non. Ad un certo punto, pensare di lasciare il Dottorato è stato uno di questi. Ma forse, voler lasciare il Dottorato è esso stesso il vero significato del Dottorato (semicit.). Ringrazio i miei professori, che mi hanno concesso molteplici opportunità professionali, permettendomi di crescere sotto diversi aspetti. Un ringraziamento particolare và al mio tutor, il prof. Primo Zingaretti, per la pazienza dimostrata nei miei confronti.

Ovviamente, questi anni sono stati alleggeriti da tutti i momenti felici passati insieme a i miei amici: partendo da quelli storici, poi quelli universitari, fino ad arrivare a quelli della squadra FC Muntijorgio (seguiteli su instagram).

Un ringraziamento speciale và alla mia fidanzata Irene, con cui ho passato e sto passando momenti stupendi, che mi ha sempre supportato fin da quando è entrata nella mia vita. La ringrazio per avermi dato l'opportunità di entrare anche nella sua famiglia, che mi ha accolto fin da subito in maniera calorosa.

Una nota di riguardo và anche alla prima regola dell'Università (cit. Tesi Magistrale), sempre attuale, in ogni circostanza.

Infine, concluderò rimarcando i ringraziamenti principali attraverso il linguaggio che è stato l'emblema di tutto il mio percorso di Dottorato:

```
family_list = ["Stefano", "Floriano", "Rosanna", \
               "Elena", "Sandro", "Beatrice", "Yuri"]
friends_list = [...]
vrai_list = [...]

for p in ["Irene", family_list, friends_list, vrai_list]:
    print("Thanks to", p)
```

Ancona, March 2022

Massimo Martini

Abstract

The understanding and description of a scene, or more generally of a space, has always been of great interest to the scientific community, as it is widely used in various fields. Applications include: 3D reconstruction of monuments in the field of cultural heritage, localisation for self-driving robots, video surveillance systems, consumer monitoring systems in the field of retail.

In recent years, these approaches have been improved by systems based on Artificial Intelligence, and specifically by using Machine Learning and Deep Learning methods. These methods are becoming very successful thanks to the growing amount of data made available by the scientific community, the development of innovative techniques, and above all, thanks to the improvement of hardware systems, which have become increasingly accessible and less expensive.

Part of the work of this thesis has been developed in collaboration with Politecnico of Torino, Bruno Kessler foundation of Trento and Grottini Lab company. The aim of this research is to improve the state of the art regarding all the aspects for Space Undertanding, both static and dynamic.

The objectives and the contributions of this thesis reflect the research activities performed in (i) the Point Cloud Semantic Segmentation of a space (ii) the Change Detection on a dynamic scene, and (iii) the Person Re-Identification on realistic environments.

In the first topic, a new deep learning method for semantic segmentation of point clouds is described. The proposed approach uses additional discriminative features, compared to the state of the art, to improve its learning and better discriminate the various classes of a space. Subsequently, the approach has been improved by adding handcrafted features, produced by experts in the field, and providing training enhancements. Extensive experiments were carried out both in cultural heritage field and indoor scenes. Finally, a framework based on generative approaches is proposed as a data augmentation technique, in order to improve the performance on unbalanced classes. The reliability of the proposed methods was evaluated on a novel dataset acquired by experts in the field, and the results obtained showed that the effectiveness of the methods outperformed the state of the art. In fact, the proposed DGCNN-Mod increases the accuracy on the two test scenes of ArCH dataset by 26.86% and 4.37%, compared to the popular DGCNN. When handcrafted features are also added, it achieves an increase of 28.44% and 6.21% on the same scenes.

In the second task, A mixed methodology between machine learning and deep learn-

ing approaches is proposed for the Change Detection task. The presented method exploits the extraction of visual and textual features, coming from the acquisition of RGB images of a retail environment. The union of these features allows to train a final classifier that will give an overall result about the state of the space. The reliability of the proposed methods was investigated using a novel dataset validated by experts in the field and by studying the behaviour of consumers. The results are interesting and the proposed approach can be considered a good starting point to be improved in its various parts.

In the third topic, it is presented an algorithm for person re-identification using rgb-d video, with a top-view configuration, is presented. The proposed approach is designed to work both in a closed world and in a more realistic open world environment. A further improvement is made by integrating all the features made available by the data: visual, spatial and temporal. The proposed method has been validated by acquiring two new datasets of RGB-D video, validated by experts in the field, for both a retail and a museum environment. The results obtained showed that the effectiveness of the methods outperformed the state of the art approaches. In fact, the proposed TVOW improves the accuracy on the new TVPR2 dataset by 2.72% and the accuracy on the TVPR dataset by 0.45%.

Sommario

La comprensione e la descrizione di una scena, o più in generale di uno spazio, è sempre stata di grande interesse per la comunità scientifica, poiché è ampiamente utilizzata in vari campi. Le applicazioni includono: Ricostruzione 3D di monumenti nel campo dei beni culturali, localizzazione per robot a guida autonoma, sistemi di videosorveglianza, sistemi di monitoraggio dei consumatori nel campo del Retail.

Negli ultimi anni, questi approcci sono stati migliorati da sistemi basati sull'intelligenza artificiale, e in particolare utilizzando metodi di Machine Learning e Deep Learning. Questi metodi stanno avendo molto successo grazie alla crescente quantità di dati messi a disposizione dalla comunità scientifica, allo sviluppo di tecniche innovative e, soprattutto, grazie al miglioramento dei sistemi hardware, che sono diventati sempre più accessibili e meno costosi.

Parte del lavoro di questa tesi è stato sviluppato in collaborazione con il Politecnico di Torino, la fondazione Bruno Kessler di Trento e la società Grottini Lab. Lo scopo di questa ricerca è quello di migliorare lo stato dell'arte per quanto riguarda tutti gli aspetti dello Space Undertanding, sia statici che dinamici.

Gli obiettivi e i contributi di questa tesi riflettono le attività di ricerca svolte in (i) metodi di Point Cloud Semantic Segmentation di uno spazio (ii) sistemi di Change Detection su una scena dinamica, e (iii) approcci di Person Re-Identification su ambienti realistici.

Nel primo tema di ricerca, viene descritto un nuovo metodo di deep learning per la segmentazione semantica delle nuvole di punti. L'approccio proposto utilizza delle feature aggiuntive, rispetto allo stato dell'arte, per migliorare il suo apprendimento e discriminare meglio le varie classi di uno spazio. Successivamente, l'approccio è stato migliorato aggiungendo feature handcrafted, prodotte da esperti del settore, e permettendo un miglioramento nell'addestramento. Sono stati condotti esperimenti approfonditi sia nel dominio dei beni culturali che nelle scene indoor. Infine, viene proposto un framework basato su approcci generativi, come tecnica di data augmentation, al fine di migliorare le prestazioni sulle classi sbilanciate. L'affidabilità dei metodi proposti è stata valutata su un nuovo dataset acquisito da esperti del settore, e i risultati ottenuti hanno mostrato che l'efficacia dei metodi supera lo stato dell'arte. Infatti, la DGCNN-Mod proposta aumenta di 26,86% e 4,37% l'accuratezza sulle due scene di test di ArCH dataset, rispetto alla popolare DGCNN. Aggiungendo anche le feature handcrafted, raggiunge un aumento del 28,44% e 6,21% sulle stesse scene.

Nella seconda tematica, viene proposta una metodologia mista tra approcci di ma-

chine learning e deep learning, per il compito di Change Detection. Il metodo presentato sfrutta l'estrazione di feature visive e testuali, provenienti dall'acquisizione di immagini RGB dall'ambito del Retail. L'unione di queste caratteristiche permette di addestrare un classificatore finale che darà un risultato complessivo sullo stato dello spazio. L'affidabilità dei metodi proposti è stata studiata utilizzando un nuovo dataset validato da esperti del settore, e studiando il comportamento dei consumatori. I risultati sono interessanti e l'approccio proposto può essere considerato un buon punto di partenza ed essere migliorato nelle sue varie parti.

Nel terzo tema di ricerca, viene presentato un algoritmo per la re-identificazione delle persone utilizzando video rgb-d, con una configurazione top-view. L'approccio proposto è progettato per funzionare sia in ambiente chiuso che in un ambiente aperto, quindi più realistico. Un ulteriore miglioramento viene fatto integrando tutte le caratteristiche rese disponibili dai dati: visive, spaziali e temporali. Il metodo proposto è stato validato acquisendo due nuovi dataset di video RGB-D, convalidati da esperti del settore, sia nell'ambito del Retail che per un ambiente museale. I risultati ottenuti hanno mostrato che l'efficacia dei metodi supera gli approcci allo stato dell'arte. Infatti, la TVOW proposta migliora del 2,72% l'accuratezza sul nuovo dataset TVPR2 e del 0,45% l'accuratezza sul dataset TVPR.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Context

The understanding and description of a scene, or more generally of a space, has always been of great interest to the scientific community, as it is widely used in various fields. First of all, it is necessary to define what the concept of Space means. In its simplest definition it can be described as a set of entities enclosed within predefined limits. The entities themselves must be located within it by using a localisation system, and above all must always be identified in some way. All these concepts make it evident that the global task of Space Understanding is full of challenges.

Computer Vision based approaches were the first to be used to solve this task. The simplest is classification, which makes it possible to associate a particular class to an entity [1, 2, 3]. But its main drawback is that it requires a phase to divide the space into all these entities. The second approach is object detection, which provides the location of the processed entity by using a bounding box [4, 5]. The most interesting approach for the comprehension of a scene is instead semantic segmentation, which allows to divide the entire space into various types of classes, at a point level, thus giving the exact positions and shapes of all the entities in the space [6, 7].

In recent years, these approaches have been improved by systems based on Artificial Intelligence, and specifically by using Machine Learning and Deep Learning methods. These methods are becoming very successful thanks to the growing amount of data made available by the scientific community, the development of innovative techniques, and above all, thanks to the improvement of hardware systems, which have become increasingly accessible and less expensive.

The understanding of space has gained much interest in various fields of application. For example, it is very useful for the definition or reconstruction of 3D models in the field of cultural heritage. It can be used for the reconstruction or maintenance of churches and monuments [8, 9, 10, 11]. It is also very popular in the field of robotics, to implement robots with autonomous driving systems [12, 13]. Indeed, it is important that the robot always locates itself within an environment, or recognises certain obstacles while in motion.

The choice of technique to be used obviously depends on the types of data available.

At the state of the art, semantic segmentation is often performed using RGB images, which only provide the visual characteristics for understanding a space. A more innovative type of data concerns the acquisition of RGB-D images, which, thanks to the Depth channel, it also allow to obtain information on the depth and distance between entities. Lately, recent approaches are relying on point clouds [14], which are easy to acquire and efficient in locating any object within a three-dimensional space. This thesis will focus on the description of new deep learning approaches for the Semantic Segmentation of Point Clouds, as it is the best method for the understanding of a static space. In particular, the proposed methods improve the state of the art by using more discriminating features for the recognition of object classes.

Another interesting challenge concerns the tracking of the temporal evolution of a space, an environment or the entities within them. At the state of the art there are several approaches concerning video surveillance or study of consumer behaviour, in the Retail domain. The temporal evolution of parts of the space can be carried out through Change Detection techniques [15, 16, 17, 18, 19]. Also in this case, the type of approach to be used is strictly related to the acquired data. Image classification is an excellent compromise to solve this task, since change detection systems have to save in memory large amounts of data, and maintain them over a long period of time. In this thesis, a new change detection approach based on images will be proposed. The structure of the framework will be based on a set of Machine Learning and Deep Learning methods. Compared to the state of the art, it will allow both visual and textual features to be exploited from the acquired RGB images.

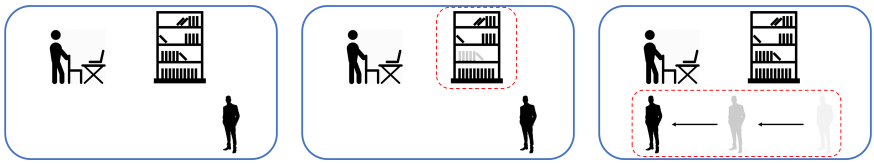
Finally, the temporal evolution of a space is also strictly related to the entities within it. Their localisation, in a short period of time, can be solved through Re-Identification techniques. At the state of the art, several methodologies exist for the re-identification of different types of entities: cars, indoor objects, animals, etc. [20, 21]. The most common and dynamic entity that can be found within a space are people. In fact, the scientific community has published several approaches on Person Re-Identification. There are approaches with frontal configuration and approaches with top-view configuration. A lot of methods use RGB images, other thermal or infrared images; hybrid approaches also exist, which integrates diffent type of data [22]. In this thesis, a new Person Re-Identification approach, based on RGB-D images with a top-view configuration, will be proposed. It will be a temporal and multimodal method based on deep learning.

1.2 Objectives and main contributions

The main objective of this thesis is to design and implement AI algorithms to improve the Space Understanding concept, in all its fundamental aspects. As a first definition, space can be described as a static concept, defined simply by its spatial limits and all that it contains within it. In this definition, the first things to be understood are

therefore the entities that can be found within it, their spatial location and, above all, their type. Once these concepts are defined, the second step is to study the temporal evolution of space. First of all, it is necessary to understand how parts of an environment can change during a long period of time. And above all, it is necessary to keep track of these changes. Finally, it is also necessary to understand how entities within a space can evolve in the short term period: this concept can include both the interaction between entities and space and between the identities themselves.

All these fundamental aspects concerning the understanding of space are described in Figure 1.1.



(a) A static space, described by its boundary and entities.

(b) Temporal evolution of a part of a dynamic space.

(c) Temporal evolution of an entity of a dynamic space.

Figure 1.1: All the concepts regarding the Space Understanding task.

The second objective of this thesis is to correctly translate these aspects into the corresponding Computer Vision tasks, and study their potential applications in various fields. In the literature, the first aspect concerning the understanding of space as a static concept is often developed with semantic segmentation approaches. In this way, it is possible to define both the spatial boundaries of an environment and, at the same time, recognise the position and type of all objects within them. The second aspect, relating to the long-term temporal evolution of the various parts of a space, is addressed in the literature as a Change Detection task. Finally, the study of the temporal evolution of the entities that move within an environment is often defined through tracking or re-identification techniques. Since the most common and dynamic entities that can be found in a space are people, the author translated this aspect into a Person Re-Identification task.

The third objective of the thesis concerns the choice of the type of data to be used for the understanding of all these sub-problems. The thesis will describe the advantages and disadvantages of various data types, starting with the easiest to acquire, i.e. RGB images, up to its natural evolution related to RGB-D images and finally point cloud acquisitions. The choice of the type of data to be used depends on several factors, including the ease of acquisition, the amount of useful information that can be used, and the speed of processing this data.

All these problems are based on processes that are time consuming and sometimes require manual procedure. To make AI techniques tailored for the aforementioned

challenging applications, considerations such as computational complexity reduction, hardware implementation and software optimization have been outlined. The overarching goal of this work consists of presenting a pipeline to select the model that best fits with the given observations; nevertheless, it does not prioritize in memory and time complexity when matching models to observations. Beside this, it is well known that, to conduct a comprehensive performance evaluation, it is critical to have meaningful datasets. Overcoming this limitation, this thesis adds to the body of knowledge by collecting and sharing representative labelled datasets. While much progress has been made in recent years regarding efforts in sharing codes and datasets, it is of great importance to develop libraries and benchmarks to gauge state-of-the-art datasets. Newly challenging datasets were specifically collected for the tasks described in this study. In fact, each described application involved the collection of one dataset, which was used as the input. Thus, the learning methods described were evaluated according to the following proposed datasets and their respective application domain: a dataset of point cloud outdoor scenes for Point Cloud Semantic Segmentation in Cultural Heritage domain (Section 3.1.2), an images dataset for Change Detection task in Retail enviroment (Section 3.2.1), an RGB-D video dataset for the Person Re-Identification task in crowded environments (Section 3.3.2). In particular, the techniques and methods for each type of research are analysed, the main paths that most approaches follow are also summarised, and their contributions are indicated. Thereafter, the proposed approaches are categorised and compared from multiple perspectives, including methodologies, functions, and an analysis of the pros and cons of each category.

The questions the author endeavours to answer with this thesis work regarding the Space Understanding tasks are summarised below:

1. How can Computer Vision algorithms be applied to the different subtasks regarding Space Understanding?

2. The integration of data of a different nature could help the understanding of a Space, or the entities contained within it?

3. Is it worth following the trend of using pure deep learning methods respect the machine learning ones? Or could a mixed approach bring improvements?

4. Are the proposed algorithms better than the standard algorithms widely used in the literature?

1.3 Thesis overview

This thesis aims to answer the questions reported above designing and developing machine learning algorithms for the task of Space Understanding.

The thesis is organized into six chapters according to the following organization:

- **Chapter 1** is the introduction to this thesis. It describes the objectives and motivations that led the author to carry out this work.
- **Chapter 2** aims to focus on the state of the art of methods relating to the understanding of a space, all the entities it contains and its temporal evolution. The author outlines the strengths and the drawbacks of the existing achievements, focusing on what questions they answer and what they do not.
- **Chapter 3** presents the materials and the proposed ML algorithms for the data processing/analysis stage. In detail, the methods implemented are organized into algorithms for Point Cloud Semantic Segmentation, Change Detection on RGB images, Person Re-Identification on RGB-D videos.
- **Chapter 4** presents the experimental setup and metrics for assessing the proposed methods. Then, the author provides the experimental results for evaluating the performance of the algorithms.
- **Chapter 5** discusses the obtained results and provides further useful analyses.
- **Chapter 6** outlines the conclusions and future works.

Chapter 2

State of the art

In this chapter, the author gives background information on the concepts introduced in Section 1.1. Then, the state of the art related to the Comprehension of a Space is summarised in Section 2.1. The literature review regarding the Temporal Evolution of a Space is described in Section 2.2. Finally, Section 2.3 addresses the problem of Short-term Evolution of Space's entities and what methods are available to detect them.

2.1 Comprehension of a Space

The understanding and description of a scene, or more generally of a space, has always been of great interest to the scientific community, as it is widely used in various fields. First of all, it is necessary to define what the concept of Space means. In its simplest definition it can be described as a set of entities enclosed within predefined limits. The entities themselves must be located within it by using a localisation system, and above all must always be identified in some way. All these concepts make it evident that the global task of Space Understanding is full of challenges.

Computer Vision based approaches were the first to be used to solve this task. The simplest is classification, which makes it possible to associate a particular class to an entity. The problem of classification has been widely studied in the database, data mining, and information retrieval communities [1, 2, 3]. Its performance is closely related to the type of data to be classified. Unfortunately, in the space understanding problem, its main drawback is that it requires a phase to divide the space into all these entities.

The second approach is object detection, which provides the location of the processed entity by using a bounding box. It has been an active area of research for several decades [4]. The goal of object detection is to determine whether there are any instances of objects from given categories (such as humans, cars, bicycles, dogs or cats) in an image and, if present, to return the spatial location and extent of each object instance (e.g., via a bounding box [23, 24]). Object detection forms the basis for solving complex or high level vision tasks such as segmentation, scene understanding, object tracking, image captioning, event detection, and activity recognition [5].

The most interesting approach for the comprehension of a scene is instead semantic segmentation. Semantic segmentation is one of the most important research methods for computer vision, and regards the task to classify each pixel or point in the scene into classes that have specific features [6, 7]. In the past, semantic segmentation concerned bi-dimensional images but, due to some limitations related to occlusions, illumination, posture and other problems, the researches began to deal with three-dimensional data. This change also occurred thanks to the growing diffusion of photogrammetry and laser scanning surveys. In the 3D form of semantic segmentation, regular or irregular points are processed in the 3D space [14].

With the rapid development of 3D acquisition technologies, 3D sensors are becoming increasingly available and affordable, including various types of 3D scanners, LiDARs, and RGB-D cameras (such as Kinect, RealSense and Apple depth cameras) [25]. 3D data acquired by these sensors can provide rich geometric, shape and scale information [26, 27]. Complemented with 2D images, 3D data provides an opportunity for a better understanding of the surrounding environment for machines. 3D data has numerous applications in different areas, including autonomous driving, robotics, remote sensing, and medical treatment [28]. In computer vision and remote sensing, point clouds can be acquired with four main techniques [29]: 1) Image-derived methods [30]; 2) Light Detection And Ranging (LiDAR) systems [31]; 3) Red Green Blue -Depth (RGB-D) cameras [32]; and 4) Synthetic Aperture Radar (SAR) systems [33]. Due to the differences in survey principles and platforms, their data features and application ranges are very diverse.
3D data can usually be represented with different formats, including depth images, point clouds, meshes, and volumetric grids. As a commonly used format, point cloud representation preserves the original geometric information in 3D space without any discretization. Therefore, it is the preferred representation for many scene understanding related applications such as autonomous driving and robotics. Recently, deep learning techniques have dominated many research areas, such as computer vision, speech recognition, and natural language processing. However, deep learning on 3D point clouds still face several significant challenges [34], such as the small scale of datasets, the high dimensionality and the unstructured nature of 3D point clouds. On this basis, this thesis focuses on the analysis of the state-of-the-art deep learning methods which have been used to process 3D point clouds.
Deep learning on point clouds has been attracting more and more attention, especially in the last years. Several publicly available datasets are also released, such as ModelNet [35], ScanObjectNN [36], ShapeNet [37], PartNet [38], S3DIS [39], ScanNet [32], Semantic3D [40], ApolloCar3D [41], and the KITTI Vision Benchmark Suite [42, 43]. These datasets have further boosted the research of deep learning on 3D point clouds, with an increasingly number of methods being proposed to address various problems related to point cloud processing, including 3D shape classification, 3D object detection and tracking, 3D point cloud segmentation, 3D point cloud registra-

tion, 6-DOF pose estimation, and 3D reconstruction [44, 45, 46].
For all these reasons, a variety of surveys of deep learning on 3D data have been written in recent years to explore these topics in more detail [47, 48, 29, 49, 50].

2.1.1 AI Approaches in the Cultural Heritage domain

Surely, the automatic interpretation of 3D point clouds by semantic segmentation in the cultural heritage (CH) context represents a very challenging task. Digital documentation is not easy to obtain, but it is necessary to disseminate cultural heritage [51]. Shapes are complex and the objects, even if repeatable, are unique, handcrafted and not serialised. Notwithstanding, the understanding of 3D scenes in digital CH is crucial, as it can have many applications such as the identification of similar architectural elements in large dataset, the analysis of the state of conservation of materials, the subdivision of the point clouds in its structural parts preliminary for scan-to-BIM processes, etc. [52].

In the literature, there is a restricted number of applications that use machine learning methods to classify 3D point clouds in different objects belonging to cultural heritage scenes, even if, according to [53], these methods had great progress to this regard. Indeed, in their study the authors explore the applicability of supervised machine learning approaches to cultural heritage by providing a standardised pipeline for several case studies.

In this domain, the research of [54] has two main objectives: providing a framework that extracts geometric primitives from a masonry image, and extracting and selecting statistical features for the automatic clustering of masonry. The authors combine existing image processing and machine learning tools for the image-based classification of masonry walls and then make a performances comparison among five different machine learning algorithms for the classification task. The main issue of this method is that each block of the wall is not individually characterised.

The research presented in [55] wants to overcome this limitation, presenting a novel automatic segmentation algorithm of masonry blocks from a 3D point cloud acquired with LiDAR technology. The image processing algorithm is based on an optimisation of the watershed algorithm, also used to improve segmentation algorithms in other works [56, 57], to automatically segment 3D point clouds in 3D space isolating each single stone block.

In their research, authors of [58] propose a strategy to classify heritage 3D models by applying supervised machine learning classification algorithms to their UV maps. To verify the reliability of the method, the authors evaluate different classifiers over three heterogeneous case studies.

In [59] the authors explore the relation between covariance features and architectural elements using supervised machine learning classifier (Random Forest), finding in particular a correlation between the feature search radii and the size of the element.

A more in-depth analysis of the previous approach [60] demonstrates the capability of the algorithm to generalise across different unseen architectural scenarios. The research conducted by Murtiyoso et al. [61] aims to help the manual point clouds labeling of large training data set required from machine learning algorithms. Moreover, the authors introduce a series of functions that allow the automatic processing for some issues of segmentation and classification of CH point clouds. Due to the complexity of the problem, the project considers only some important classes. The toolbox uses a multi-scale approach: the point clouds are processed from the historical complex to architectural elements, making it suitable for different types of heritage.

Mainly in recent years, deep learning has received increasing attention from the researches and has been successfully applied to semantically segment 3D point clouds in different domains [14, 62]. In the context of cultural heritage there are still few studies that use deep learning approaches to classify 3D point clouds. The need to have a large scale well-annotated dataset can limit its development, blocking the research in this direction. In some cases this problem can be solved using synthetic dataset [63, 64]. However, the researches conducted so far have yielded encouraging results.

Deep learning approaches are properly used for directly managing the raw data of point clouds without considering an intermediate processing that allows a more regular representation. For this purpose the first approach is proposed in [34]. An extended version of the previous network considers not only each point separately, but also its neighbors, in order to exploit the local features and thus obtain more efficient classification results [65].

The framewor proposed in [66] uses PointNet++ to semantically segment 3D point clouds of CH dataset. The aim of the paper is to demonstrate the efficiency of chosen deep learning approaches to process point clouds of CH domain. Moreover, the method is evaluated on a suitably created CH dataset manually annotated by domain experts.

An alternative to these approaches is based on the point clouds Convolutional Neural Network (PCNN) [67], a novel architecture that uses two operators (extension and restriction). The extension maps functions defined over the point cloud to volumetric functions, while the restriction operator does the inverse.

An approach inspired by PointNet is proposed by [68] where the difference is to exploit local geometric structures using a neural network module, EdgeConv, that constructs a local neighborhood graph and applies convolution-like operations. Moreover the model, named DGCNN (Dynamic Graph Convolutional Neural Network), dynamically updates the graph, changing the set of k-nearest neighbors of a point from layer to layer of the network.

Inspired by this architecture, the author of this thesis will propose to semantically segment 3D point clouds using an augmented DGCNN [69] by adding features such as normals and the radiometric component. In fact, these features have very discriminating and useful information for the segmentation task, and are never used by other

state-of-the-art methods. This modified version, described in Section 3.1 has the aim to simplify the management of DCH assets that have complex geometries, extremely variable and defined with a high level of detail. The author will also propose a novel publicly available dataset to validate the novel architecture making a comparison between other DL methods.

Another study that uses DL to classify objects of CH is presented in [52]. The authors make a performances comparison between machine and deep learning methods in the classification task of two different heritage datasets. Using machine learning approaches (Random Forest and One-versus-One) the performances are excellent in almost all the identified classes, but there is no correlation between the characteristics. Using DL approaches (1D CNN, 2D CNN and RNN Bi-LSTM) the 3D point clouds are considered as a sequence of points. However ML approaches overcome DL, because according to the authors the DL methods implemented are not very recent. In Section 3.1.2, the author of this thesis will propose a comparison similar to that described in [52], but using much more recent Deep Learning methods. This comparison will be useful to design an improved version of the proposed DGCNN-Mod network.

2.1.2 Point Clouds Generative Approaches

The semantic segmentation task that classifies different parts associating to each part a label, is an approach most used in the cultural heritage field [56, 70, 53]. The problem is that the training of deep learning networks requires a great amount of data mainly considering CH dataset with unique and unrepeatable elements. Moreover, the dataset has unbalanced classes and then the networks are not able to correctly recognize the objects (e.g. wall, roof, column, vault, etc.) due to the lack of data [69]. For this reason, mainly in the last years, the use of Generative Networks have filled this gap since they are able to generate novel suitable data as well as learn [71].

Synthetic data are used in several application fields. One of the pioneer works is proposed by [72] that demonstrates that the use of simulated data has significant advantages. This work introduces a simulator that describes and stochastically generates a sequence of DNA with different repeated structures. Most recent work is proposed by [73] that produces synthetic point clouds departing from a single 2D-dimensional image and reconstructing the 3D geometry of the complete object. To do this task they use an algorithm based on deep learning. A 3D point cloud is automatically generated in the work of [74] and used to train a traditional random forest network. This approach is based on supervised learning to classify 3D real urban scenes. The work of [75] extracts synthetic point clouds of urban scenes (road scene) from the Grand Theft Auto famous videogame to augment a standard benchmark dataset (KITTI). This approach has the aim to increase the semantic segmentation task based on Convolutional Neural Networks. Street scenes are also dealt in the work of [76] where virtual LiDAR sensors are used to acquire synthetic point clouds. The work simulates several point

clouds acquisition tools. The urban context is also the object of the work proposed by [64], where there is a need of a great amount of data to train a deep neural network to classify a 3D point cloud. The authors create a synthetic dataset and intend to demonstrate that the network trained with the dataset is able to generalize correctly the points cloud. In the work of [77] the authors propose a GAN to generate 3D points cloud departing from RGB-D images and corresponding to a single red, green, blue image. The method involves two phases. In the first phase, a generative adversarial network generates a depth image estimation from a single RGB image. In the second the 3D point cloud is calculated from the depth image. During the experimental phase, they demonstrate that the method provides high-quality 3D point clouds from single 2D images.

The work of [71] proposes a novel GAN that creates in an unsupervised way dense coloured 3D point clouds of various classes of objects. They propose a point transformer that using a graph convolution to increase in progress of the network, this to overcome the problem to acquire complex details with high resolution. The aim of the paper is to create a network able to produce coloured point clouds with fine details at multiple resolutions. In the context of cultural heritage interesting is the work of [78] where they propose a model that combines the latent-space GAN and Laplacian GAN architectures to form a multi-scale model capable of generating 3D point clouds at augmenting levels of detail. During the experimental phase they demonstrate that the method is better than other object of comparison. The work of [79] shows different training approaches with the task to classify optical character in historical documents. To solve the problem to have a great amount of annotated data, they summarize several methods to create a synthetic dataset. Moreover, they train a convolutional recurrent neural network using input a synthetic dataset and validate the approach using a real annotated dataset.

In this context, the aim of this thesis is to propose a framework based on DL that synthetically generates additional architectural elements to increase segmentation accuracy. To generate novel scenes, the author will use three different generative networks: PointGrow [80], PointFlow [81], and PointGMM [82]. Moreover, to compare the best performances, the proposed DGCNN-Mod network will be trained to classifies the synthetically generate scenes [69]. The experiments have been performed to ArCH dataset described in [83]. In contrast to many existing datasets, it has been manually labelled by domain experts, thus providing a more precise dataset.

2.2 Temporal Evolution of a Space

Another interesting challenge concerns the tracking of the temporal evolution of a space, an environment or the entities within them. At the state of the art there are several approaches concerning video surveillance or study of consumer behaviour, in the Retail domain. The temporal evolution of parts of the space can be carried out

through Change Detection techniques. Also in this case, the type of approach to be used is strictly related to the acquired data.

Change detection using an image aims to detect the changed areas by utilizing images of the same region captured at different points in time [84]. Many change detection methods have been proposed to detect changing areas, with machine learning being used in recent years [15, 16, 17, 18, 19]. The work of [15] proposed a convolutional neural network (CNN) architecture that measures changes in a region using an implicitly learned metric with a contrastive loss threshold [16]. To detect precise temporal changes in a region, a superpixel segmentation method that integrates CNN features has been introduced [17]. The work of [18] proposed a novel model for change detection in UAV images based on a supervised deep Siamese CNN. [19] introduced an object-based change detection method that uses multitemporal UAV images. Change detection in urban areas is performed by finding an elevation difference based on point cloud data from the aerial images acquired at different times.

The works just described concerns the task of change detection in open environments, using unmanned aerial vehicle (UAV) images. Completely different approaches will be used for smaller environments, where the resolution of the cameras is much lower but allows for much more complex surveillance systems.

This section will mainly describe Scene Change Detection algorithms. Scene Change Detection (SCD) aims to compare images captured at different times to identify changes that occur in the image. Till now, scene change detection has found various application scenarios, such as land cover monitoring [85], medical diagnosis [86], urban landscape analysis and autonomous driving [87, 88, 89]. With the development of learning systems for image classification and semantic segmentation, Convolutional Neural Networks (CNNs) have been widely used in computer vision tasks. The state-ofthe-art networks, such as VGG [90] or ResNet [91], can extract well-learned feature maps to gain excellent performance. Since the architectural advance of fully convolutional networks [92], image segmentation can be performed in an end-to-end fashion. Similar to dense image semantic segmentation, scene change detection also addresses pixelwise detection. Because of this, current networks designed for scene change detection are mostly based on CNNs and encoder-decoder-architectures [85, 87, 89, 93, 94].
On account of the specification of scene change detection tasks, as for coping with input images at two temporal points, some frameworks [85, 87, 88] choose to concatenate the input paired images at the beginning. In addition, some methods [17] first extract the feature maps in the two branches and later process the fusion to determine the change areas. Further, more research works [89, 93, 94] have put efforts on approach to fusion of the feature maps extracted from paired images. These fusion mechanisms are called early-fusion, late-fusion or correlation-fusion.
Recently more often, research works [93, 94] focus on the relation of feature maps at the same level. The stateof-the-art method is driven by feature correlation [89, 94], which estimates the likelihood of finding the similarity in a fixed neighbourhood,

along with paired channel fusion [93], which combines the same-level feature maps by a cross feature stack to make the channels interweave. However, existing approaches may be incapable of handling objects with various sizes and forms due to their regular and limited sampling ranges. Both quantitative and visual results indicate that the relation of feature maps still has huge research space.
In order to achieve better prediction performance, the network for SCD tends to become deeper, such as CSCDNet [89] has the encoder based on the ResNet block, but more deeper, CDNet++ [94] inserts five correlation layers, each at the corresponding feature level. However, deeper network will inevitably lead to a lower efficiency. Therefore, the network for SCD requires further work on improving the trade-off between efficiency and performance. Moreover, the estimated change masks in most networks only have the rough outline, yet more detailed change detection should be exploited. For a better comprehension, only the generic Change Detection term will be used in this thesis.

2.2.1 Change Detection on the Retail field.

At present, there are many works concerning the task of Change Detection in the retail world. In fact, systems of this type are widely used both for video surveillance and for marketing, concerning the study of consumer behaviour. These types of systems can be divided into two types: multi-camera systems, which are expensive and difficult to monitor, and single-camera systems supported by robots. The latter are becoming very successful as they also allow a certain level of interaction with consumers.

This section provides an overview of various works where robots have been used in the retail environment and works related to image classification using DCNNs. With the goal to change people's lifestyles for the better, robots are being deployed in various fields, such as construction, transportation, services, cleaning, surveillance, welfare, etc [95]. Robots are also being increasingly deployed in the retail environment, for both indoor and outdoor services. In [96], a virtual reality-based system is proposed for automating data collection and surveying in retail stores using mobile robots with the economic deployment solution. While sensors and cameras are employed in some cases to analyse the pre-purchase behaviour of the customers [97], in the others, the physical robots are used to assist the customers. An interactive system as a ubiquitous networked robot system is presented in [98], with the communication robots installed both outside and inside a shop. Matsuhira et al. [99] have implemented a robotic transport system to assist people during shopping. Another example of robot-assisted shopping is given in [100], where RoboCart assists visually impaired customers in navigating a typical grocery store and carrying purchased items with RFID and laser technologies. A robot used for the outdoor environment was presented in [101] to assist with shopping delivery and garbage collection, operating on domestic or condominium WLAN for easier communication and equipped with PC.

A human-robot interaction was reported in [102], where the robot could accept input from a customer and give output verbally or in written messages. In [103], the SugarTrail was developed, a system that does not require maps but learns traversable path-clusters to build a navigable virtual roadmap of the environment.

In an intelligent retail environment, two strategies are currently adopted for monitoring the customers' behaviour and trajectories: UWB technologies and RGB-D cameras. UWB technology for indoor tracking is considered the most promising solution in terms of accuracy, reliability, system cost, etc [104]. The system using UWB technology can monitor consumer movement in stores and send tracking data to a cloud server. However, several papers have proposed the use of RGB-D cameras to understand shopper's behaviour in front of a shelf [105], [106].

Ever since convolutional networks (ConvNets) were introduced by [107] in the early 1990s, they have demonstrated excellent performance in tasks such as hand-written digit classification and face detection. In the past years, convolutional networks have been applied in more challenging visual classification tasks, i.e., large-scale image and video recognition, with great success [108], [109], [110] due to large public image repositories, high-performance computing systems, and/or large-scale distributed clusters [111]. ImageNet Large-Scale Visual Recognition Challenge (ILSVRC) has played an important role in advancing deep visual recognition architectures, serving as a testbed for a few generations of large-scale image classification systems [108]. A notable result is shown in [108] with the record-beating performance on the ImageNet 2012 classification benchmark and the proposed ConvNet model achieving an error rate of 16.4%.

The main task of convolutional networks is recognising the dominant object's category in an image, where ConvNets have been applied for many years, whether the objects are handwritten characters [107], house numbers [112], traffic signs [113], or objects from the 1000-category ImageNet dataset [108]. Ever since the ImageNet 2012 win, the authors' network AlexNet has been successfully applied to various computer vision tasks, such as object-detection [114], semantic segmentation [92], human pose estimation [115], etc.

In DCNNs, there are typically two phases: feature extraction and classification [116]. Two-dimensional spatial convolutional kernel function layers, along with normalisation and pooling layers [108] are used for the feature extraction phase. The classification phase consists of 1-D linear connectivity in fully connected layers. In addition to the AlexNet, there is a number of other popular classifiers, such as CaffeNet [117], GoogleNet [118], ResNet [119], and VGG net [90]. CaffeNet has a similar structure to AlexNet, with a reversed order of pooling and normalisation. ResNet is a collection of DCNN architectures inspired by the philosophy of VGG nets and employs residual connections, which allows errors to better propagate backwards through the network.

Section 3.2 will describe a novel solution for an intelligent retail environment, based on a robot called ROCKy [120], specifically designed to facilitate automatic change

detection and map SOOS and PA events. Another important contribution is the introduction of an approach to estimate the classification of stock assortment by shelf pictures acquired by ROCKy.

2.3 Short-term Evolution of Space's entities

Finally, the temporal evolution of a space is also strictly related to the entities within it. Their localisation, in a short period of time, can be solved through Re-Identification techniques. At the state of the art, several methodologies exist for the re-identification of different types of entities: cars, indoor objects, animals, etc. [20, 21].

The task of re-identification (re-id) [121] has long been a task of extracting features or representations from two observations and measuring how similar these features are. Since different variations affect these features, many works have introduced different methods to improve their extraction. Initially, these features were hand-crafted and include spatio-temporal information such as color, width and height, and salient edge histograms. Some work have also tried to use different input modalities such as depth [122, 123, 122], infrared [124], LiDAR [125], or Inductive Loop Detectors for vehicles [126]. These features, however, fail drastically when dealing with unexpected scenarios. To remedy this problem and with the advent of machine learning, researchers are now benefiting from the strength of deep learning to be able to extract more general and more discriminative features allowing them to reach high performance. Since then, an arms race of methods was built on top of this by making use of different object-specific characteristics (e.g., human semantic segmentation, pose [127]) and by learning features through the supervision of a cross-entropy loss [121].
The most common and dynamic entity that can be found within a space are people. In fact, the scientific community has published several approaches on Person Re-Identification. There are approaches with frontal configuration and approaches with top-view configuration. A lot of methods use RGB images, other thermal or infrared images; hybrid approaches also exist, which integrates diffent type of data [22].

2.3.1 Person Re-Identification

Person re-identification (re-ID) is the task of recognising individuals at different locations and times, which involves different camera views, poses and lighting [127]. This topic has gained increasing interest in the computer vision community due to its challenging nature, and its important practical role underpinning many visual surveillance functionalities, including person searching and tracking across disjointed cameras [128]. Person re-ID has been adopted in several domains ranging from video surveillance to retail [129].

In a common real-world application, a watch list of known people is given as the

gallery/target set for searching through a large volume of video recording locations where where those selected people are likely to return. This aspect is fundamental in retail to understand how customers schedule their shopping. The identification of regular and occasional customers allows temporal purchasing profiles to be defined, which can correlate customer temporal habits with other information, such as expenditure amounts and numbers of purchased items. This knowledge enables novel marketing strategies tailored to the temporal and systematic behavior of each customer, as well as new innovative services and increased customer awareness based on shopping schedule recommendations [130, 131].

Video captured by store cameras usually contain people who are not part of a watch list. Moreover, a target person can appear similar to a non-target person whilst dissimilar to target gallery videos due to significant changes lighting and view angle conditions across camera views. To further aggravate the problem, there may only be a single gallery image (a one-shot) available for each target person, preventing the effective learning of a target's appearance variations. Facing re-ID issues becomes difficult in a crowded retail environment with many occlusions [132], especially where probe sets contain mostly irrelevant (non-target) people. This problem is called open-world re-ID [133, 134, 135]. For such a challenging problem, depending on a fully automated system to provide exhaustive accurate verification against each targeted individual is neither scalable nor tractable. Nonetheless, it is adequate to expect an automated system to produce some screening by dealing with an easier problem: checking if a targeted person is in a given set (group-based person verification), whilst leaving the more challenging task of individual identification within the set for a human operator. Since watch lists are typically small, human verification can be carried out quickly and more robustly. Many approaches investigate either the best feature representation [136], [137], [138], [139], [140] or the best matching metrics [141], [142] when using person re-ID under difficult appearance changes across camera views.

They are not suitable to re-identify people in the retail environment as they assume a closed-world setting with probe sets containing exactly the same people in the gallery set. For probe sets consisting of mostly non-target people (many more than those in the gallery set), the re-ID problem becomes more arduous. They also do not consider retail environments where analytic interactions and re-ID are developed with the aim of learning shopper skills based on occlusion-free RGB-D cameras in a top-view configuration [106, 143, 144].

Furthermore, re-identifying a person in more crowded situations is a problem that remains largely unresolved due to many serious issues, such as the exhibition of persistent occlusion, appearance changes and dynamic or complex backgrounds. All of these issues cause extreme problems when encountered with a crowded environment, since conventional surveillance technologies have difficulty understanding video (Figure 2.1).

Open-set re-ID is much closer to practical video surveillance applications but its

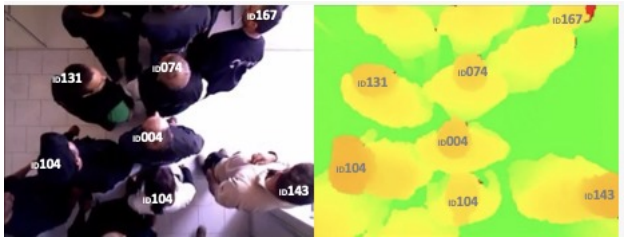


Figure 2.1: Example of RGB-D videos acquired in a persistent crowded environment with person identification. The figure depicts both RGB (left) and Depth (right) streams, showing that the top-view approach allows to avoid occlusions between people, a situation where the frontal approach often fails.

low recognition rates under low false accepted rates of existing results show that this setting is very challenging [128]. Historically, the scientific community has been devoted mainly to closed-set re-ID, a mature technology [145] that is convenient and fair for conducting research given its various baselines, datasets and evaluations.

However, open-set re-ID is a realistic approach that considers irrelevant people (those not part of a gallery) during recognition [146]. It can be defined as a person verification task instead of person identification, allowing verification of those who are part of a gallery and images in which those subjects appear [147]. The evaluation metrics are different respect the closed world ones. In fact, two metrics were defined in [133], namely high true target recognition (TTR) and low false target recognition (FTR), which focus on calculating the likelihood of target and non-target numbers of images being verified as target identities.

The first work for open-set re-ID was proposed in [148]. The authors showed a transfer ranking framework for set-based verification. Another approach [149], was based on online conditional random field inference. In [134], open-set re-ID was decomposed into detection and identification while also being presented as two generic evaluation metrics (i.e., identification rate and false acceptance rate). In [150], the authors tested a regularised kernel subspace learning model for one-shot verification by learning crossview identity-specific information from just unlabeled data. Authors of [133] presented clearer descriptions of open-set challenges and standard evaluation metrics, describing a group-based setting and a transfer local relative distance comparison model for addressing label scarcity. For performance evaluation, they used TTR and FTR. [135] proposed a hashing approach (cross-view identity correlation) and introduced a large-scale setting characterised by huge size probe images and an open person population.

Common re-ID approaches are usually based on frontal image datasets, but sensors installed in top-view configuration have been revealed as especially effective in crowded environments [132]. The latter configuration has several advantages because it prevents occlusion due to objects and other people while ensuring personal privacy,

as faces are not recorded. The work of [151] proposed a method to extract anthropometric features through image processing techniques, then training machine-learning algorithms for re-ID tasks. Their tests were carried out on a dataset of 100 people acquired using a top-view RGB-D camera.

In [152], authors developed an attention-based model that deduces human body shape and motion dynamics by using depth information. Their approach was a combination of convolutional and recurrent neural networks leveraging unique 4-D spatio-temporal signatures to identify small discriminative regions indicative of human beings. Their tests were assessed on a DPI-T dataset, which consisted of 12 persons appearing in 25 videos while wearing different sets of clothing and holding different objects. In [153], the authors started with a two-flow Convolutional Neural Network (CNN) (one for RGB and one for depth) and a final fusion layer. They improved on this approach with a multimodal attention network [154], adding an attention module to extract local and discriminative features that were fused with globally extracted features. Another work [155] presented a SLATT network with two types of attention modules (one spatial and one layer-wise). The authors collected also the OPR dataset from a university canteen, which was composed of 64 persons captured twice (entering and leaving a room). However, these datasets are not publicly available.

Recently, the person re-ID task is often solved using a triplet loss function, with excellent performance. In the work of [156], the authors propose a batch hard function expecially designed for person re-ID problem: they show that, for both models trained from scratch or pretrained ones, using a well designed triplet loss can outperform most state-of-the-art methods. [157] present a triplet loss that achieves good performance with large-scale re-ID datasets and has direct transferability with unseen datasets. The framework proposed in this thesis uses a triplet loss function based on the work of [156].

In Section 3.3, the author of this thesis will presents the first attempt to solve a more realistic re-ID setting, facing these important issues using top-view open-world (TVOW) person re-ID. Its backbone is based on a pretrained deep convolutional neural network (DCNN) that has been fine-tuned on a dataset acquired via a top-view configuration. This framework is trained by using a triplet loss function based on the work of [156], optimising the embedding space by approximating the features of frames of the same person while distancing the features of frames of different people. Similar to [156], triplet loss allows end-to-end learning between input images and a desired embedding space. At the inference phase, test people are compared by computing the Euclidean distance of their embeddings. In addition to the normal metrics used in a closed-set environment, particular metrics were defined, employed and evaluated for an open-set environment. The TVOW framework will be evaluated on a new publicly dataset, called TVPR2 [158], both in closed world and open world configurations.

Museum environment

In Cultural Heritage (CH) domain, as well as in museum environment, understanding and analysing users' activities and behaviours is becoming imperative. Users' behaviour can provide important statistics and insights on what happens inside this space, which are the successful exhibitions, and which are the interactions with the artworks [159]. Nowadays, it is possible for museum curators and personnel obtaining feedback on museums thanks to online purchases, social networks and other communication channels [160]. However, there is a limited knowledge about the circumstances that occur during the visit, and the layout, the arrangement of works, the management of flows can be designed according to the real needs of users, after collecting their information. Museum exhibits are usually arranged considering the target of users. This condition emerges from the obstacle in understanding a priori visitors' interests [161]. The concept of Smart Environment identifies a place able to acquire and apply knowledge about the environment and its inhabitants, in order to improve both their experience (i.e. by automatically reacting to some events is even more attractive and challenging way) and the knowledge of the space itself (i.e. by providing managers with useful information for security or arrangement reasons). By means of innovative technological applications, it is possible to leverage novel human space interaction paradigms over the existing proxemic interaction space model-based user interfaces, nowadays determined by the purely aesthetic and essentially passive fruition of cultural objects [162].

Currently, there is a lack of reliable solution which can fulfil such important tasks. In fact, these data represent a precious value for the museum curators, and they are one of the parameters need to be assessed [163]. For this reason, a data-driven approach for collection and analysis provide an objective and reliable source of information. One of the current trend is to configure location-aware services, i.e., applications driven by location information, in particular, by users movements in the environment [164], [165]. Recently, RGB-D cameras have demonstrated their suitability for solving this task. In fact, this kind of solution provides affordable, additional rough depth information coupled with visual images, offering enough resolution and accuracy for indoor applications. Furthermore, RGB-D camera in a top-view configuration reduces the problem of occlusion, it allows precise people counting, and it has the advantage of preserving privacy by not recording faces, and it is easier to set up on ceiling installation [166].

Considering museum environments, an automatic re-ID system can provide important information to improve user experiences. The re-ID of users that move within museum enables understanding which artworks are most attractive, the displacement inside the spaces, and possible stops, as well as classifying different users groups and targets. Several previous researches have adopted the top-view configuration because it facilitates the extraction of trajectory features and ensures greater robustness. Furthermore, reliable depth maps can provide valuable additional information that can significantly improve detection and tracking results [143]. In a crowded environment

(more than three people per square metre), an RGB-D system with Top-View configuration provides high accuracy [167].

In Section 3.3.2, it will be presented SeSAME (Senseable Self Adapting Museum Environment) a novel system for collecting and analysing the behaviours of visitors inside a museum environment. SeSAME uses re-identification (re-ID) techniques to perform visual profiling of visitor interest.

Chapter 3

Methodologies proposed for static and dynamic space understanding frameworks

This chapter describes the methodologies and tools used by the author of the thesis for the Space Understanding.

Initially, starting from space as a static concept, its understanding is defined through the computer vision task called semantic segmentation. Moreover, the author has chosen to process point cloud data because they are currently easy to acquire and effectively define the location of any entity in the 3D space. For these reasons, Section 3.1 will describe a framework for the semantic segmentation of point clouds.

Subsequently, the author defines methodologies to study the temporal evolution of the environment and the entities within it, so it will be described as a dynamic concept. In Section 3.2 an innovative approach to solve the change detection task will be presented, using RGB images acquired over time, from which further information will be extracted in order to define temporal changes of the space.

Finally, in Section 3.3, the author will design a new framework to study the short-term evolution of entities within a space. Since the most common and dynamic entities that can be found in a space are usually people, the author will define a framework to solve the Person Re-Identification task.

3.1 Point Clouds Semantic Segmentation for a static space

As already mentioned, this section describes a new framework for the semantic segmentation of point clouds. Initially, the author presents in Section 3.1.1 a new Deep Learning method based on a dynamic graph network, improved by the use of new point cloud attributes to achieve better embedding feature learning. In order to perform the experiments, the author contributed to the generation of a new dataset concerning scenes from the Cultural Heritage field. The approach is tested on both an indoor

and an outdoor scene dataset, allowing a proper generability analysis. Subsequently, through a comparison with the current state-of-the-art methods of Machine Learning and Deep Learning, the author defines in Section 3.1.2 a hybrid approach between the two domains in order to obtain a better understanding of the space. Finally, since the weaknesses of generic deep learning approaches for semantic point cloud segmentation are not only related to methods but also to dataset issues, the author describes in Section 3.1.3 a framework to solve these issues based on generative approaches.

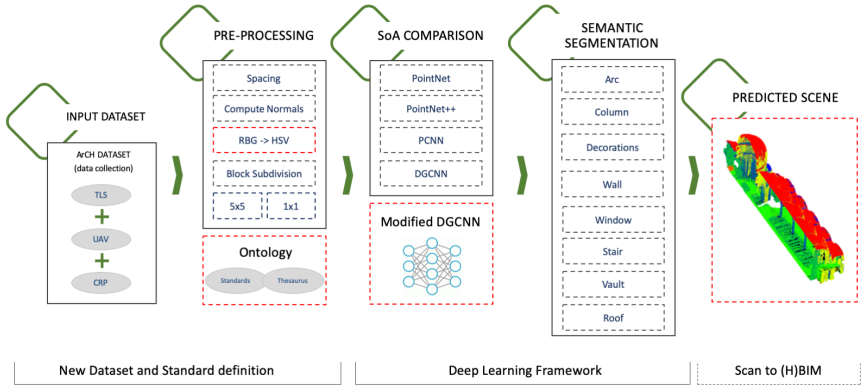


Figure 3.1: DL Framework for Point Cloud Semantic Segmentation.

3.1.1 DGCNN-Mod Network

The understanding of a static scene is solved by the author of this thesis by proposing a DL framework for Point Cloud segmentation, inspired by the work presented in [68]. Instead of employing individual points like PointNet [34], the approach proposed in [68] exploits local geometric structures by constructing a local neighborhood graph and applying convolution-like operations on the edges connecting neighboring pairs of points. This network has been improved by adding relevant features such as normal vectors and HSV encoded color. The experiments has been performed to a completely new and recent dataset regarding the Cultural Heritage domain. This dataset comprises different labeled points clouds, derived from the union of several single scans or from the integration of the latter with photogrammetric surveys. The involved scenes are both indoor and outdoor, with churches, chapels, cloisters, porticoes and archades covered by a variety of vaults and beared by many different types of columns. They belong to different historical periods and different styles, in order to make the dataset the least possible uniform and homogeneous (in the repetition of the architectural elements) and the results as general as possible. In contrast to many existing datasets, it has been manually labelled by domain experts, thus providing a more precise dataset. The resulting network achieves promising performance in recognizing elements. A comprehensive overall picture of the developed framework

is reported in Figure 3.1. This approach has been published on [69] and its pipeline might represent a baseline for further experiments from other researchers dealing with Semantic Segmentation of Point Clouds with DL approaches.

ArCH Dataset

In the state of the art, the most used Point Clouds datasets to train neural networks are: ModelNet 40 [35] with more than 100k CAD models of objects, mainly furnitures, from 40 different categories; KITTI [168] that includes camera images and laser scans for autonomous navigation; Sydney Urban Objects [169] dataset acquired with Velodyne HDL-64E LiDAR in urban environments with 26 classes and 631 individual scans; Semantic3D [170] with urban scenes as churches, streets, railroad tracks, squares and so on; S3DIS [39] that includes mainly office areas and it has been collected with the Matterport scanner with 3D structured light sensors and the Oakland 3-D Point Cloud dataset [171] consisting of labeled laser scanner 3D Point Clouds, collected from a moving platform in a urban environment. Most of the current datasets collect data from urban environments, with scans composed of around 100 K points, and to date there are still no published datasets focusing on immovable cultural assets with an adequate level of detail.

During this thesis work, a collaboration between the author's research group (VRAI group), the Polytechnic of Turin and the Fondazione Bruno Kessler of Trento was born. This collaboration has led to the development of a new dataset of point clouds called ArCH (Architectural Cultural Heritage) [172]. This dataset is actually composed of 17 labeled scenes, derived from the union of several single scans or from the integration of the latter with photogrammetric surveys. In the first experiments of this work, the first version of the dataset will be used, containing only the first 11 scenes that were labelled by the authors. Figure 3.2 shows the acquired point clouds and their ground truth, including the classes that have been segmented.

The involved scenes are both indoor and outdoor, with churches, chapels, cloisters, porticoes and archades covered by a variety of vaults and beared by many different types of columns. They belong to different historical periods and different styles, in order to make the dataset the least possible uniform and homogeneous (in the repetition of the architectural elements) and the results as general as possible.

Different case studies are taken into exam and are described as follows: The Sacri Monti (Sacred Mounts) of Ghiffa (SMG) and Varallo (SMV); The Sanctuary of Trompone (TR); The Church of Santo Stefano (CA); The indoor scene of the Castello del Valentino (VA). The ArCH Dataset is publicly available [1] for research purposes.

[1] http://archdataset.polito.it/

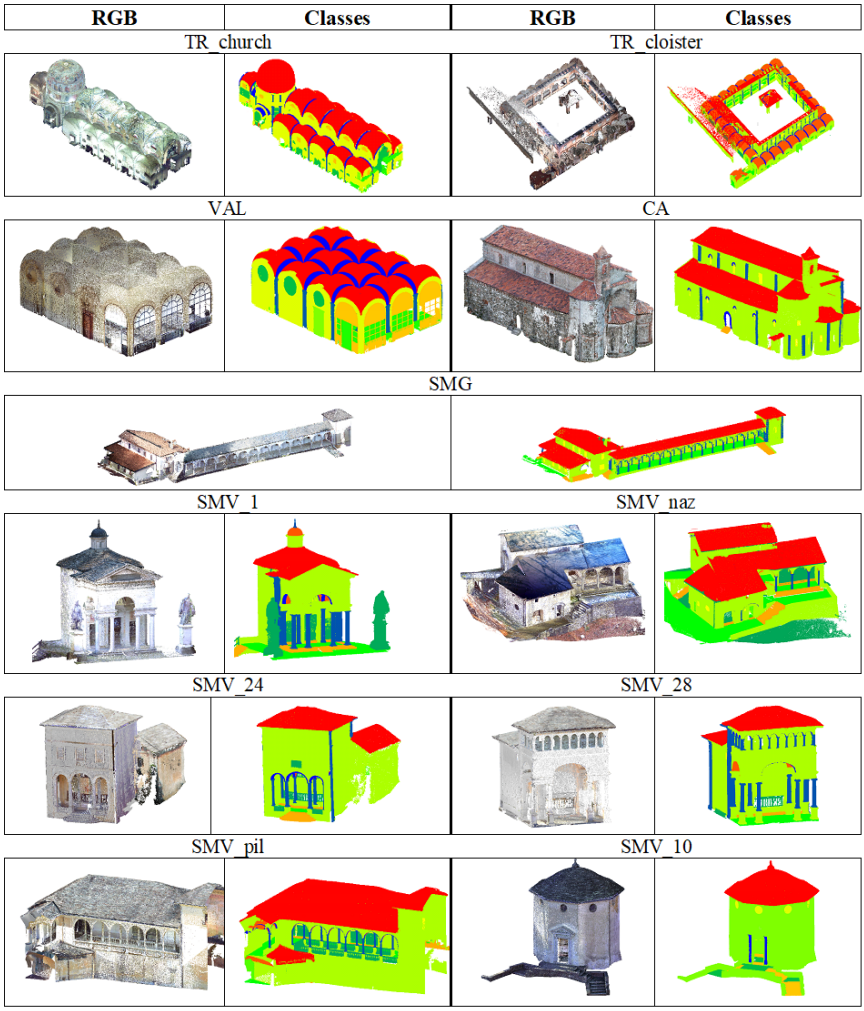


Figure 3.2: ArCH dataset. On the left column the RGB point clouds and on the right the annotated scenes. 10 classes have been identified: Arc, Column, Door, Floor, Roof, Stairs, Vault, Wall, Window and Decoration. The Decoration class includes all the points unassigned to the previous classes, as benches, balaustrades, paintings, altars and so on.

In the majority of cases, the final scene was obtained through the integration of different Point Clouds, those acquired with the terrestrial laser scanner (TLS), and those deriving from photogrammetry (mainly aerial for surveying the roofs), after appropriate evaluation of the accuracy. This integration results in a complete Point Cloud, with different density according to the sensors used, however leading to increasing the overall Point Cloud size and requiring a pre-processing phase for the NN.

The common structure of the Point Clouds is therefore based on the sequence of the

coordinates *x, y, z* and the *R, G, B* values.

In the future, other point clouds will be added to the ArCH dataset, to improve the representation of complex CH objects with the potential contribution of all the other researchers involved in this field.

Data Pre-Processing

To prepare the dataset for the network, pre-processing operations have been carried out in order to make the cloud structures more homogeneous. The pre-processing methods, for this dataset, have followed 3 steps: translation, subsampling and choice of features.

The *spatial translation* of the Point Clouds is necessary because of the georeferencing of the scenes, the coordinate values are in fact too large to be processed by the deep network, so the coordinates are truncated and each single scene is spatially moved close to the cardinal point (0,0,0). This operation on the one hand has led to the loss of georeferencing, on the other hand, however, it has made possible to reduce the size of the files and the space to be analyzed, thus also leading to a decrease in the required computational power.

The *subsampling* operation, which became necessary due to the high number of points (mostly redundant) present in each scene ($>$ 20M points), was instead more complex. It was in fact necessary to establish which of the three different subsampling options was the most adequate to provide the best typology of input data to the neural network. The option of random subsampling was discarded because it would limit the test repeatability, then both the other two methods have been tested: octree and space. The first is efficient for nearest neighbor extraction, while the second provides, in the output Point Cloud, points not closer than the distance specified. As far as space is concerned, it has been set a minimum space between points of 0.01 m, in this way a high level of detail is ensured, but at the same time it is possible to considerably reduce the number of points and the size of the file, in addition to regularize the geometric structure of the Point Cloud

As for the octree, applied only in the first tests on half of the Trompone Church scene, level 20 was set, so that the number of final points was more or less similar to that of the scene subsampled with the space method. The software used for this operation is CloudCompare. An analysis of the number of points for each scene is detailed in Table 3.1, where it is possible to see the lack of points for some classes and the highest total value for the 'Wall' class.

Table 3.1: Number of points per class and overall for the whole scene. The point cloud of the Trompone church has been split into right (r) and left (l) part according to the tests conducted in Section 4.1.1

Scene/class	Arc	Column	Decoration	Floor	Door	Wall	Window	Stairs	Vault	Roof	TOTAL
	0	1	2	3	4	5	6	7	8	9	
TR_cloister	900,403	955,791	765,864	1,948,029	779,019	10,962,235	863,792	2806	2,759,284	1,223,300	21,160,523
TR_church_r	466,472	658,100	1,967,398	1,221,331	85,001	3,387,149	145,177	84,118	2,366,115	0	10,380,861
TR_church_l	439,269	554,673	1,999,991	1,329,265	44,241	3,148,777	128,433	38,141	2,339,801	0	10,022,591
VAL	300,923	409,123	204,355	1,011,034	69,830	920,418	406,895	0	869,535	0	4,192,113
CA	17,299	172,044	0	0	30,208	3,068,802	33,780	11,181	0	1,559,138	4,892,452
SMG	309,496	1,131,090	915,282	1,609,202	18,736	7,187,003	137,954	478,627	2,085,185	7,671,775	21,544,350
SMV_1	46,632	314,723	409,441	457,462	0	1,598,516	2011	274,163	122,522	620,550	3,846,020
SMV_naz	472,004	80,471	847,281	1,401,120	42,362	2,846,324	16,559	232,748	4,378,069	527,490	10,844,428
SMV_24	146,104	406,065	154,634	20,085	469	366,2361	6742	131,137	305,086	159,480	4,992,163
SMV_28	36,991	495,794	18,826	192,331		1,965,782	4481	13,734	184,261	197,679	3,109,879
SMV_pil	584,981	595,117	1,025,534	1,146,079	26,081	7,358,536	313,925	811,724	2,081,080	3,059,959	17,003,016
SMV_10	0	16,621	0	125,731	0	1,360,738	106,186	113,287	0	499,159	2,221,722
TOTAL	**3,720,574**	**5,789,612**	**8,308,606**	**10,461,669**	**1,095,947**	**47,466,641**	**2,165,935**	**2,191,666**	**17,490,938**	**15,518,530**	**114,210,118**

The *extraction of features* directly from the Point Clouds is instead an open and constantly evolving field of research. Most of the features are handcrafted for specific tasks and can be subdivided and classified into intrinsic and extrinsic, or also used for local and global descriptors [173, 174]. The local features define statistical properties of the local neighborhood geometric information, while the global features describe the whole geometry of the Point Cloud. Those mostly used are the local ones, such as eigenvalues based descriptors, 3D Shape context and so on, however in our case, since the last networks developed [65, 68] tend to let the network itself learn the features and since the main goal of this work is to generalize as much as possible, in addition to reduce the human involvement in the pre-proccessing phases, the only features calculated are the normals and the curvature. The normals are calculated by using Cloud Compare software and have been computed and orientated with different settings depending on the surface model and 3D data acquisition. Specifically a plane or quadric 'local surface model' as surface approximation for the normals computation has been used and a 'minimum spanning tree' with kNN=10 has been set for their orientation. The latter has been further checked on MATLAB software.

Deep Learning approaches

State-of-the-art deep neural networks are specifically designed to deal with the irregularity of Point Clouds, directly managing raw Point Cloud data rather than using an intermediate regular representation. In this contribution, the performances obtained with the ArCH dataset of some state-of-art architectures are therefore compared and then evaluated with regards to the DGCNN that the author has modified. The NNs selected are:

- *PointNet* [34], as it was the pioneer of this approach, obtaining permutation in-

variance of points by operating on each point independently and applying a symmetric function to accumulate features.

- *PointNet++* [65] (extension of PointNet) that analyzes neighborhoods of points in preference of acting on each separately, allowing the exploitation of local features even if with still some important limitations.

- *PCNN* [67], a DL framework for applying CNN to Point Clouds generalizing image CNNs. The extension and restriction operators are involved, permitting the use of volumetric functions associated to the Point Cloud.

- *DGCNN* [68] that addresses these shortcomings by adding the EdgeConv operation. EdgeConv is a module that creates edge features describing the relationships between a point and its neighbors rather than generating point features directly from their embeddings. This module is permutation invariant and it is able to group points thanks to local graph, learning from the edges that link points.

The author's first contribution is the design of a new neural network called DGCNN-Mod [69], build upon the DGCNN implementation provided by [68]. Such an implementation of DGCNN uses k-nearest neighbors (kNN) to individuate the k points closest to the point to be classified, thus defining the neighboring region of the point. The edge features are then calculated from such a neighboring region and provided as input to the following layer of the network. Such a edge convolution operation is performed on the output of each layer of the network. In the original implementation, at the input layer kNN is fed with normalized points coordinates only, while in this novel implementation the network use all the available features. Specifically, the author added color features, expressed as RGB or HSV channels, and normal vectors.

Figure 3.3 shows the overall structure of the network. A scene block has been given in input, composed of 12 features for each point: XYZ coords, X'Y'Z' normalized coords, color features (HSV channels), normals features. These blocks pass through 4 EdgeConv layers and a max-pooling layer to extract global features of the block. The original XYZ coordinates are kept to take into account the positioning of the points in the whole scene, while the normalized coordinates represent the positioning within each block. The KNN module is fed with normalized coordinates only and both original and normalized coordinates are used as input features for the neural network. RGB channels have been converted to HSV channels in two steps: first they are normalized to values between 0 and 1, then they are converted to HSV channels using the rgb2hsv() function of the scikit-image library implemented in python. This conversion is useful because the individual channels H, S and V are independent one from the other, each of them has a different typology information, making them independent features. Channels R, G and B are conversely somehow related to each other, they share a part of the same data type, so they should not be used separately.

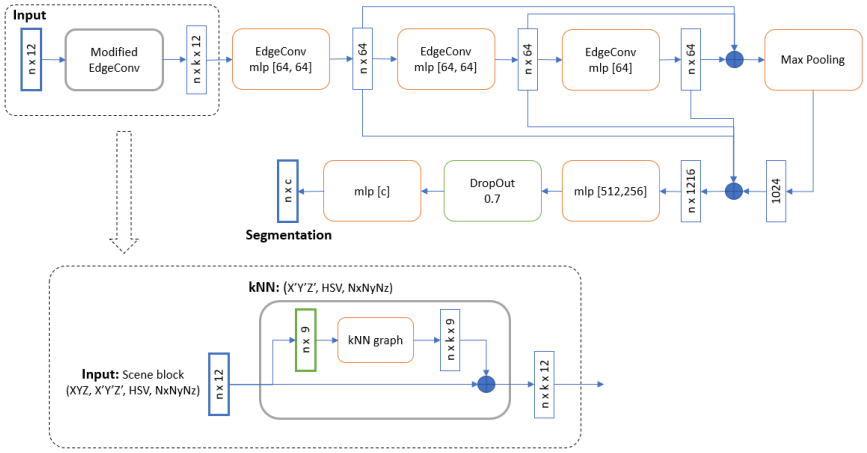


Figure 3.3: Illustration of the DGCNN-Mod architecture.

The choice of using normals and HSV features is based on different reasons. On one side the RGB component, based on the sensors used in data acquisition, is most of the time present as a property of the Point Cloud and therefore it has been decided to fully exploit this kind of data; on the other the RGB components define the radiometric properties of the point cloud, while the normals define some geometric properties. In this way different kinds of information has been used as input for the NN. Moreover, the decision to convert the RGB into HSV is borrowed from other research works [175] that, even if developed for different tasks, show the effectiveness of this operation.

The author's basic idea is to support the NN, in order to increase its accuracy, with these common features that could be easily obtained by any user. These features are then concatenated with the local features of each EdgeConv. The author has modified the first EdgeConv layer so that the kNN phase could also use color and normal features to be able to select k neighbors for each point. Finally, through 3 convolutional layers and a dropout layer, the NN outputs the same block of points but with segmentation scores (one for each class to be recognized). The output of the segmentation will be given by the class with the largest score.

Generalisation ability on indoor scenes

The ArCH dataset consists mainly of outdoor scenes, with classes describing objects related to the Cultural Heritage domain. It is interesting to test the proposed approach in other areas, such as indoor environments. The author of this thesis has therefore chosen the public dataset S3DIS [39], for evaluating the generalisation skill of the proposed method. S3DIS dataset is collected in 6 large-scale indoor areas that originate from 3 different buildings of mainly educational and office use. It includes the

colored 3D point cloud data of these areas with the total number of 695,878,620 points. The annotations are instance-level and correspond to 14 object classes. These classes represents structural elements like ceiling, floor, wall, beam, column, window, door, stairs, and commonly items like table, chair, sofa, bookcase, board, clutter. Figure 3.4 shows the 6 areas regarding this dataset.

Figure 3.4: The 6 indoor areas of S3DIS dataset.

3.1.2 A mixed approach between ML and DL methodologies

This Section describes how the author is able to improve the approach proposed in the previous Section. It starts with a comparison of state-of-the-art methods, both based on classical Machine Learning approaches and recent Deep Learning approaches. From this comparison, the best qualities of both types of approaches are derived and then merged into the next framework proposed by the author. This work has been published in [176].

Figure 3.5 shows the workflow of the comparison between the two methodologies, as well as the classifiers and scenes used for the three experiments presented in Section 4.1.2.

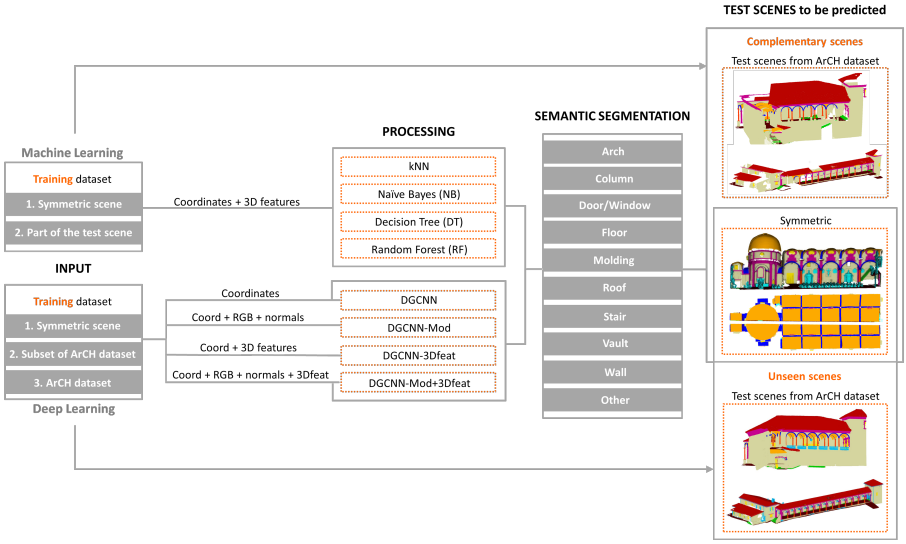


Figure 3.5: Workflow for the machine learning (ML) and deep learning (DL) framework comparison.

For a fair comparison between segmentation algorithms, it would be necessary to use the same training data. In this context, some initial experiments using the same number of scenes in the training phases for both DL and ML algorithms have been performed. However, the ML classifiers did not achieve satisfactory results compared with those obtained using reduced annotated portions of the test scenes. Therefore, as the aim of the work is discussing the best approaches for heritage segmentation, a comparison between ML and DL approaches is presented, where the training data are different.

Three different experiments have been performed as follows. In the first experiment both the different ML and DL classifiers have been trained on the same portion of a symmetrical scene: half of the point cloud is used for training and validation, and half for the final test. In the second and third experiment the samples used to train the ML and DL classifiers are different. On one hand, for the ML approach, a reduced portion of the test scene is annotated and used during the training phase, leaving the remaining part for the prediction phase. On the other, for the DL approach, different annotated scenes are used for the training phase, while for the testing totally new data are presented to the networks. Further details are given in the following subsections.

Benchmark for Point Cloud Semantic Segmentation

For the experiments of this part of the thesis, it was possible to use the final version of the ArCH dataset [172], already introduced in Section 3.1.1.

This benchmark represents the current state of the art in the field of annotated cul-

tural heritage point clouds, with 15 point clouds of architectural scenarios for training and 2 for test. Although other benchmarks and datasets for point clouds' classification and semantic segmentation already exist [169, 39, 177, 40], the ArCH dataset is the only one specifically focused on the CH domain and with a higher level of detail, therefore it has been chosen for the tests here presented.

For the experiments, three test scenes are used (Table 3.2): (i) the symmetrical point cloud of the Trompone Church, (ii) the Palace of Pilato of the Sacred Mount of Varallo - SMV (a two-floor building, not symmetrical and not linear), and (iii) the portico of the Sacred Mount of Ghiffa - SMG (a simpler and quite linear scene). For the DL approach, the symmetrical point cloud is used for an initial evaluation of the hyperparameters. Whilst the other two scenes allow to evaluate the generalisation ability of state-of-art neural networks by testing them on different cases: a complex one, SMV, and a simpler one, SMG. Moreover, these two scenes are the same as those defined as tests for the benchmark.

Table 3.2: Experiments performed with relative test and training sets.

Experiment	**Test Set**	**Training Set**		
		ML	**DL**	
1 *Overall Results in Table 4.8 and Figure 4.8*	Trompone Church - symmetrical half part -	Remaining half part	Remaining half part (Training and Validation)	/
2 *Overall Results in Table 4.9 and Figure 4.10*	SMV scene (Sacred Mount of Varallo)	16% of the test scene	10 scenes for Training and 1 for Validation	14 scenes for Training and 1 for Validation (whole ArCH dataset)
3 *Overall Results in Table 4.10 and Figure 4.12*	SMG scene (Sacred Mount of Ghiffa)	20% of the test scene	10 scenes for Training and 1 for Validation	14 scenes for Training and 1 for Validation (whole ArCH dataset)

Machine Learning Classifiers

Over the past ten years, different Machine Learning approaches have been proposed in the literature for point cloud semantic segmentation such as k-Nearest Neighbour (kNN) [178], Support Vector Machine (SVM) [179, 180], Decision Tree (DT) [181, 182], AdaBoost (AB) [183, 184], Naive Bayes (NB) [185, 186], and Random Forest (RF) [187]. Among them, in this work, kNN, NB, DT, and RF classifiers have been implemented in Python 3, starting from the available Scikit-learn Python library [188], in order to solve multi-class classification tasks. For each case study the four classifiers have been trained through selected features and small manually annotated portions of the datasets.

With regard to the kNN classifier, the k value being highly data-dependent, a few preliminary test with increasing values have been run, in order to find the best fit solution. Best results were achieved with low values of k ($k = 5$).

The NB classifier used is the GaussianNB [189], a variant of Naive Bayes that follows Gaussian normal distribution and supports continuous data.

For the DT, different maximum depths of the tree have been tested. Results confirmed that the default parameter *max-depth=None*, by which nodes are expanded until all leaves are pure, allows for higher accuracy results.

Within the RF classifier two parameters have been initially tuned considering the best F1-score computed on the evaluation set: the number of decision trees to be generated *Ntree* and the maximum depth of the tree *Mtry* [187]. The reported results refers to the use of 100 trees with *max-depth=None*.

Features Selection

In order to effectively train the different ML classifiers a composition of 3D geometric features have been used, including normal-based (Verticality), height-based (Z coordinates), and eigenvalue-based features (also defined covariance features).

The covariance features [190] are shape descriptors obtained as a combination of eigenvalues ($\lambda_1 > \lambda_2 > \lambda_3$) which are extracted from the covariance matrix, a 3D tensors that describe the distribution of point within a certain neighbourhood. Through statistical analysis, the Principal Component Analysis (PCA), it is possible to extract from this matrix the three eigenvalues representing the local 3D structure. According to Weinmann et al. [174], different strategies can be applied to recover the local neighbourhood for points belonging to a 3D point cloud. It can generally be computed as a sphere or a cylinder with a fixed radius or be described by the number of the kNN. In this work, considering the studies presented in [59, 60], only a few covariance features (Omnivariance, Surface Variation and Planarity) have been calculated on spherical neighbourhoods at specific radii in order to highlight the architectural components.

As one can see in Figure 3.6, different features emphasises different elements. Verticality makes easier the distinction between vertical and horizontal surfaces, allowing the recognition of walls and columns as well as floors, stairs and vaults. The feature planarity becomes useful for isolating columns and cylindrical elements if extracted at radii close to the diameter dimensions. Finally, surface variation and omnivariance, calculated within a short radius, emphasises changes in shapes facilitating, for example, the detection of moldings and windows.

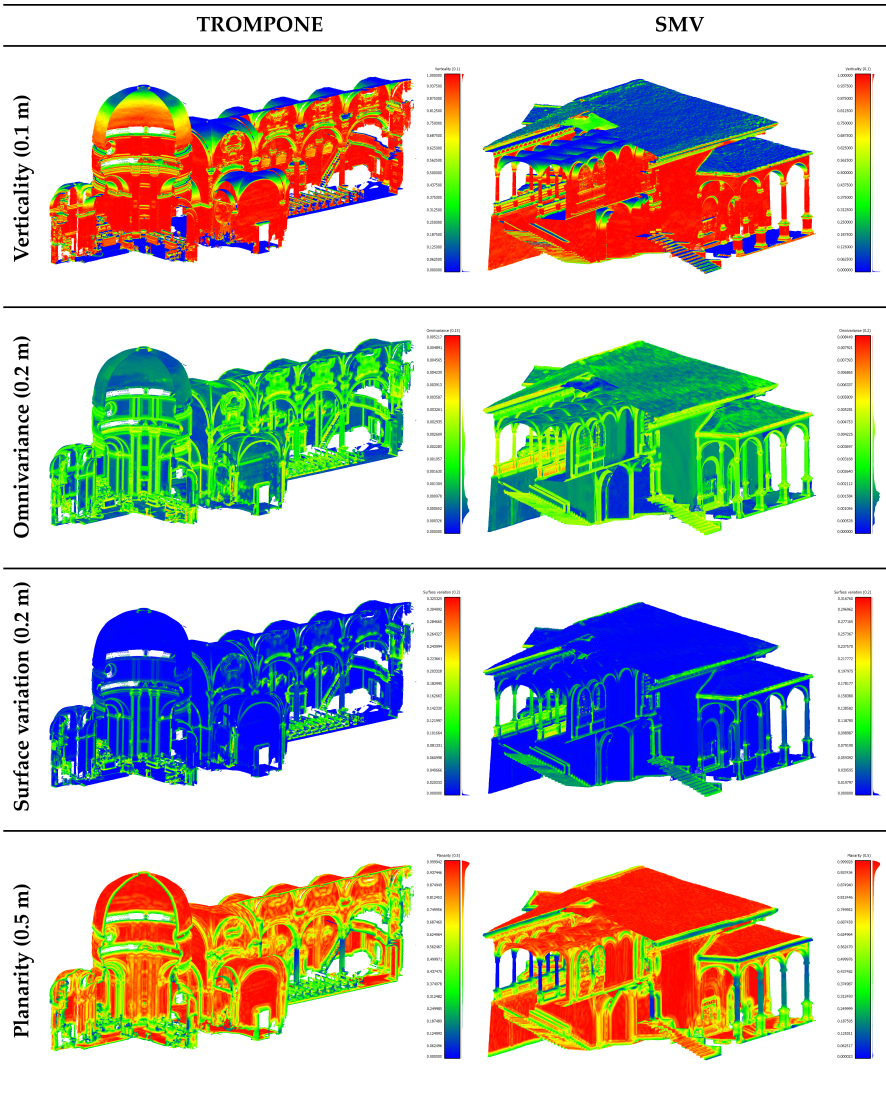
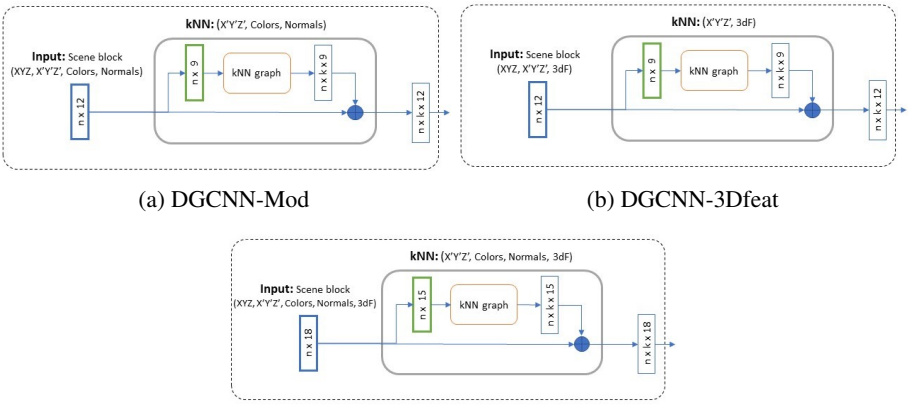


Figure 3.6: Three-dimensional features used to train the ML and DL classifiers. The colour of the plot represents the feature scale. The used search radii are reported in brackets.

Deep Learning approaches

The choice of Deep Learning methods obviously focused on the DGCNN-Mod approach presented in [69]. As already described in Section 3.1.1, this implementation includes several improvements, compared to the DGCNN original version: in the input layer, kNN phase considers coordinates of normalised points, color features transformations like HSV, and normal vectors. Moreover, the performance of the DGCNN-

Mod is compared with two novel versions of this network: the DGCNN-3Dfeat and the DGCNN-Mod+3Dfeat that take into consideration other important features aforementioned. In particular, the DGCNN-3Dfeat adds to the kNN the 3D features. Instead, for a complete ablation study the DGCNN-Mod+3Dfeat comprises all the available features. Figure 3.7 represents the configurations of the EdgeConv layer with the various feature combinations.



(a) DGCNN-Mod

(b) DGCNN-3Dfeat

(c) DGCNN-Mod+3Dfeat

Figure 3.7: Modified EdgeConv layer for DGCNN-based approaches.

Compared to the DGCNN-Mod, two types of pre-processing techniques are tested: Scaler1 and Scaler2. The Scaler1 standardises features by removing the mean and scaling to unit variance. The standard score of a sample x is determined as:

$$z = \frac{x - \mu}{\sigma} \tag{3.1}$$

where μ is the mean of the training samples and σ is the standard deviation of the training samples. Instead, Scaler2 scales features using statistics that are robust to outliers. This pre-processing phase removes the median and scales the data according to the quantile range (IQR: InterQuartile Range). The IQR is the range between the 1st quartile (25th quantile) and the 3rd quartile (75th quantile). Centering and scaling happen independently on each feature by computing the relevant statistics on the samples in the training set. Median and interquartile range are then stored to be used on the validation and test set.

In addition, the original DGCNN network uses the Cross Entropy Loss. Since the ArCH is really an unbalanced dataset, the author of this thesis decides to use Focal Loss [191] as well. This particular function has been implemented just to solve unbalance issues.

All deep learning approaches have been implemented using Python 3 and the well-known framework called Tensorflow. Pre-processing techniques on features, i.e.,

Scaler1 and Scaler2, have been implemented through the Scikit-Learn library [188], also implemented in Python.

Performance Evaluation Metrics

In the experimental section (Section 4.1.2), the employed state-of-the-art approaches are compared using the most common performance metrics for semantic segmentation. The Overall Accuracy (OA), along with weighted Precision, Recall and F1-Score are calculated regarding the test set, as these are very good performance indicators to understand if the approaches are able to generalise in a proper way. Please consider that OA and Recall have the same values, since the metrics are weighted.

It is worth noting that, in the scenes to be classified, the number of points varies according to the two approaches involved. In fact, with ML the total number of points both in the input and output scene are used, while with DL the unseen point cloud is subsampled with respect to the original one, for computational reasons. The number of subsampled points could be arbitrarily set, the most used is 4096 for each analysed block, but higher values can be chosen (e.g., 8192) at training time expense. In this work 4096 points per block have been set as subsampling parameter.

3.1.3 Point Clouds Generative Approaches

While the previous Section improve the architecture of the approach, in this one the author improve the proposed framework by solving data issues.

The author proposes a novel framework depicted in Figure 3.8 and published on [192]. The purpose of the proposed method is to improve the semantic segmentation of point clouds of an unbalanced dataset, generating new objects through generative approaches. Three generative networks have been trained for point cloud generation in the CH Domain.

The workflow starts with an initial pre-processing phase to give in input proper objects to the networks, and ends with a final phase for the comparison of two different training for the semantic segmentation of point clouds. Further details are given in the following subsections. The framework is comprehensively evaluated on the ArCH dataset.

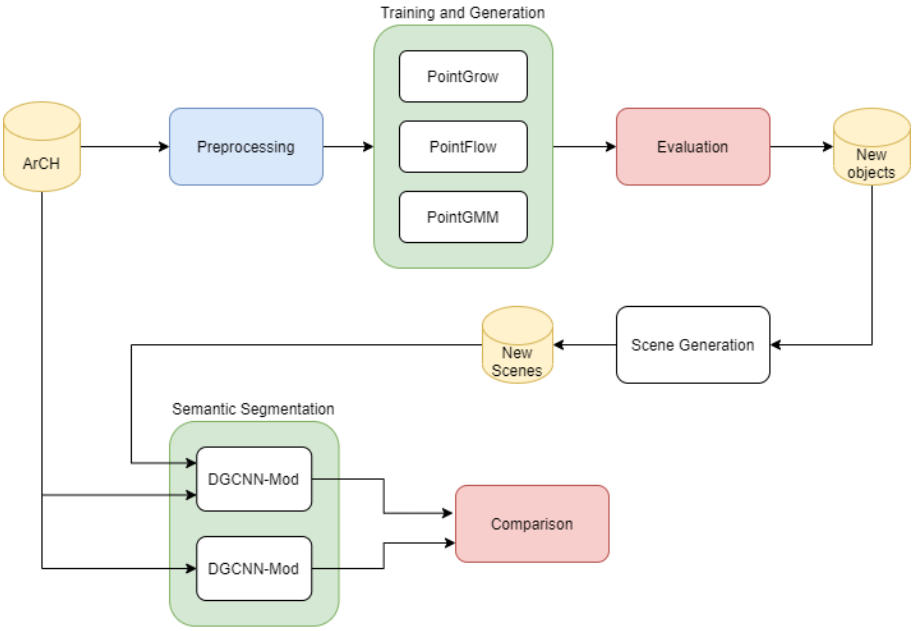


Figure 3.8: Workflow of the proposed method, based on generative approaches.

A subset of ArCH Dataset

The experiment section is based on the scenes of the ArCH benchmark [83]. This work is based on the tests carried out in [176], specifically the experiment of the Trompone's scene is reproduced using only the coordinates as a feature. Figure 3.9 shows this particular scene, representing original features and the relative ground truth.

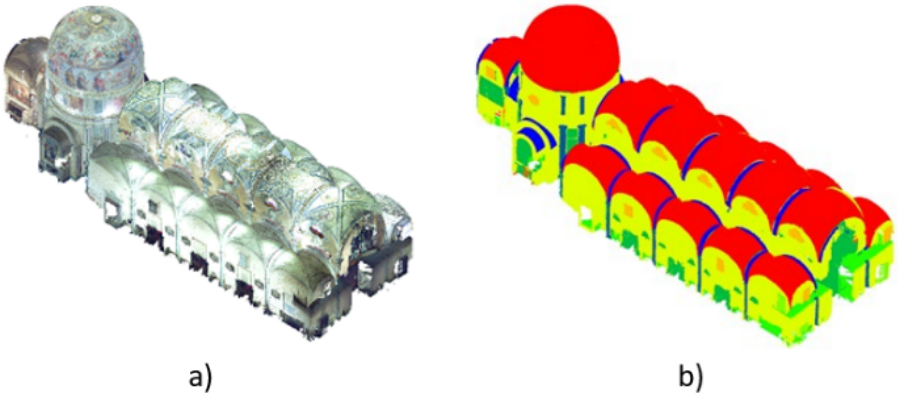


Figure 3.9: Trompone's scene from ArCH dataset: a) the scene features; b) the ground truth.

The objects of this dataset are given as input to generative approaches, in order to generate new objects and counterbalance the dataset. However, this dataset does not provide labels of the instances but only for the classes. Then, every single object has been manually divided. ArCH dataset provides the ground truth for 10 classes, but only the classes that are more difficult to recognize has been selected, as indicated in [176]. For this reason, only the objects regarding Column and Window classes have extrapolated from all scenes. The dataset used to train generative approaches comprises:

- 234 columns;
- 390 windows.

In the last phase of the workflow, ArCH dataset is used to compare two different trainings: the first one using the original dataset and the last one using the augmented dataset through the generated objects.

Generative Approaches for Point Cloud Generation

This subsection describes the generative approaches used to generate new ArCH dataset objects and counterbalance those classes that have only few instances. Before the training of the networks, a pre-processing phase is performed on the data, consisting of 2 steps: subsampling and data normalization. These networks can only take in input objects of 1024 points, so a random subsampling has been done for each instance of the dataset. Then, the objects are spatially centred at the (0,0,0) point and normalized to obtain values in the range of 0 and 1. The pre-processed data are then used to train three different generative networks: PointGrow [80], PointFlow [81], and PointGMM [82]. These networks have been chosen because they are very recent state of the art approaches and have very good performance in point clouds generation, which are then used to improve related tasks such as classification and segmentation.

The first tested approach is PointGrow [80], an autoregressive method for generating recurrently every point. This network estimates a conditional distribution of the point by considering all its preceding points. Taking into account the irregularity of the point clouds, the authors of this paper propose two point cloud-based self-attention modules for dynamically aggregating long-range dependencies from the other points. There are several ways to train these networks: in this thesis, the Unconditional PointGrow approach is used. To facilitate the generation process, training points are sorted according to their z coordinates; in this way, the shape should be encouraged to be generated mainly along its primary axis during the test phase.

The second network of the proposed framework is called PointFlow [81]. It is a variational auto-encoders (VAE) based approach, it is composed by three modules:

- An Encoder, that encodes a point cloud into a shape representation z;

- A Prior Module $P(z)$ over shape representations z;
- A Decoder $P(X|z)$ that models the distribution of points given the shape representation.

This particular network learns a two-level hierarchy of distributions: the first one is the distribution of shapes and the second one is the distribution of points given a shape. A continuous normalizing flow is used to learn these particular levels of distributions.

The third network, PointGMM [82], is also composed of encoders and decoders: the first receives point clouds as input and generates a features map as an output, the second build a GMM representation from the previous latent vector. GMM is a Gaussian mixture model, usually used as an alternative representation for 3D objects. This approach learns a hierarchical GMM (hGMM), performing coarse-to-fine learning to improve performance, instead of a common single scale GMM.

All three approaches take the input data from the preprocessing phase, generate new objects, and then are compared according to appropriate metrics.

Performance Evaluation Metrics

To evaluate the performance of the generative networks the following metrics have been adopted: Minimum Matching Distance based on Chamfer Distance (MMD-CD), Minimum Matching Distance based on Earth Mover's Distance (MMD-EMD) and Jensen-Shannon Divergence (JSD) [193]. First of all, Chamfer Distance (CD) and the Earth Mover's Distance (EMD) should be introduced: they are two symmetric distance metrics to measure the distance between two points clouds. In this work, the two points clouds are the original and the generated ones. Given two points clouds, S1 and S2, CD metric measures the squared distance between each point in S1 to its nearest neighbour in S2. So, the Chamfer Distance between S1 and S2 is defined in Eqs. 3.2 and 3.3:

$$d_{CD}(S_1,S_2) = \frac{1}{|S_1|}\sum_{x\in S_1}\min_{y\in S_2}\|x-y\|_2^2 + \frac{1}{|S_2|}\sum_{y\in S_2}\min_{x\in S_1}\|x-y\|_2^2 \tag{3.2}$$

Instead, the EMD is defined as:

$$d_{EMD}(S_1,S_2) = \min_{\phi:S_1\to S_2}\sum_{x\in S_1}\|x-\phi(x)\|_2 \tag{3.3}$$

where ϕ is a bijection between S1 and S2.

Then, a method to measure the similarity of A (set of generated point clouds) with respect to B (set of original point clouds) is needed. To this end, every point in B is matched with the closer one in A, by using minimum distance (MMD) and reporting the average of distances in the matching. CD and EMD metrics can be used as pointset distance for MMD, yielding MMD-CD and MMD-EMD [193]. The last comparison

metric is JSD, which measures if point clouds of A tend to occupy similar spaces as those of B, and the degree of this similarity. Given the empirical distributions (P_A, P_B), JSD metric is described as follow in Eq. 3.4:

$$JSD(P_A \,||\, P_B) = \frac{1}{2} D(P_A \,||\, M) + \frac{1}{2} D(P_B \,||\, M) \tag{3.4}$$

where $M = \frac{1}{2}(P_A + P_B)$ and $D(\cdot \,||\, \cdot)$ the KL-divergence between the two distribution.

Another state of the art method to evaluate the quality of the generated points clouds is to use PointNet [34] as a shape classifier [80]. In fact, if the generated objects will have very discriminating features for their relative class, it means that a classification model trained on the original objects should have good performance in classifying those generated, and vice-versa. In this work, after generating 100 points clouds per category, two classification activities are conducted:

1. Training on original data and testing on generated shapes;

2. Training on generated shapes and testing on original data.

3.2 Change detection on a dynamic space

This section begins to address the understanding of the dynamism of a space by defining its long-term changes. The author proposes a new framework to solve the Change Detection task from RGB image acquisition. The application field concerns the retail world and the problem of Out of Stock, strictly related to the concept of dynamic space. To test the proposed approach, the author acquired a new dataset of shelf images from a real store.

In the retail industry, the occurrence of Shelf Out of Stock (SOOS) situations is a significant problem. SOOS events are often strongly related to planogram design, where a planogram represents the way that stock keeping units (SKUs) are organised among the shelves [194]. The global average out-of-stock rate is about 8%, meaning that retailers have about 4% losses in sales. Out-of-stock situations happen for several reasons; the main one is defective shelf replenishment practices (surveying and restocking), which result in 70-90% of cases leading to SOOS. Another 10-30% result from problems in the supply chain, leading to store-OOS [195]. Promo activities also strongly influence shoppers' behaviour and result in SOOS with a strong impact on the overall retail turnover.

The method proposed by the author of this thesis is part of a broader approach, based on a robot called ROCKy and published in [130]. ROCKy, which stands for "Retail Out of Stock", is a low-cost mobile robot for detecting SOOS events both in real-time and on-demand, which has been described in [120]. In addition to the identification of SOOS and misplaced items, ROCKy can survey promotions and discounts,

Figure 3.10: TurtleBot robotic platform with cameras and UWB tags for localisation. The robot can navigate a retail environment and gather pictures to classify and localise SOOS and PA on grocery shelves.

model changes in shelf planogram (i.e., vertical product displacement) and store layouts, or monitor the warehouse during night time. The system can navigate the store using a modified potential field approach that tracks shopping carts to find the most visited areas. Also, the proposed approach enables retailers to analyse the store performance by comparing different shelf layouts and to address issues such as ease of selection, trading up, and overall shopping experience.

ROCKy consists of a TurtleBot, on-board RGBD camera for navigation, a low-power netbook for running fundamental algorithms, and a top-mounted GoPro HD personal cameras for shelf images collection. It makes use of six cameras to capture images and videos of the shelves on either side of the robot. These cameras take 12 MP photos (with a resolution of 4000×3000 and a horizontal FOV of 122.6 degrees) every five seconds. Six cameras are mounted on the top of ROCKy. The TurtleBot robotic platform and its cameras are shown in Figure 3.10 with a standard grocery retail environment used for testing and some shelf pictures that the robot collected during real world experiments.

ROCKy relies on a real-time locating system (RTLS) based on ultra-wideband (UWB) technology. The same localisation system is used to track human customers all day and to evaluate a store heat map. ROCKy starts from these grid-based heat map to move around, giving priority to hot zones (red areas) and using a potential field approach for its navigation. ROCKy captures images of the store's shelves during

buisness hours with a multiple-view camera to record consumer behaviours. These images are classified into 3 different categories and mapped in the grid-based store map for retail surveying:

- *SOOS*: Images of SOOS (incorrect scenario with high priority).
- *PA*: Images of the shelf with products and PA (scenario to be checked or updated with the store's promotional plan).
- *Normal*: Images of planogram in a standard layout (correct scenario).

To classify these pictures as SOOS, Normal, and PA, it is essential to judge both the visual elements and the included text at once. While a picture showing cookies with the phrase "Special Offer 50% off" is considered PA, the same picture containing the words "Gluten Free" might be considered Normal. These categories can be important indicators of the shelf availability: to monitor the daily situation, to measure SOOS at store level effectively and accurately, and to control and manage the total impact of promotions and offers. SOOS leads to disappointed customers, but the disappointment is even bigger when the customer goes to the store because of an advertised promoted product and does not find the offer on the shelf. The approach introduced in [196] to estimate the overall content of the images based on both visual and textual information is used in this work on the images acquired by ROCKy.

To classify the pictures, a machine learning classifier based on visual and textual features extracted from two specially trained Deep Convolutional Neural Networks (DCNNs) is implemented.

For the visual feature extractor, VGG-16 net [90], AlexNet [108], CaffeNet [117], GoogLeNet [118], and ResNet [119] with 50 layers and ResNet with 101 layers were used and applied to the whole image, trained by fine-tuning a model pre-trained on the ImageNet dataset. For the textual feature extractor, a DCNN architecture was used, proposed by [197] and created by fine-tuning a model that has been previously trained on synthesised planogram images. The DCNN's performance was compared with the ones of long short-term memory (LSTM) recurrent neural networks. However, before extracting text features, a text extraction and recognition phase is required.

Once both types of features have been extracted, six state-of-the-art classifiers, i.e., kNearest Neighbors (kNN) [198], [199], Support Vector Machine (SVM) [200], Decision Tree (DT) [201], Random Forest (RF) [202], Naive Bayes (NB) [203], and Artificial Neural Network (ANN) [204], were evaluated to classify the overall planogram image content.

The previously described classifiers are applied to the Shelf Management Assortment (SMART) Dataset, containing pictures acquired by ROCKy in a real retail environment during business hours. Both visual and textual elements concerning shelves in the targeted store were present in the dataset with a total of 14.244 images. Ground

truth has been manually evaluated by three human annotators to make it more reliable. The SMART Dataset is publicly available [2] for research purposes.

3.2.1 The proposed methodology

The approach presented in [196], i.e., the combination of the visual and textual features, has been used and extended for the development of the proposed system. The framework for joint visual and textual analysis, as well as the novel retail dataset (SMART Dataset) used for evaluation, was comprised of three main components: the visual feature extractor, the textual feature extractor, and the fusion classifier (see Figure 3.11). Two trained DCNNs were used for visual and textual feature extraction. Then, the two features were combined and fed into the fusion classifier. To estimate the overall content of the image, common machine learning algorithms were compared. Further details on the visual and textual feature extractor and fusion classifier are given in the following sections. Details on the data collection and ground truth labelling are discussed in Section 3.2.1.

SMART Dataset

The framework is comprehensively evaluated on the "Shelf Management Assortment" (SMART) Dataset, a visual and textual retail dataset made of pictures that ROCKy acquired during different experiments in different stores. The SMART Dataset is composed of 14.244 shelves images. SMART is the first dataset in that field built for these purposes and is publicly available [3]. As previously described, these images are divided into three categories, which include:

- *4,748 SOOS images*: Images of SOOS (incorrect scenario with high priority).
- *4,748 PA images*: Images of the shelf with products and PA (scenario to be checked with the promotional plan of the store).
- *4,748 Normal images*: Images of planogram in a standard layout (correct scenario).

The true content has been manually estimated to provide a more precise and less noisy dataset. All pictures are annotated with respect to their visual, textual, and overall content.

Figure 3.12 shows three examples of pictures in the SMART Dataset. Figure 3.12a is an example of SOOS situation, Figure 3.12b represents an image with normal shelf layout, and Figure 3.12c is a picture with promo activities. As can be seen, the overall content depends not just on the visual content of the picture, but also on the textual

[2] http://vrai.dii.univpm.it/content/smart-dataset
[3] https://vrai.dii.univpm.it/content/smart-dataset

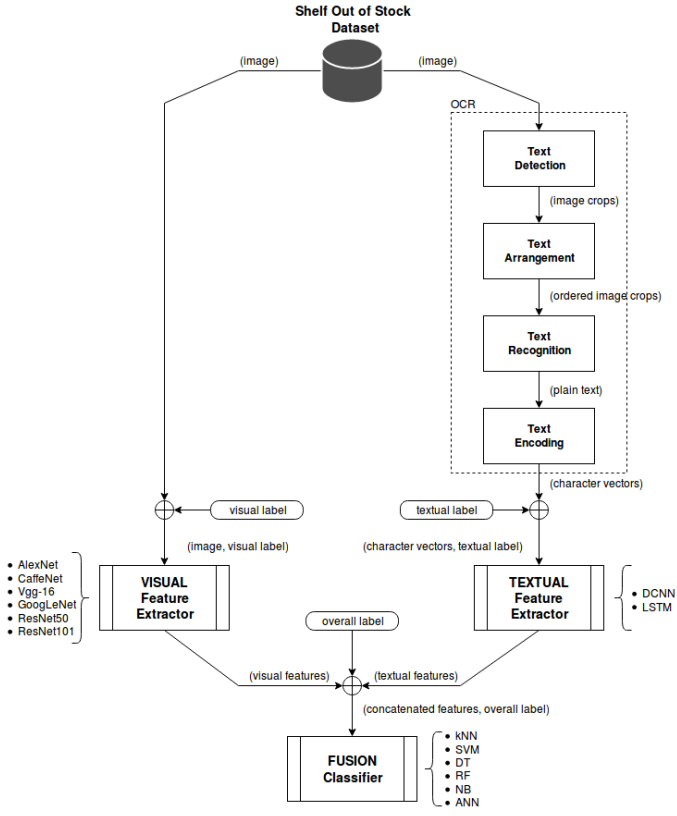


Figure 3.11: Deep learning visual and textual analysis workflow. The overall classification process mixed two different deep learning methods for visual and textual features, using a fusion classifier based on a classic machine learning approach to estimate three different classes (Normal, SOOS, and PA).

content. The current dataset is used to test and deploy the proposed methodology and compare different approaches. Further details will be presented in Section 4.2.

Visual Feature Extractor

The visual feature extractor provides information about the visual part of the picture. For this purpose, it is trained with image labels that indicate the visual category of the images. The training is performed by fine-tuning a DCNN. Different DCNNs were tested to chose the ones with the best performance: VGG-16 net [90], an AlexNet [108], a CaffeNet [117], a GoogleNet [118], and a ResNet [119] with 50 layers, and a ResNet with 101 layers. The DCNNs have been pre-trained on the ImageNet dataset [108] to classify images into 1,000 categories. The fine-tuning is done by cutting off the final classification layer and replacing it with a fully connected layer

(a) (b) (c)

Figure 3.12: Shelf pictures of SMART Dataset. Figure 3.12a is an example of SOOS situation, Figure 3.12b represents an image with normal shelf layout, and Figure 3.12c is a picture with promotion.

Table 3.3: Number of visual features extracted from the DCNNs

DCNNs	Number of visual features
VGG-16 [90]	4096
AlexNet [108]	4096
CaffeNet [117]	4096
GoogleNet [118]	1024
ResNet-50 [119]	2048
ResNet-101 [119]	2048

that has three outputs (one for each category class); the learning rate multipliers are increased for that layer. Loss and accuracy layers are adapted to take the input from the newly created final layer. The author used the SoftmaxWithLoss caffe layer, as it is conceptually identical to a softmax layer followed by a multinomial logistic loss layer, but provides a more numerically stable gradient. The output of the next to last layer is passed to the fusion classifier (fc7 layer for VGG-16, AlexNet, and CaffeNet; pool5 for GoogleNet and ResNet50). The image feature extractor is implemented using standard Caffe[4] tools. The number of visual features is illustrated in Table 3.3.

Textual Features Extractor

The textual feature extractor provides information about the textual category of a picture. It is trained with image labels that indicate the textual category of the images. Multiple components make up the textual feature extractor. The central component is a character-level CNN [197], extended for this analysis by one additional convolution layer. This extra layer, inserted before the last pooling layer, has a kernel size of 3 and produces 256 feature maps. Two training phases have been applied for the textual feature extractors: i) training a base model on synthesised planogram images, and ii) fine-tuning the base model on SMART dataset. The text must be transformed into

[4] http://caffe.berkeleyvision.org/

characters before being processed by the character-level DCNN, since it is embedded in the picture as pixels. To do so, the following steps have been performed:

1. *Text Detection*: Individual text boxes are detected in an image with the TextBoxes model [205].

2. *Text Arrangement*: Detected text boxes are put in order based on a left-to-right, top-to-bottom policy, thus forming logical lines.

3. *Text Recognition*: Each text box is processed by the OCR model [206] to transcribe the text of the box.

4. *Text Encoding*: The recognised text is encoded into one-hot vectors based on the alphabet of the character-level DCNN.

The textual features of the next to last layer of the character-level DCNN are passed to final fusion classifier.

In the state of the art regarding text classification, recurrent networks of the LSTM type are also very popular. For this reason, the author provides a comparison between DCNN and a recurrent network.

The performance of a LSTM model is evaluated using the Keras deep learning library[5]. Each sentence is mapped into a real vector domain, a technique that is called ”word embedding”, useful for processing text. This is a technique where words are encoded as real-valued vectors in a high dimensional space; the similarity between word meanings translates to closeness in the vector space [207]. Keras provides a convenient way to convert positive integer representations of words into a word embedding through an Embedding layer. Each word is mapped onto a 32 length real-valued vector. The total number of words interested in modelling was also limited to the 4,500 most frequent words and the rest were zeroed out. The sequence length (number of words) in each sentence varies, so each sentence is constrained to 20 words, truncating long sentences and padding the shorter sentence with zero values. Results are compared with those of character-level DCNN. The number of textual features extracted these models is illustrated in Table 3.4.

Fusion Classifier

Fusion classifier estimates the overall content of an image on the basis of the visual and textual features. Hence, the visual and textual features extracted from DCNN were pooled in the predictor vector and the machine learning model it is trained with the overall category of the images. Based on all features, six state-of-the-art classifiers were compared to recognise the overall content of the images: k-Nearest Neighbor (kNN), Support Vector Machine (SVM), Decision Tree (DT), Random Forest (RF),

[5]http://keras.io/

Table 3.4: Number of Textual features

Model	Layer	Number of features
DCNN [197]	ip4	1024
LSTM [207]	lstm	100

Naive Bayes (NB), and Artificial Neural Network (ANN). For what concerns the kNN, the euclidean distance has been employed as metric function. The Gaussian kernel was used for SVM. The author selected the optimal hyper-parameters for the machine learning methods (i.e., kNN: number of neighbours, SVM: kernel scale and box constraint, RF: number of weak classifiers, ANN: number of hidden layers), implementing a grid-search and optimising the F1-score in five-fold cross-validation within the training set. The testing performances were evaluated in terms of precision, recall, and F1-score.

3.3 Person Re-Identification on a dynamic space

This section addresses the second part of understanding space as a dynamic concept. More precisely, the author describes a new framework to study the dynamism of the most common entities that can be found in a space, namely people. A new deep learning method for person re-identification is proposed, tested both in closed world and in the more realistic open world domain. Moreover, the proposed approach is based on the acquisition of RGB-D videos, with a Top-View configuration. The approach is initially tested in the retail environment, in order to perform the finetuning of the hyperparameters. Subsequently, the author tests its generalisation ability in another context, regarding a museum. To test the proposed approaches, two new datasets concerning RGB-D video with a Top-View configuration were acquired. The first was acquired in a real store, and is publicly accessible, while the other was acquired from a multi-camera system in a museum. This methodology has been published on [158].

3.3.1 TVOW framework

This section introduces a new framework for person re-identification, called TVOW framework, as well as the dataset used for evaluation. The framework is depicted in Figure 3.13. The author uses a novel modified DCNN for re-ID that is composed of the following phases:

- *Data Acquisition*: The dataset is acquired through the use of an RGB-D camera.
- *Person Detection*: Using the depth channel, people can be detected.
- *Preprocessing*: By combining depth information with RGB information, the background is removed from the image and only the important information (the

person) remains.

- *Triplet Loss DCNNs*: Data augmentation techniques are used to fine-tune the networks, which are pretrained on the ImageNet dataset [208]. The triplet loss function is used for network training.
- *Evaluation*: Defining and evaluating specific metrics for this work.

Further details are given in the following subsections. The framework is comprehensively evaluated using the publicly available TVPR2 dataset.

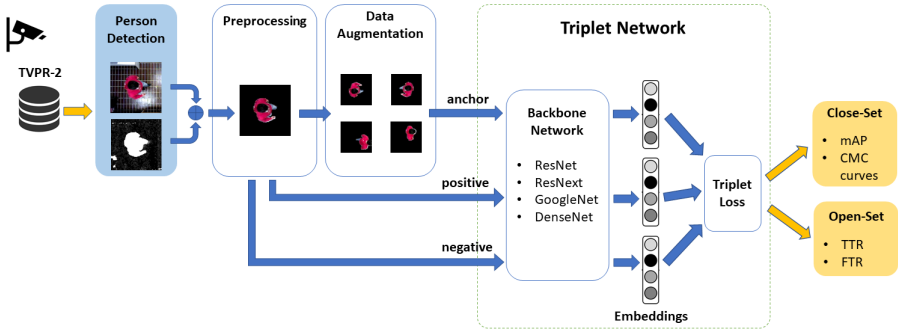


Figure 3.13: TVOW framework. Four phases are followed: Data acquisition, Person Detection, Data processing, Training of the Triplet Loss DCNN and performance evaluation.

TVPR2 Dataset

In this work, the author collected a new dataset for person re-ID called TVPR2 (Top-View Person Re-identification 2). It was acquired following the procedure outlined in [151], which is closer to realistic settings. The dataset comprises 235 videos containing RGB and depth information. Each person during a recording session walked with an average gait within the area under the camera in one direction, then it turned back and repeated the same route in the opposite direction. The number of people present in the videos also varies from one to eleven with the entire dataset comprising 1027 unique individuals. Figure 3.14 shows an example of frames extracted from both the RGB channel and the corresponding Depth channel. TVPR2 dataset[6] is publicly available for research purposes.

Preprocessing

The first problem to solve in a crowd environment is how to isolate individual people in each frame. Once isolated, one can proceed for extracting personal features and

[6]https://vrai.dii.univpm.it/content/tvpr2-dataset

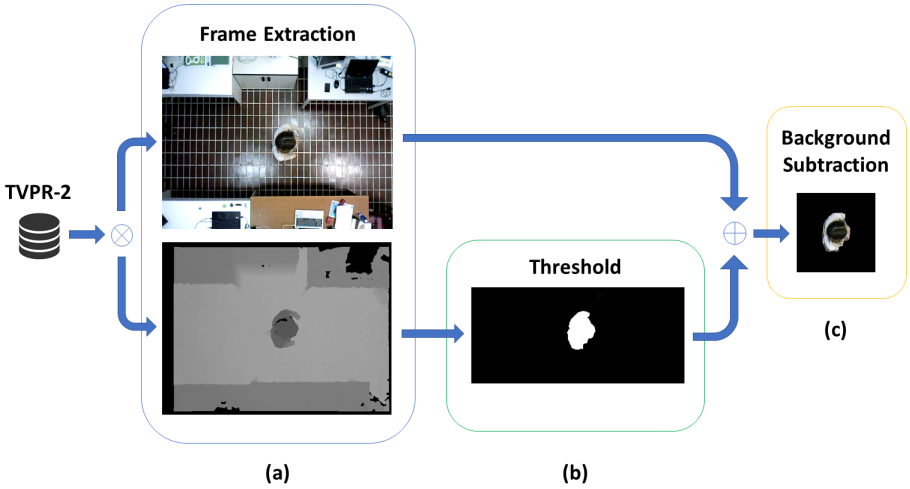


Figure 3.14: Preprocessing phase for the people detection task on an example frame of TVPR2 Dataset. (a) Frame Extraction for both streams. (b) Threshold on the Depth channel based on person's height. (c) Background subtraction by using the contour with the biggest area.

performing re-ID. Before using the dataset, individual frames were subjected to a preprocessing phase, shown in Figure 3.14. Firstly, RGB and Depth frames are extracted from their related streams, which was temporally and spatially synchronized. These have dimensions 320×240 pixels. A person detection algorithm made a crop of each person using a 150×150 pixel bounding box. This was made possible by using the depth channel and a threshold for a person's minimum height. In this way, noise produced by the frame background was removed to allow focusing on more important details (i.e., the person). The 150×150 pixel size was chosen experimentally, given the average dimensions of people in the dataset of between 80×80 and 125×125 pixels. As a further improvement, it was possible to use the depth information to remove the background inside the cropped image. This step was implemented using the previous mask to determine the outline with the largest area and then remove everything outside of that area. These cropped images were then used as input for the proposed deep learning method.

Triplet Loss DCNNs

The approach proposed by the author is based on a comparison of the most common state-of-the-art CNNs for image classification, used by the author as a backbone for the entire methodology. The choice of CNNs has focused on: ResNet-50 [91], ResNext-50 [209], DenseNet-161 [210], GoogleNet [211]. These backbones are pretrained using the ImageNet dataset, then fine-tuned on the TVPR2 dataset.

Before training the networks, data augmentation techniques were applied to increase the dataset and improve network performance. Subsequently, images were given as input to a DCNN. In this phase, various state-of-the-art networks were tested, pretrained on the public ImageNet dataset and then retrained on the TVPR2 dataset using the fine-tuning technique. Networks have been trained using a triplet hard loss. With this technique, the input image a (*anchor*) is transformed into a feature embedding space. An image p of the same class (defined as *hard positive*) is taken as an image n of a different class (defined as *hard negative*). The network is subsequently trained to bring the anchor a closer to the hard positive p while simultaneously moving it away from the hard negative n. For the triplet loss function, the batch hard function proposed in [156] has been used, designed for person re-ID tasks: authors show that, for both models trained from scratch or pretrained ones, using a well designed triplet loss can outperform most state-of-the-art methods. Batches of PK frames are created by randomly sampling P person IDs and K frames of each person. The triplet used to calculate the loss function is determined by selecting the hardest positive and the hardest negative samples within the batch for each sample a of the batch itself. The TVOW triplet loss is defined as follows:

$$L_{Triplet} = \overbrace{\sum_{i=1}^{P}\sum_{a=1}^{K}}^{all\ anchors} [m + \overbrace{\max_{p=1..K} D(a^i, p^i)}^{hardest\ positive} - \overbrace{\min_{\substack{j=1..P \\ n=1..K \\ j\neq 1}} D(a^i, n^j)}^{hardest\ negative}] \tag{3.5}$$

where the hard positive samples refer to poses of the same person in different frames and hard negative samples refer to similar-looking people.

Evaluation Metrics

In literature, there are few evaluation metrics for the open-set environment. The only existing ones were studied in [22] and are described below. Liao et al. [134] proposed calculating the cumulative matching curve (CMC) rate (usually used for close-set re-ID) at a fixed false accept rate to indicate the likelihood of misidentifying a person. In the work of Wang et al. [150], the false accept rate evaluation was used on two standard datasets for frontal person re-ID. According to [22], neither of these two studies worked well because the CMC metric is dependant on similar identity correspondence in a closed-set scenario. A completely different approach was used by Zheng et al. [133] and Zhu et al. [135] which adopted the True Target Rate (TTR) and False Target Rate (FTR) for evaluation in an open-set environment. The author of this thesis uses these metrics in the proposed approach. Several non-target persons should be placed in the probe population, the aim is not only to measure performance based on how well target probe persons are matched, but also how badly non-target

persons pass through the verification process. To evaluate the performance of different open-set environment methods, their measured TTR values will be compared.

To evaluate various approaches under a different verification standard, the author compare their TTR values against a series of given FTR values. TTR and FTR values can be described in the following equations:

$$TTR = \frac{N_{t2t}}{N_t}; \tag{3.6}$$

$$FTR = \frac{N_{nt2t}}{N_{nt}}; \tag{3.7}$$

where TTR is the number of accurate verifications N_{t2t} (target probe images are matched in the gallery) divided by the number of probe images from target persons N_t and FTR is the number of false verifications N_{nt2t} (non-target probe images treated as target persons) divided by the number of probe images from non-target persons N_{nt}.

Although TTR differs from the CMC rate, it also indicates the probability of the correct target, which means they can considered comparable to some extent.

For evaluation, according to [22], TTR values (with certain FTR values) are preferred over traditional CMC rates. TTR values can measure performance by verifying target and non-target persons, and are independent of one-to-one identity correspondence (a closed-set hypothesis).

TTR and FTR values must be calculated using the same test, and their purpose is to show how the network behaves in the presence of people unknown to it. The optimal result would be a high value of TTR with a low FTR value, showing that the network succeeded in correctly separating targets from non-targets.

A high FTR value, regardless of the corresponding TTR value, would mean the network had failed to exclude non-targets and subsequently assigned them identities of targets. To calculate pairs of these parameters, the following matching algorithm is used:

- Calculate the Euclidean distance $d(\widetilde{x^p}, \widetilde{x_i^g})$, between the probe $\widetilde{x^p}$ and all gallery vectors $\widetilde{x_i^g}$ defined as $i^* = argmin_i d(\widetilde{x^p}, \widetilde{x_i^g})$ considering the i^*th element of the gallery as the identity to assign to the probe.

- Given a threshold ϕ_m, we identify person $\widetilde{x^p}$ as target if $d(\widetilde{x^p}, \widetilde{x_{i^*}^g}) < \phi_m$. Otherwise, the person is a non-target.

- We consider a person a target when $d(\widetilde{x^p}, \widetilde{x_{i^*}^g}) < \phi_m$ and simultaneously $\widetilde{x_{i^*}^g}$ and $\widetilde{x^p}$ belongs to that person. Otherwise, $\widetilde{x^p}$ is treated as a non-target.

- The steps are repeated for each vector of probe.

- TTR and FTR values are calculated according to Equations 3.7.

3.3.2 A multimodal Person Re-Identification framework.

The approach described in the previous section is improved by the author using also the possibility of extracting further information obtained from the RGB-D videos acquired for the TVPR2 dataset. The basic idea is to use all types of information that can be extracted from the input data. Starting from RGB-D videos, the author can extract 3 types of information: visual data coming from the RGB colour, data regarding the depth acquired through the D channel, temporal data related to the acquisition of the video frames. This new methodology, called SeSAME is first tested on the TVPR2 dataset for finetuning hyperparameters and then used in a new domain, the museum, to test its ability to generalise.

SeSAME is based on a temporal multimodal deep learning approach[7] to extract the anthropometric and the appearance features from from RGB-D videos for RGB-D person re-ID. RGB-D images contain more information than RGB images, but there are two main issues to solve: how to combine these two modalities and how to extract efficient discriminative features from the depth channel. To convert depth images to RGB, the proposed approach is based on [212], and to integrate these two information uses, the proposed approach approach is based on [213]. Moreover, the author follows the work of [214] for video-based person re-ID, who used deep neural networks. Given the depth and RGB images, he considers a convolutional neural network (CNN) with two branches to extract their appearance features. The two branches will then be merged using a specific module to get overall image-level features as output. This approach will be tested on four different temporal modelling methods to aggregate a sequence of image-level features into clip-level features.

Following the procedure outlined in [166], a new dataset has been collected in a real museum environment, which is Palazzo Buonaccorsi, an historical building in Macerata, in Marche Region in the center of Italy. It is composed of videos that contained RGB and Depth channels. Each video recorded people on forward paths (left to right) for half the time and recorded the same people on return paths (right to left) the other half of the time, though not necessarily in that order. The results of person re-ID are used for evaluating important indicators and statistics as the time spent in each floor of Palazzo Buonaccorsi, the most visited floor of this building as well as the attention and the interaction with the artworks.

Figure 3.15 depicts the architecture of SeSAME, based on a re-id system adapted to a multi-camera museum environment.

As mentioned above, SeSAME is a vision-based person re-ID system for developing a tailored user experience in museum environment. SeSAME is based on multimodal information and it is able to work in a multi-camera environment. As Figure 3.16 shows, there are three main modules of the implemented system: a frame-level feature extractor, a temporal modelling and fusion module to aggregate previous

[7] https://github.com/vrai-univpm/temporal_reid

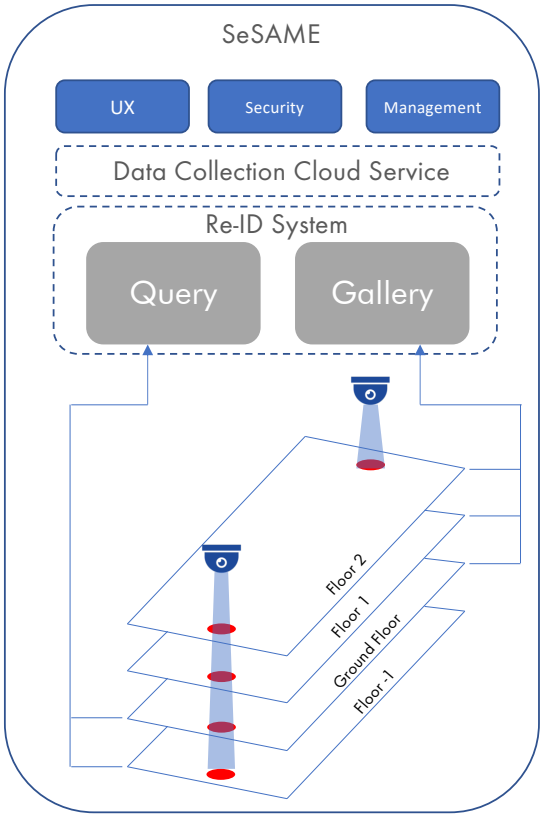


Figure 3.15: Workflow of the museum multi-camera system.

features, and a loss function [214].

The re-ID framework is comprehensively evaluated using TVPR2 [158] dataset. Four different temporal approaches was tested on this dataset. Finally, the best approach was chosen for SeSAME in order to perform users' re-ID for personalising the visit experience. SeSAME has been applied on a newly collected dataset acquired in a real museum environment, which is Palazzo Buonaccorsi, in the center of Italy. A detailed description of the data collection and ground-truth labelling is presented in the Section 3.3.2, including a preprocessing phase for the dataset.

Feature Extractor

A feature extractor typically employs a CNN. In this work, the author tested two types of CNNs: a 3D-CNN and a 2D-CNN with a temporal aggregation method. The first type takes a video as an input and gives a feature vector f_v of the entire video as an

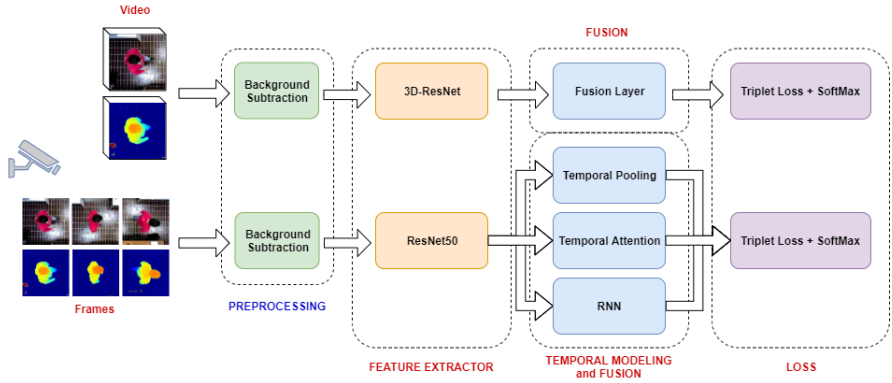


Figure 3.16: The proposed temporal multimodal framework comprised three main components: a feature extractor, a temporal modelling module and a loss function. The author tested 4 different methods based on this framework: a 3D-CNN, which does not need a temporal modelling method, and a 2D-CNN combined with three different temporal modelling modules. The last component was always a loss function designed to improve the network training. The dataset was initially processed in a preprocessing step to remove the backgrounds from the frames.

output, while a 2D-CNN takes a sequence of frames as an input, produces a sequence of frame-level features $\{f_v^t\}$, then aggregates the entire sequence in a feature vector f_v using a temporal aggregation method.

First, the 3D ResNet model proposed in [215] was tested. It was a 3D-CNN designed for the task of action classification formed with ResNet architecture [91] and 3D convolutional kernels. ResNet is a well-known architecture in the field of image classification, a residual network which implements skip connections, allowing deeper architecture while maintaining high performance.

The 3D ResNet model was pre-trained using a Kinetics dataset [216], a kinetics human-action video dataset. The final classification layer was replaced with an adapted layer to classify people based on the used dataset. The last layer before the classification was used as a feature-extractor, so its output was the representation of the recognised person.

For the 2D-CNN approach, a ResNet-50 model was tested. This network takes a sequence of frames as an input and gives a sequence of frame-level features $\{f_v^t\}$ as an output, which was fed into a temporal modelling module and produced the same output as the 3D-CNN.

For both CNNs, a second branch of the network was developed to extract depth features: the main idea is to duplicate the architecture of the RGB network and use it for other modalities, like the depth channel. The next step was to merge the feature maps extracted from each architecture, similar to the approach proposed in [212].

The merging layer is based on an element-wise summation of the features maps. Figure 3.17 shows the Resnet-50 with two branches and the final merging layer used in the 2D-CNN approach.

Another problem to solve was how to feed the network with depth frames. The network was designed to receive RGB images, that is, 3-channel images. Its depth duplicate must also receive images in that format. To convert depth images to RGB, the proposed approach was based on [212]. As a first step, all depth values was normalised to the range [0, 255] by choosing a threshold for the maximum value to compare with 255. Then, a jet colormap was applied to this value matrix, which transformed it from a single to a 3-channel matrix (colorising the depth). This method essentially mapped every distance value to a pixel RGB value, ranging from blue (small distance) to red (large distance). Figure 3.17 shows an example of a jet-encoded frame. According to [212], the jet colormap is the best of all those used to convert depth information, even better than the HHA [217] method, based on height above ground, horizontal disparity, and pixelwise angle between a normal surface and the direction of gravity.

Temporal Modelling and Fusion Module

Three different temporal modelling methods were tested:

- Temporal pooling
- Temporal attention [218][219]
- Recurrent neural network (RNN) [220][221].

Each method was tested by using the best parameter combination of [214]. The temporal pooling module performs the average pooling of N frames:

$$f_v = \frac{1}{N}\sum_{t=1}^{N} f_v^t \tag{3.8}$$

In the temporal attention model, temporal attention scores was computed for each frame using a temporal generation network formed by a spatial and a temporal convolutional layer. Then, after computing the final attention score a_v^t by a softmax function, an attention-weighted average was applied:

$$f_v = \frac{1}{N}\sum_{t=1}^{N} a_v^t f_v^t \tag{3.9}$$

Finally, the tested RNN module was formed by long short-term memory cells only. The RNN outputs $\{o^t\}$ were averaged to produce the final feature vector:

$$f_v = \frac{1}{N}\sum_{t=1}^{N} o_v^t \tag{3.10}$$

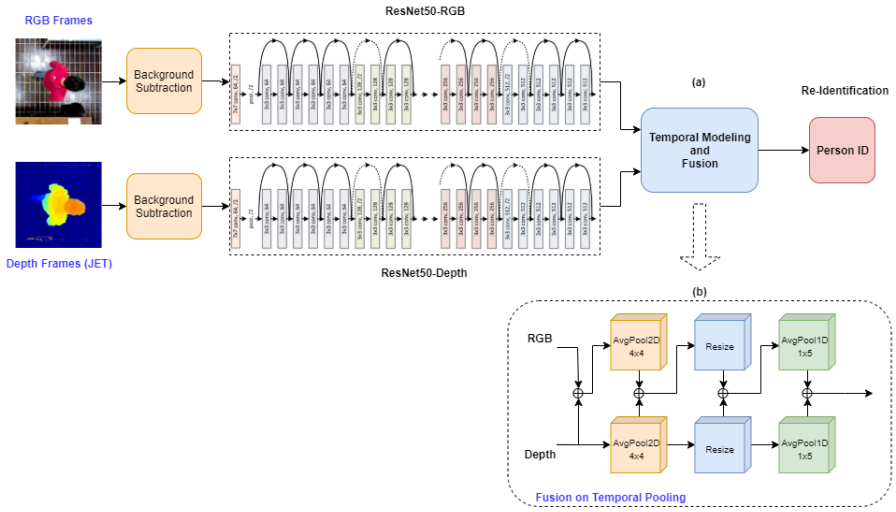


Figure 3.17: In the 2D-CNN approach, the Resnet50 was duplicated for extracting both RGB and depth features. (a) A temporal modelling and fusion layer were then used to aggregate the overall feature map. This framework was fed by RGB frames and depth frames encoded by jet colormap. (b) The fusion of both streams for the Temporal Pooling approach.

This module also included the merging phase between the RGB and depth features based on the approach of [213]: it contained two branches to extract temporal features from RGB and depth data, in addition depth feature maps were constantly fused into the RGB branch. In the proposed framework, the fusion layer was implemented as an element-wise summation, as demonstrated in Figure 3.17(b). This was the implementation of the temporal pooling approach. The fusion of the features was performed at four different points of the temporal approach: the first point was the output of the Resnet-50 (with the classification part removed), while the following points were those between the average pooling layers. The same method was used both for the temporal attention approach and for those based on the RNN. In the approach in which 3D-ResNet was used, there was no phase for temporal modelling, as it was already part of the 3D network. The fusion of the features was thus implemented in five main points within the network itself.

Loss Function

The networks were trained using a combination of a triplet loss function and a softmax cross-entropy function, as described in [214]. This is another improvement respect the TVOW framework described above.

The $L_{Triplet}$ function, described in Equation 3.5, was combined with a classic soft-

max loss function, which helped the network correctly recognise the person in the input frames. So, the final loss is calculated as follows:

$$L = L_{Triplet} + L_{Softmax} \tag{3.11}$$

Museum multi-camera system

Finally, the best person re-ID approach has been chosen for the multi-camera system of a museum. The museum system consists of 6 cameras on 4 different floors. Each camera detects the passage of a person using a threshold in the Depth channel frames. The saved frames also contain the timestamp of the detection, which is useful for obtaining important visitors statistics.

The network pretrained on TVPR2 was used as a feature extractor for people entering the building. These features are used to build the person gallery. The frames extracted from the RGB-D cameras are called query frames: the network extracts features from each of these frames, then compare them with all the gallery frames. The comparison is done by Euclidean distance. The result of the classification is the class of the gallery frame that has the smallest distance respect the query frame.

The architecture of SeSAME is showed on Figure 3.15.

Figure 3.18: Top-View Visitors' Museum Dataset: frame examples for every camera of the museum surveillance system.

Top-View Visitors' Museum Dataset

A dataset containing museum's visitors has been specifically collected for this work, with 6 RGB-D cameras placed in 4 different floors: one camera at the entrance of floor 0, one camera for the entrance and exit of floor -1, one camera for the entrance

and one for the exit of floor 1, one camera for the entrance and one for the exit of floor 2. The cameras record an RGB-D data flow with 640x480 pixel resolution. The proposed dataset consists of 55 days of video recordings from 10 July 2020 to 13 November 2020. People detected are 6200. Three days of registrations have been used as training set.

Figure 3.18 shows some frame examples for every camera of the museum surveillance system.

Chapter 4

Results

This chapter presents the results of Point Clouds Semantic Segmentation (see Section 4.1), Change Detection framework (see Section 4.2) and Person Re-Identification framework (see Section 4.3) for the Space Understanding tasks.

4.1 Point Clouds Semantic Segmentation framework.

This Section describes the results of all experiments concerning the proposed approaches for the understanding of space as a static concept.

4.1.1 DGCNN-Mod Network

In this section, the results of the first experiments conducted on ArCH Dataset are reported. In addition to the performance of the proposed DGCNN-Mod architecture, the author presents the performance of PointNet [34], PointNet++ [65], PCNN [67] and DGCNN [68] which form the basis of the improvement of the proposed method.

The experiments are separated in two phases. The first one attempt at tuning the networks, choosing the best parameters for the task of semantically segmenting the ArCH dataset. To this end the author considered a single scene and used an annotated portion of it for training the network, evaluating the performances on the remaining portion of the scene. He has chosen to perform such first experiments on the TR_church (Trompone) as it presents a relatively high symmetry, allowing to subdivide it in parts with similar characteristics, and it includes almost all the considered classes (9 out of the 10). Such an experimental setting addresses the problem of automatically annotating a scene that has been only partially annotated manually. While this has in fact practical applications, and could speed up the process of annotating an entire scene, the main goal is to evaluate the automatic annotation of an scene that was never seen before. The author addresses this more challenging problem, in the second experimental phase, where he trains the networks with 10 different scenes and then attempt at automatically segmenting the remaining one. Segmentation of the entire Point Cloud

into sub-parts (blocks) is a needed pre-processing step for all the analysed neural architectures. For each block a fixed number of points have to be sampled. This is due to the fact that neural networks need a constant number of points as input and that it would be computationally unfeasible to provide all the points at once to the networks.

Segmentation of Partially Annotated Scene

Two different settings have been evaluated in this phase: a k-fold cross-validation and a single splitting of the labeled dataset into a training set and a test set. In the first case the overall number of test samples is small and the network is trained on more samples. In the second case, an equal number of samples is used to train and to evaluate the network, possibly leading to very different results. For completeness, both settings were experimented.

In the first setting, the TR_church scene was divided into 6 parts and a cross-validation with 6 Fold was performed, as shown in Figure 4.1. The author has tested different combinations of hyperparameters of the various networks to be able to verify which was the best, as described in Table 4.2, where the mean accuracy is derived from calculating the accuracy of each test (fold), then averaging them.

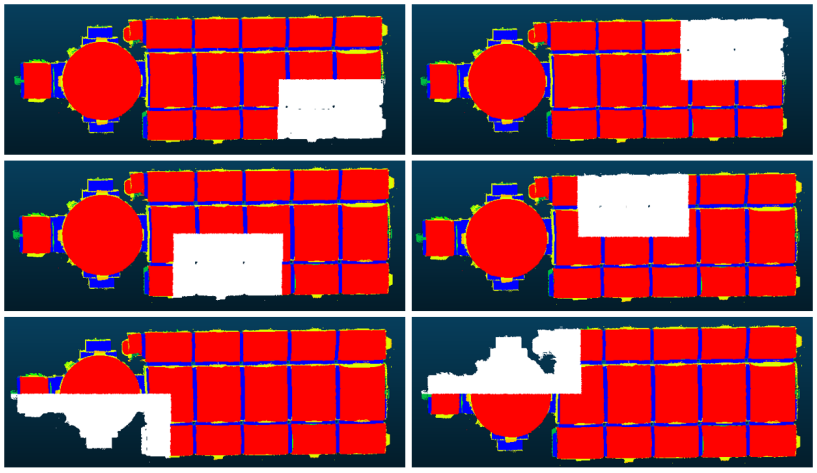


Figure 4.1: 6-Fold Cross Validation on the TR_church scene. The white fold in every experiment is the scene part used for the test.

Table 4.1: 6-Fold Cross-Validation on the Trompone scene. Different combinations of hyperparameters are used for the various state-of-the-art networks.

Network	Features	Mean Acc.
DGCNN [68]	XYZ+RGB	0.897
PointNet++ [65]	XYZ	0.543
PointNet [34]	XYZ	0.459
PCNN [67]	XYZ	0.742
DGCNN-Mod [69]	XYZ+Norm	0.781
DGCNN-Mod [69]	**XYZ+HSV+Norm**	**0.918**

Regarding the pointcloud preprocessing steps, which consists in segmenting the whole scene into blocks and, for each block, sampling a number of points, the author used, for each evaluated model, the settings described in the corresponding original paper. PointNet and PointNet++ used blocks are of size 2×2 meters and 4096 points for cloud sampling. In the case of the DGCNN, has been used blocks of size 1×1 and 4096 points per block. Finally, the PCNN network was tested using the same sampling as the DGCNN (1×1), but using 2048 points, as this is the default setting used in the PCNN paper. This network was also tested providing 4096 points per block, but results were slightly worse. The author also notices that the performances improve slightly using the color features represented as HSV color-map. The HSV (hue, saturation, value) representation is known for more closely aligning with the human perception of colors and, by representing colors as three independent variables, allows, for example, to take into account variations, e.g., due to shadows and different light conditions.

Table 4.2: 6-Fold Cross-Validation on the Trompone scene. Different combinations of hyperparameters are used for the various state-of-the-art networks.

Network	Block	Points	Features	Mean Acc.
DGCNN [68]	5x5	4096	XYZ+RGB	0.786
DGCNN [68]	1x1	4096	XYZ+RGB	0.897
PointNet++ [65]	2x2	8192	XYZ	0.701
PointNet++ [65]	2x2	4096	XYZ	0.543
PointNet [34]	2x2	8192	XYZ	0.611
PointNet [34]	2x2	4096	XYZ	0.459
PCNN [67]	5x5	2048	XYZ	0.693
PCNN [67]	1x1	2048	XYZ	0.742
DGCNN-Mod [69]	1x1	4096	XYZ+Norm	0.781
DGCNN-Mod [69]	1x1	4096	XYZ+HSV+Norm	**0.918**

In the second setting, the author splits the TR-cloister scene in half, choosing the left side for the training and the right side for the test. Furthermore, he splits the left side into a training set (80%) and a validation set (20%). The validation set is used to test overall accuracy at the end of each training epoch, then the author performed evaluation on the test set (right side). In Table 4.3, the performance of state-of-the-art networks are reported. The reported results are obtained with the hyperparameters combinations that best performed in the cross-validation experiment.

Table 4.3: The scene was divided into 3 parts: Train, Validation, Test. This table shows the average of the metrics calculated on the different parts: accuracy for Train, Validation and Test; precision, recall, F1-score and support for the Test.

Network	Train Acc.	Valid Acc.	Test Acc.	Prec.	Rec.	F1-Score	Supp.
DGCNN [68]	0.993	0.799	0.733	0.721	0.733	0.707	1,437,696
PointNet++ [65]	0.887	0.387	0.441	0.480	0.487	0.448	1,384,448
PointNet [34]	0.890	0.320	0.307	0.405	0.306	0.287	1,335,622
PCNN [67]	0.961	0.687	0.623	0.642	0.608	0.636	1,254,631
DCNN-Mod [69]	0.992	0.745	0.743	0.748	0.742	0.722	1,437,696

Table 4.4: The scene was divided into 3 parts: Train, Validation, Test. This table describes the metrics for every class, calculated on the Test set.

Network	Metrics	Arc	Col	Dec	Floor	Door	Wall	Wind	Stair	Vault
	Precision	0.484	0.258	0.635	0.983	0.000	0.531	0.222	0.988	0.819
	Recall	0.389	0.564	0.920	0.943	0.000	0.262	0.013	0.211	0.918
DGCNN [68]	**F1-Score**	0.431	0.354	0.751	0.963	0.000	0.351	0.024	0.348	0.865
	Support	69,611	36,802	240,806	287,064	8562	285,128	20,619	14,703	47,4401
	IoU	0.275	0.215	0.602	0.929	0.000	0.213	0.012	0.210	0.764
	Precision	0.000	0.000	0.301	0.717	0.000	0.531	0.000	0.000	0.654
	Recall	0.000	0.000	0.792	0.430	0.000	0.284	0.000	0.000	0.765
PointNet++ [65]	**F1-Score**	0.000	0.000	0.437	0.538	0.000	0.370	0.000	0.000	0.705
	Support	74,427	59,611	235,615	230,033	12,327	334,080	40,475	13,743	384,137
	IoU	0.000	0.000	0.311	0.409	0.000	0.215	0.000	0.000	0.681
	Precision	0.000	0.000	0.155	0.588	0.000	0.424	0.175	0.000	0.600
	Recall	0.000	0.000	0.916	0.422	0.000	0.078	0.004	0.000	0.387
PointNet [34]	**F1-Score**	0.000	0.000	0.265	0.492	0.000	0.132	0.008	0.000	0.470
	Support	30,646	11,020	29,962	43,947	1851	69,174	3212	1057	87,659
	IoU	0.000	0.000	0.213	0.406	0.000	0.051	0.003	0.000	0.311
	Precision	0.426	0.214	0.546	0.816	0.000	0.478	0.193	0.178	0.704
	Recall	0.338	0.474	0.782	0.754	0.000	0.231	0.012	0.188	0.744
PCNN [67]	**F1-Score**	0.349	0.294	0.608	0.809	0.000	0.281	0.021	0.306	0.779
	Support	65,231	32,138	220,776	212,554	8276	253,122	18,688	12,670	431,176
	IoU	0.298	0.273	0.592	0.722	0.000	0.210	0.010	0.172	0.703
	Precision	0.574	0.317	0.621	0.991	0.952	0.571	0.722	0.872	0.825
	Recall	0.424	0.606	0.932	0.920	0.002	0.324	0.006	0.284	0.907
DGCNN-Mod [69]	**F1-Score**	0.488	0.417	0.746	0.954	0.005	0.413	0.011	0.428	0.865
	Support	69,460	36,766	240,331	286,456	8420	285,485	20,542	14,790	475,446
	IoU	0.322	0.263	0.594	0.913	0.002	0.260	0.005	0.272	0.761

Table 4.4 shows the metrics in the test phase for each class of the Trompone's right side. This table reports, for each class, precision, recall, F1-Score and Intersection over Union (IoU). This table allows to understand which are the classes that are best

discriminated by the various approaches, to understand which are their weak points and their strengths (as broadly discussed in Section 4.1.1).

Finally, Figure 4.2 depicts the manually annotated test scene (ground truth) and the automatic segmentation results obtained with our approach.

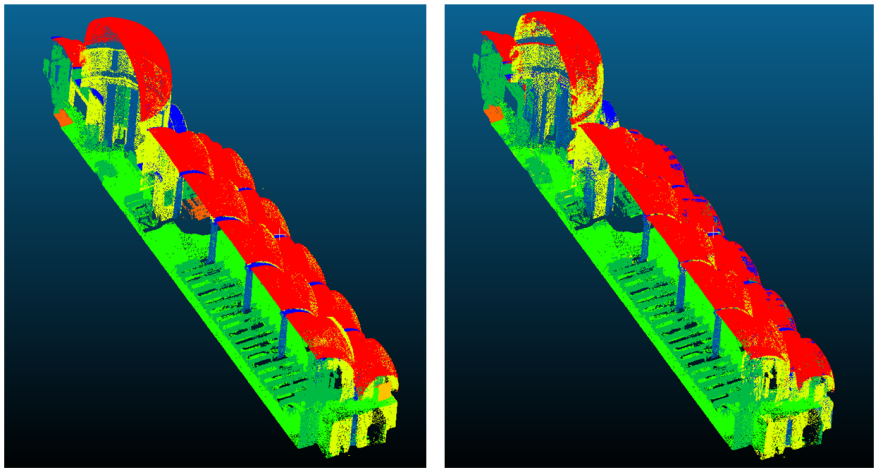


Figure 4.2: Ground Truth and Predicted Point Cloud, by using the proposed Approach on Trompone's Test side.

Segmentation of an Unseen Scene

In the second experimental phase, all the scenes of the first version of ArCH Dataset are used: 9 scenes are used for the Training, 1 scene as Validation (Ghiffa scene), 1 scene for the Final Test (SMV). State-of-the-art networks are evaluated, comparing the results with the DGCNN-Mod approach. In Table 4.5, the overall performances are reported for each tested model, while Table 4.6 reports detailed results on the individual classes of the test scene.

Table 4.5: Results of the tests performed on an unknown scene, training the network on the others.

Network	Valid Acc.	Test Acc.	Prec.	Rec.	F1-Score	Supp.
DGCNN [68]	0.756	0.740	0.768	0.740	0.738	2,613,248
PointNet++ [65]	0.669	0.528	0.532	0.528	0.479	2,433,024
PointNet [34]	0.453	0.351	0.536	0.351	0.269	2,318,440
PCNN [67]	0.635	0.629	0.653	0.622	0.635	2,482,581
DGCNN-Mod [69]	0.831	0.825	0.809	0.825	0.814	2,613,248

The performance gain provided by the proposed approach is more evident than in

previous experiments, leading to an improvement of around 0.8 in overall accuracy as well as in F1-score. The IoU also increases. In Table 4.6, it can be seen that the proposed approach outperforms the others in the segmentation of almost all classes. For some classes values of Precision and Recall are lower than the original DGCNN. However, DGCNN-Mod generally improves performance in terms of F1-score. This metric is a combination of Precision and Recall, thus it allows to better understand how the network is learning.

Table 4.6: Tests performed on all scenes of the dataset in terms of Precision, Recall, F1-Score and Support of each class for the Test scene.

Network	Metrics	Arc	Col	Dec	Floor	Door	Wall	Wind	Stair	Vault	Roof
	Precision	0.135	0.206	0.179	0.496	0.000	0.745	0.046	0.727	0.667	0.954
	Recall	0.098	0.086	0.407	0.900	0.000	0.760	0.007	0.205	0.703	0.880
DGCNN [68]	**F1-Score**	0.114	0.121	0.249	0.640	0.000	0.752	0.012	0.319	0.684	0.916
	Support	54,746	37,460	71,184	182,912	2642	642,188	18,280	172,270	288,389	1,143,177
	IoU	0.060	0.064	0.142	0.470	0.000	0.603	0.006	0.190	0.520	0.845
	Precision	0.000	0.000	0.124	0.635	0.000	0.387	0.000	0.000	0.110	0.738
	Recall	0.000	0.000	0.002	0.012	0.000	0.842	0.000	0.000	0.091	0.639
PointNet++ [65]	**F1-Score**	0.000	0.000	0.004	0.023	0.000	0.530	0.000	0.000	0.099	0.685
	Support	52,866	49,826	88,578	161,741	3032	756,905	26,682	165,169	245,929	882,296
	IoU	0.000	0.000	0.002	0.009	0.000	0.514	0.000	0.000	0.074	0.608
	Precision	0.000	0.000	0.240	0.763	0.000	0.299	0.000	0.000	0.298	0.738
	Recall	0.000	0.000	0.001	0.354	0.000	0.984	0.000	0.000	0.566	0.106
PointNet [34]	**F1-Score**	0.000	0.000	0.001	0.484	0.000	0.458	0.000	0.000	0.391	0.186
	Support	51,280	46,836	85,920	155,271	2880	726,628	25,614	158,562	236,091	829,358
	IoU	0.000	0.000	0.001	0.294	0.000	0.411	0.000	0.000	0.337	0.094
	Precision	0.119	0.181	0.143	0.441	0.000	0.633	0.041	0.582	0.580	0.801
	Recall	0.086	0.070	0.330	0.783	0.000	0.608	0.006	0.164	0.605	0.783
PCNN [67]	**F1-Score**	0.103	0.108	0.217	0.544	0.000	0.654	0.010	0.268	0.616	0.824
	Support	52,008	35,587	67,624	173,766	2509	610,078	17,366	163,656	273,969	1,086,018
	IoU	0.072	0.062	0.198	0.482	0.000	0.581	0.004	0.082	0.468	0.658
	Precision	0.288	0.391	0.270	0.798	0.000	0.729	0.035	0.707	0.806	0.959
	Recall	0.107	0.157	0.173	0.806	0.000	0.868	0.010	0.692	0.810	0.940
DGCNN-Mod [69]	**F1-Score**	0.156	0.224	0.211	0.802	0.000	0.791	0.015	0.699	0.808	0.950
	Support	54,746	37,460	71,184	182,912	2642	642,188	18,280	172,270	288,389	1,143,177
	IoU	0.085	0.126	0.118	0.669	0.000	0.655	0.008	0.538	0.678	0.905

Figure 4.4 and Figure 4.3 depict the confusion matrix and the segmentation results of the last experiment: 9 scenes for Training, 1 scene for validation and 1 scene for test.

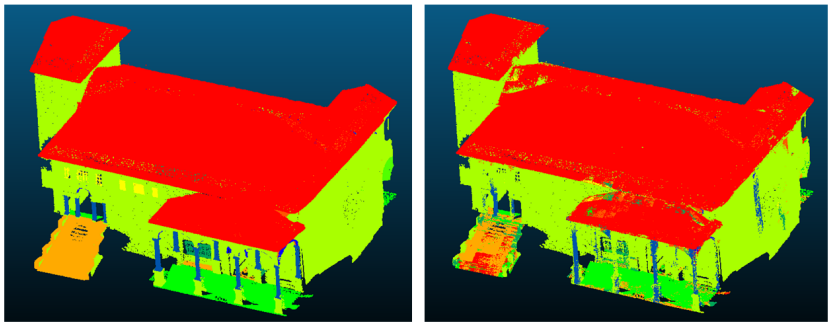


Figure 4.3: Ground truth (a) and predicted Point Cloud (b), by using our approach on the last experiment: 9 scenes for Training, 1 scene for Validation and 1 scene for Test.

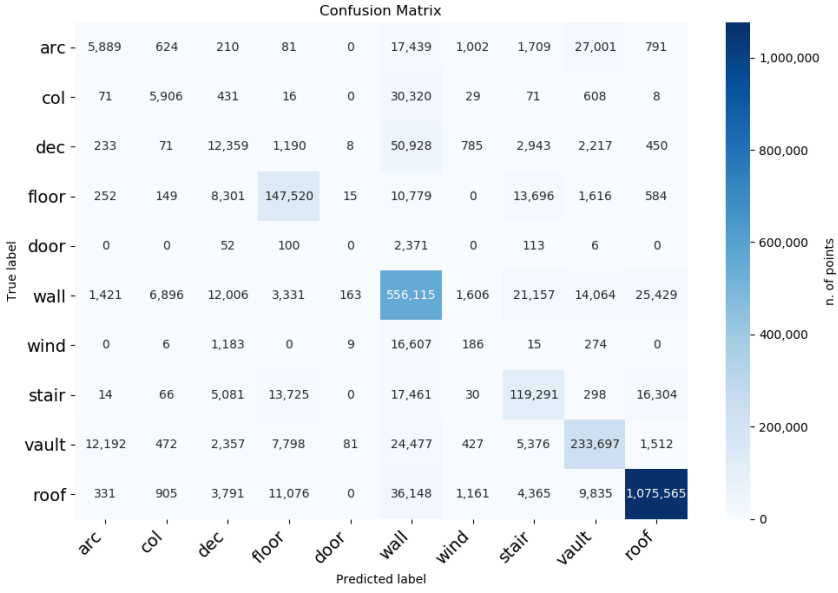


Figure 4.4: Confusion matrix for the last experiment: 9 scenes for Training, 1 scene for Validation and 1 scene for Test. The darkness of cells is proportional to the number of points labeled with the corresponding class.

Results Analysis

This research rises remarkable research directions (and challenges) that is worth to deepen. First of all, looking at the first experimental setting, performances are worse than those obtained in the K-fold experiment (referring to Table 4.3). This is probably do to the fact that the network has less points to learn on. As in the previous exper-

iment, the results on the test set is obtained with the proposed approach, confirming that HSV and Normals does in fact help the network to learn higher level features of the different classes. Besides, as reported in Table 4.4 and confirmed in Figure 4.2, it can be noticed that using the proposed setting helps in detecting *vaults*, increasing precision, recall and IoU, as well as *columns* and *stairs*, by sensibly increasing recall and IoU.

Dealing with the second experimental setting (see Section 4.1.1), it is worth to notice that all evaluated approaches fail in recognizing classes with low support, as *doors*, *windows* and *arcs*. Beside, for these classes it can be observed a high variability in shapes across the dataset, this probably contributes to the bad accuracy obtained by the networks.

More insights can be drawn from the confusion matrix, shown in Figure 4.4. It reveals, for example, that *arcs* are often confused with *vaults*, as they clearly share geometrical features, while *columns* are often confused with *walls*. The latter behaviour can be possibly due to the presence of half-pilasters, which are labeled as columns but have a shape similar to walls. The unbalanced nature of the number of points per class is clearly highlighted in Figure 4.5.

Furthermore, if the classes are individually considered, it can be seen that the lowest values are in Arc, Dec, Door and Window. More in detail:

- Arc: the geometry of the elements of this class is very similar to that of the vaults and, although the dimensions of the arcs are not similar to the latter, most of the time they are really close to the vaults, almost a continuation of these elements. For these reasons the result is partly justifiable and could lead to the merging of these two classes.

- Dec: in this class, which can also be defined as “Others” or “Unassigned”, all the elements that are not part of the other classes (such as benches, paintings, confessionals, etc.) are included. Therefore it is not fully considered among the results.

- Door: the null result is almost certainly due to the very low number of points present in this class (Figure 4.5). This is due to the fact that, in the proposed case studies of CH, it is more common to find large arches that mark the passage from one space to another and the doors are barely present. In addition, many times, the doors were open or with occlusions, generating a partial view and acquisition of these elements.

- Window: in this case the result is not due to the low number of windows present in the case study, but to the high heterogeneity between them. In fact, although the number of points in this class is greater, the shapes of the openings are very different from each other (three-foiled, circular, elliptical, square and rectangular)

(Figure 4.6). Moreover, being mostly composed of glazed surfaces, these surfaces are not detected by the sensors involved such as the TLS, therefore, unlike the use of images, in this case the number of points useful to describe these elements is reduced.

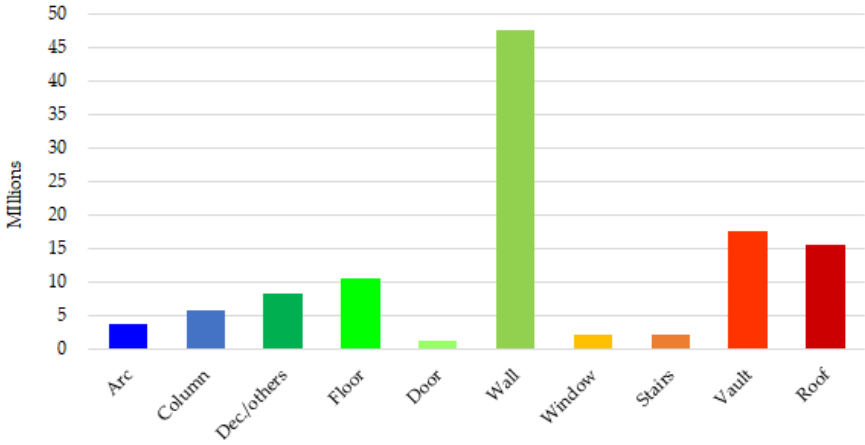


Figure 4.5: Number of points per class.

Figure 4.6: Different typologies of windows and doors. For the latter, their opening has sometimes affected the points acquisition.

Generalisation ability on indoor scenes

The proposed approach was also tested with the indoor dataset S3DIS, in order to verify its generalisation ability. The dataset contains point clouds described by spatial coordinates and RGB values only. In order to use the DGCNN-Mod network,

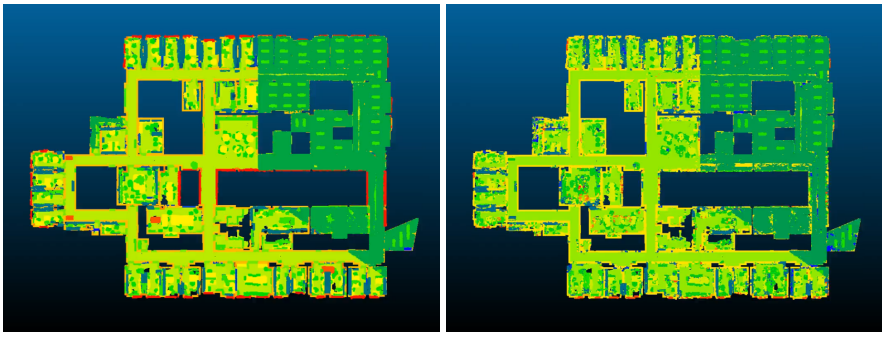
the author calculated the value of the normals vectors using the CloudCompare software. Subsequently, the DCNN-Mod network was compared with its original version, the DGCNN. In this way, it can be verified again whether the addition of additional features, i.e. normals, results in better performance.

Table 4.7: Results of the tests performed on S3DIS dataset.

Network	Train Acc.	Test Acc.	Prec.	Rec.	F1-Score	Supp.
DGCNN [68]	0.974	0.705	0.698	0.705	0.667	1,126,400
DGCNN-Mod [69]	0.973	0.723	0.715	0.723	0.692	1,126,400

The networks were trained for 100 epochs, using 5 areas as the training set and the remaining area as the test set. Table 4.7 shows the results in terms of Train Accuracy, and then Test Accuracy, Precision, Recall and F1-Score. It can be seen that the addition of the normals to the network improved performance in all the metrics described.

Figure 4.7 also shows the qualitative results of the semantic segmentation relative to the DGCNN-Mod. It can be seen that the predictions are very similar to ground truth. The errors in the predictions mainly concern the most unbalanced classes.



(a) Ground Truth (b) Prediction

Figure 4.7: Test Scene of S3DIS dataset. (a) Ground Truth (b) Prediction of DGCNN-Mod.

4.1.2 A mixed approach

This section describes several experiments performed with the ML and DL methods presented in Section 3.1.2, included the novel mixed approach proposed by the author. The experiment described in Section 4.1.2 regards the segmentation of the Trompone symmetrical scene, starting from the partial annotation of the same scene. In the second and third experiments, the training samples change according to the adopted

classification strategy (ML or DL). Still, the same scenes are tested: SMV scene for Subsection 4.1.2 and SMG scene for Section 4.1.2.

Segmentation of a Partially Annotated Scene

In this setting, the Trompone scene is initially split into two parts, choosing one side for the training and the symmetrical one for the test. Then, the side used for the training phase is further split into training set (80%) and validation set (20%). The validation set is used to test the OA at the end of each training epoch while the evaluation is performed on the test set. For this test, 9 architectural classes have been considered. Unlike the next experiments (Sections 4.1.2 and 4.1.2), the class "Other" was used during the training as it could be uniquely identified with the furnishing of the church (mainly benches and confessionals). No points from the class "roof" were tested, this being an indoor scene.

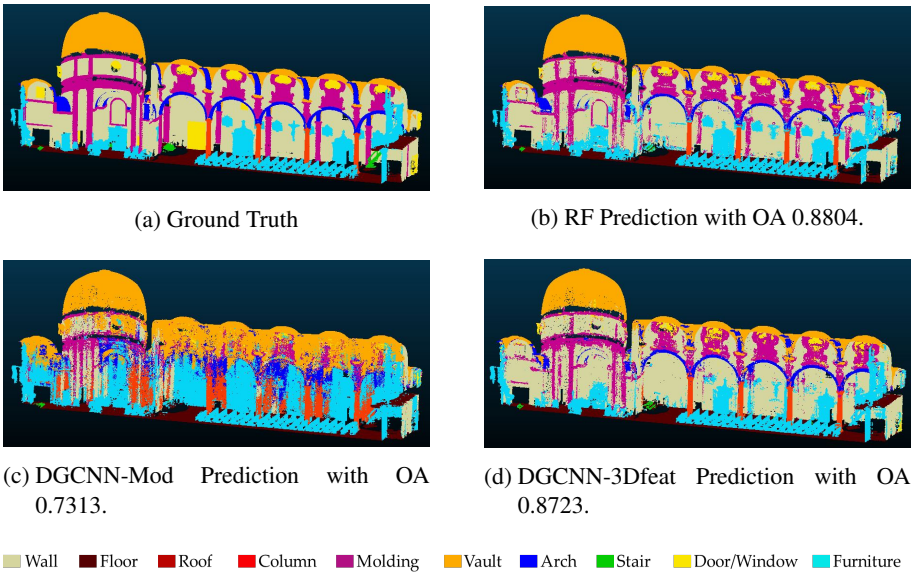
Original DGCNN uses its standard hyper-parameters: normalised XYZ coordinates for the kNN phase and XYZ + RGB for the feature learning phase, with 1×1 m block size. This latter parameter defines only the size of the block base, since the height is considered "endless"; in this way the whole scene can be analysed and the lowest number of blocks is defined. For the other DGCNN-based approaches it has been used the Scaler1 pre-processing setting for the features, as it resulted to be the best configuration among all the various tests performed. In addition, for the DGCNN-Mod+3Dfeat network, the best result was achieved using Focal Loss function.

In Table 4.8, the performances of the state-of-the-art approaches are reported. As it can can seen, the best performance in terms of accuracy metrics come from the RF approach. In addition, the other approaches exceeding 0.80 of accuracy are DT, DGCNN-3Dfeat, and DGCNN-Mod+3Dfeat, which all have in common the use of the 3D features. It can, therefore, deduced that this type of features allows for an improvement of the original DGCNN performances as they are very representative for the classes under investigation.

Table 4.8: Weighted metrics computed for the Test set of the Trompone scene divided into 3 parts: Training, Validation, Test.

Model	Overall Accuracy	Precision	Recall	F1-Score
kNN [178]	0.7438	0.7337	0.7438	0.7345
NB [185]	0.6639	0.6406	0.6639	0.6364
DT [181]	0.8345	0.8313	0.8345	0.8312
RF [187]	0.8804	0.8796	0.8804	0.8754
DGCNN [68]	0.7117	0.7400	0.7117	0.7040
DGCNN-Mod [69]	0.7313	0.7344	0.7313	0.6963
DGCNN-3Dfeat [176]	0.8723	0.8705	0.8723	0.8676
DGCNN-Mod+3Dfeat [176]	0.8290	0.8271	0.8290	0.8215

Finally, Figure 4.8 depicts the manually annotated test scene (ground truth) and the automatic segmentation results, obtained with best approaches. From this visual result it can be noticed again the issues with the class Stair (in green), and Window-Door (in yellow) (e.g., in none of the approaches it has been possible to identify the door at the center of the scene).



(a) Ground Truth

(b) RF Prediction with OA 0.8804.

(c) DGCNN-Mod Prediction with OA 0.7313.

(d) DGCNN-3Dfeat Prediction with OA 0.8723.

Figure 4.8: Ground Truth and predicted point clouds, by using best approaches on Trompone's Test side.

Segmentation of an Unseen Scene, SMV

In the second and third experiments, as previously anticipated, the training samples change according to the classification strategy adopted (ML or DL). Moreover, based on the experience of [65], the class "Other" is excluded from the classification, as the objects included are too variegated and it would confuse the NN. The portion of scene used to train the different ML classifiers consists of 2 526 393 points out of 16 200 442 points (approx. 16%) (Figure 4.9), while for the NNs 12 scenes of ArCH dataset have been used according to the previous tests performed in [69].

Same state-of-the-art approaches as in the previous section are evaluated.

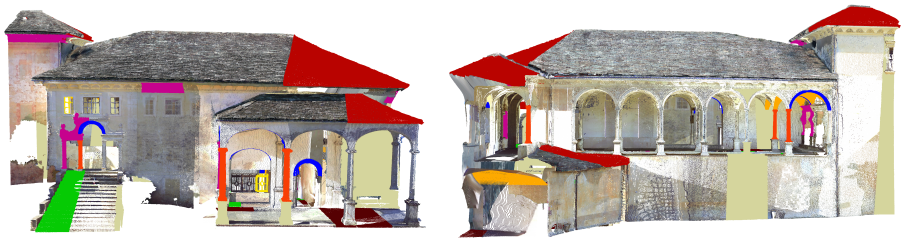


Figure 4.9: Manual annotations used to train the ML algortihms for the Sacro Monte Varallo (SMV) Scene.

In Table 4.9, the overall performances are reported for each tested model.

Original DGCNN is trained again using its standard hyperparameters. For the other DGCNN-based approaches, the best results were achieved using:

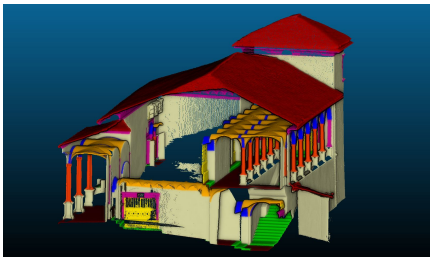
- Focal Loss for DGCNN-Mod;
- Scaler1 pre-processing for DGCNN-3Dfeat;
- Focal loss and Scaler2 pre-processing for DGCNN-Mod+3Dfeat;

Table 4.9 shows that DGCNN-Mod+3Dfeat is the best approach in terms of OA, reaching 0.8452 on the Test Scene, followed by the RF with 0.8369. The second best approach, on the contrary, gets better results on these classes, while maintaining an high average accuracy.

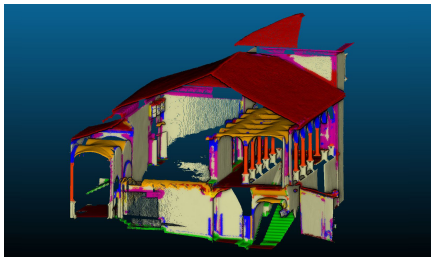
Table 4.9: Weighted metrics computed for the Test set of the SMV scene.

Model	Overall Accuracy	Precision	Recall	F1-Score
kNN [178]	0.8102	0.8588	0.8102	0.8248
NB [185]	0.7331	0.7970	0.7331	0.7584
DT [181]	0.8041	0.8522	0.8041	0.8180
RF [187]	0.8369	0.8736	0.8369	0.8467
DGCNN [68]	0.5608	0.6850	0.5608	0.5602
DGCNN-Mod [69]	0.8294	0.8216	0.8295	0.8192
DGCNN-3Dfeat [176]	0.7890	0.7776	0.7890	0.7720
DGCNN-Mod+3Dfeat [176]	0.8452	0.8287	0.8452	0.8343

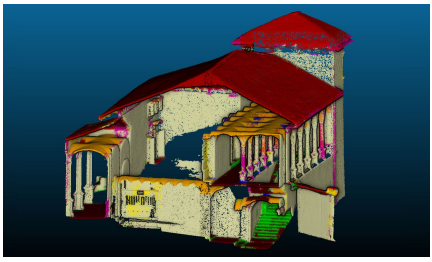
Figure 4.10 depicts the manually annotated test scene (ground truth) and the automatic segmentation results obtained with the best approaches. It is possible to notice that most of the classes have been well recognised, except for the Arch class in the DGCNN-based approaches and the Door-Window class for the RF.



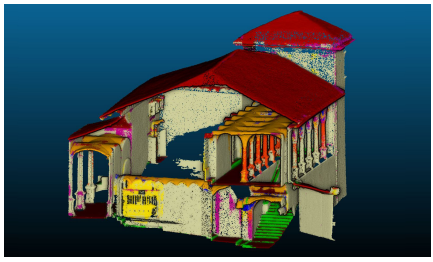
(a) Ground Truth



(b) RF Prediction with OA 0.8369.



(c) DGCNN-Mod Prediction with OA 0.8294.



(d) DGCNN-Mod+3Dfeat Prediction with OA 0.8452.

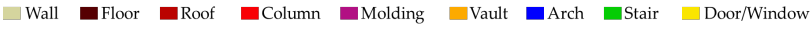


Figure 4.10: Section of Ground Truth (**a**) and the best Predictions (**b**–**d**) of the SMV scene. Please note that the point clouds deriving from the DL approach are subsampled.

Segmentation of an unseen scene, SMG

As in the previous experiments, for the ML approaches ad hoc annotations have been distributed along the point cloud (Figure 4.11), consisting of 3 545 900 points over a total of 17 798 049 points (approx. 20%).

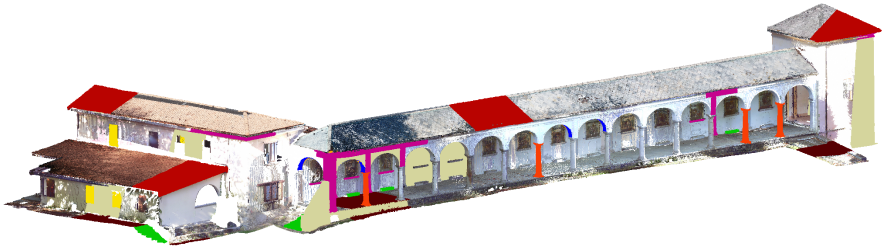


Figure 4.11: Manual annotations used to train the ML algortihms for the Sacred Mount of Ghiffa (SMG) Scene.

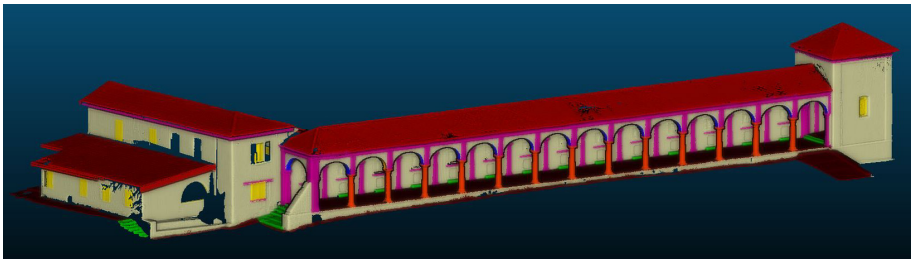
In Table 4.10, the overall performances are reported for each tested model. Best re-

sults have been achieved with RF, immediately followed by the DGCNN-Mod+3Dfeat network. However, in this case, given the higher symmetry of the point cloud, if compared to the SMV scene, the increase in OA when using the 3D features is lower, but still significant. Results are consistent with the previous test and the most problematic class is again the Door-Window, probably due to the dataset unbalance.

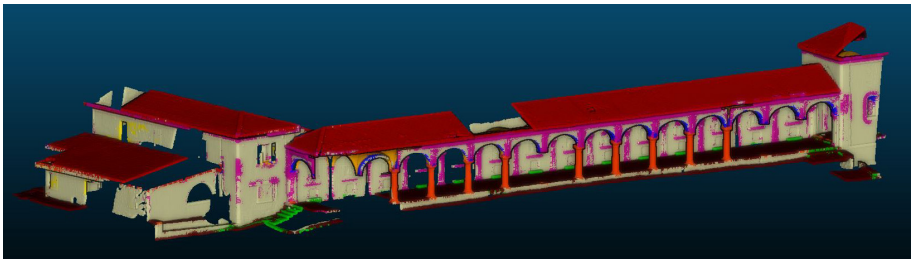
Table 4.10: Weighted metrics computed for the Test set of the SMG scene.

Model	Overall Accuracy	Precision	Recall	F1-Score
kNN [178]	0.6078	0.6565	0.6078	0.6262
NB [185]	0.7186	0.7967	0.7186	0.7422
DT [181]	0.8952	0.9014	0.8952	0.8971
RF [187]	0.9266	0.9239	0.9266	0.9243
DGCNN [68]	0.8514	0.8528	0.8514	0.8474
DGCNN-Mod [69]	0.8951	0.8887	0.8951	0.8860
DGCNN-3Dfeat [176]	0.8736	0.8887	0.8737	0.8776
DGCNN-Mod+3Dfeat [176]	0.9135	0.9165	0.9135	0.9125

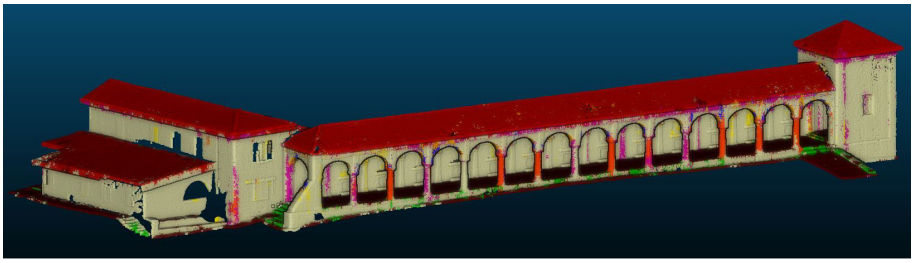
Finally, Figure 4.12 depicts the manually annotated test scene (ground truth) and the automatic segmentation results obtained with best approaches.



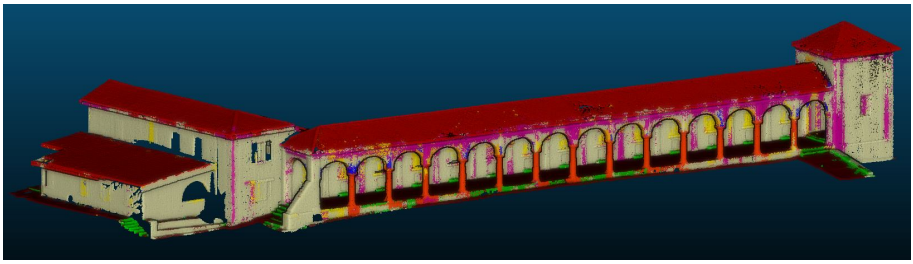
(a) Ground Truth



(b) RF prediction with OA 0.9266.



(c) DGCNN-Mod prediction with OA 0.8951.



(d) DGCNN-Mod+3Dfeat prediction with OA 0.9135.

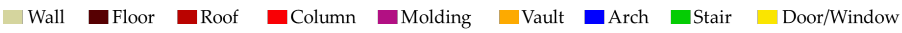


Figure 4.12: Ground Truth (**a**) and the best Predictions (**b**–**d**) of the SMG scene. Please note that the point clouds deriving from the DL approach are subsampled.

Results Analysis

The recap of the best OA achieved (Figure 4.13) highlights that the Random Forest method is slightly better in the two almost symmetrical scenes of Ghiffa and the Trompone church. In these cases, with manual annotation, it is possible to select a number of adequately representative examples of the test scene, ensuring an accurate result. The DL solutions, on the other hand, seem to work better in the non-symmetric scene, thus showing a good generalisation ability. More generally, the results of DL are satisfactory, as they demonstrate the achievement of OA similar to those of RF, although the training set is partially limited, if compared to the others present in the state of the art.

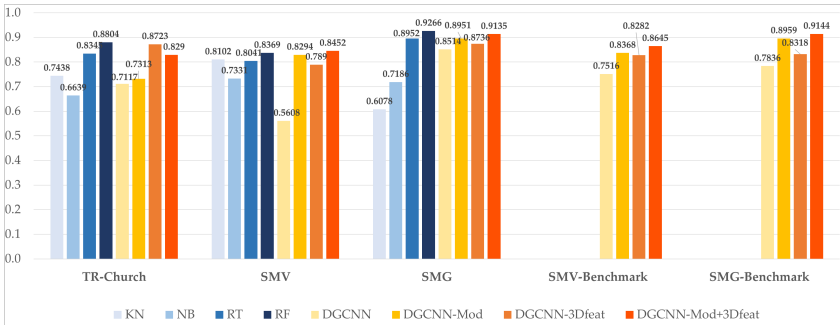


Figure 4.13: Overall Accuracy of all tests carried out.

Figure 4.14 shows the F1-Score, a combination of precision and recall, relative to the single classes. In this case, the ML approaches outperform DL for some classes such as Arch, Column, Molding and Floor, while the DL gives better results in the segmentation of Door-Window and Roof. The remaining classes of Vault, Wall and Stair are equally balanced between the results of the two techniques, with vaults and walls leaning towards the RF and stairs to the DGCNN-Mod+3Dfeat.

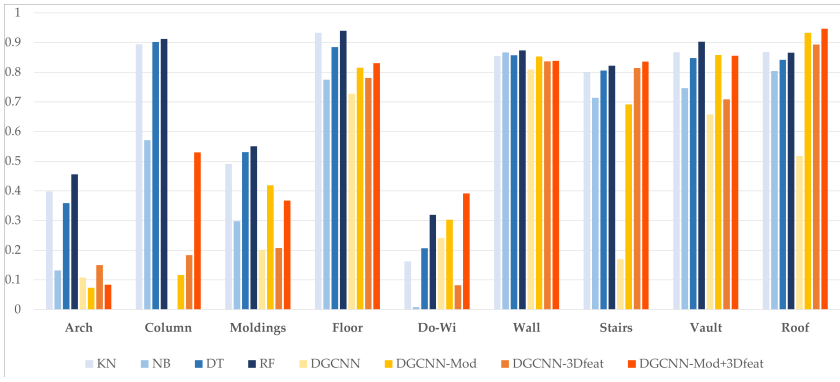


Figure 4.14: F1-Score of the different classes for the SMV scene with the different approaches.

4.1.3 Point clouds generative approaches

This section describes the configuration for the generative approaches and discusses the results of the various experiments reported in Section 3.1.3. The three generative networks are trained using Window and Column objects from the ArCH dataset. As described in Section 3.1.3, the objects are sub-sampled at 1024 points, then centred with respect to the point (0,0,0) and normalized with values between 0 and 1. The training was done by splitting the dataset as follows:

- 80% of the objects for training;
- 20% of the objects for the final test.

A part of the training was used as the validation set. The networks have been trained using the hyperparameters configuration described in the related papers. The first experiment performed concerns the training of generative networks. Table 4.11 shows the comparison of the approaches in terms of JSD, MMD-CD and MMCEMD metrics. The most satisfactory results are those obtained with PointFlow, which is the best approach for the generation of columns object, by obtaining lower JSD. Mand both MMD values. However, regarding windows objects, the two best methods for their generations seem to be PointGrow and PointGMM.

Class	Model	JSD	MMD-CD	MMD-EMD
Column	PointGrow [80]	0.1941	0.0090	0.1352
	PointFlow [81]	**0.0820**	**0.0078**	**0.1228**
	PointGMM [82]	0.0929	0.0080	0.1421
Windows	PointGrow [80]	0.2588	**0.0061**	**0.1030**
	PointFlow [81]	0.3014	0.0292	0.1613
	PointGMM [82]	**0.1704**	0.0079	0.1052

Table 4.11: Results of the generative approaches.

The second experiment regards the classification performed with PointNet. The results are shown in Table 4.12: even in this case, PointFlow approach turns out to be the best generative method, as the classification accuracy for both SG and GS tests remain congruent.

Model	SG	GS
PointGrow [80]	0.5733	0.9350
PointFlow [81]	0.7143	0.715
PointGMM [82]	0.6900	0.9200

Table 4.12: Classification accuracy using PointNet. SG: Training on ArCH Dataset and testing on generated shapes; GS: Training on generated shapes and testing on ArCH Dataset.

From the results obtained, the PointFlow network has been chosen to generate column objects, while the PointGrow network has been used to generate Window objects. Using CloudCompare software, it was possible to build 3 scenes containing these objects. To make the scene more realistic, portions of the wall were also added, so that the objects could be positioned consistently. Figure 4.15 shows the scenes created with the generated objects.

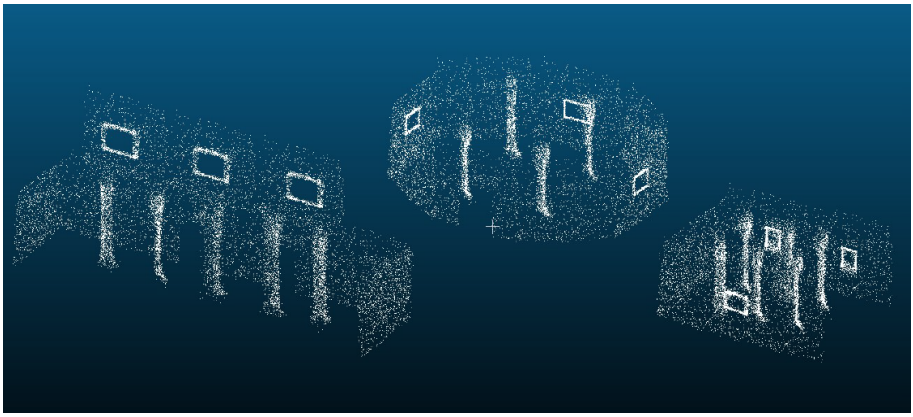


Figure 4.15: Scenes created by using generated objects.

Finally, these scenes are used to improve the training of the DGCNN-Mod network, for the semantic segmentation task on ArCH dataset. As said previously, experiments are conducted on the Trompone's scene. Two different experiments are performed: the first regards the training by using original dataset, while the last regards the training with the ArCh dataset augmented by the generated scenes. Table 4.13 and Table 4.14 show the results in terms of Precision, Recall and F1-Score, for every class of the Arch dataset. Performance of the DGCNN-Mod semantic segmentation on ArCH Dataset without the generated scenes are represented in Table 4.13. On the other hand, the performance of the DGCNN-Mod trained with the augmented dataset is shown in Table 4.14.

These experiments show that semantic segmentation performance regarding column class are improved. However, window segmentation has decreased. This is probably due to the fact that in the ArCH dataset windows are included in the same class as doors, because they are the classes with fewer elements. By increasing the number of windows, the network probably has more difficulty in classifying door points. These findings deserve a more detailed explanation. First, looking at the column class, it is worth noting that despite the precision value is higher "without" than "with" the generated scenes, F1-score and Recall increase in this latter case (see the comparison between Tables 4.13 and 4.14); such values provide comforting insights from the author's experiments, since they are meaningful and opens to future generalizations of the method. This is not true for the class Door-Win, but the reason shall is that

Object	Precision	Recall	F1-score
Arc	0.2904	0.1268	0.1765
Column	**0.9906**	**0.0856**	**0.1576**
Molding	0.5074	0.1465	0.2273
Floor	0.9485	0.9051	0.9263
Door-window	**0.0974**	**0.0467**	**0.0631**
Wall	0.5867	0.2339	0.3344
stairs	0.0424	0.0006	0.0011
Vault	0.9177	0.7408	0.8198
Forniture	0.3484	0.9241	0.5060

Table 4.13: Results of the DGCNN-Mod semantic segmentation on ArCH Dataset without the generated scenes.

Object	Precision	Recall	F1-score
Arc	0.2284	0.4445	0.3018
Column	**0.6692**	**0.2131**	**0.3232**
Molding	0.4460	0.2488	0.3194
Floor	0.9159	0.9808	0.9473
Door-window	**0.1207**	**0.0354**	**0.0547**
Wall	0.6646	0.5773	0.6179
stairs	0.3583	0.1789	0.2386
Vault	0.9380	0.7460	0.8310
Forniture	0.5556	0.8286	0.6652

Table 4.14: Results of the DGCNN-Mod semantic segmentation on ArCH Dataset with the generated scenes

this class is merged among two objects (namely doors and windows), while in the benchmark dataset such classes were considered separately.

It is possible to obtain a further comparison through the graph of Figure 4.16. In this graph, the comparison of metrics for the entire dataset is shown. The results show that the accuracy of semantic segmentation has also improved thanks to the inclusion of these new generated scenes. It is fairly straightforward to deduce the motivation behind this increase in accuracy; data generation has been performed for those classes with a lower number of points available, that is to say unbalanced classes. Consequently, by balancing the dataset, the overall performances yield better results. To the sake of completeness, the latest discussion is devoted to the generative model. The PointFlow model will be the one to design future experiments, as it gains stability for both SG and GS values. This means that the generator is able to reproduce the object regardless of the type of training-test method (see Table 4.12).

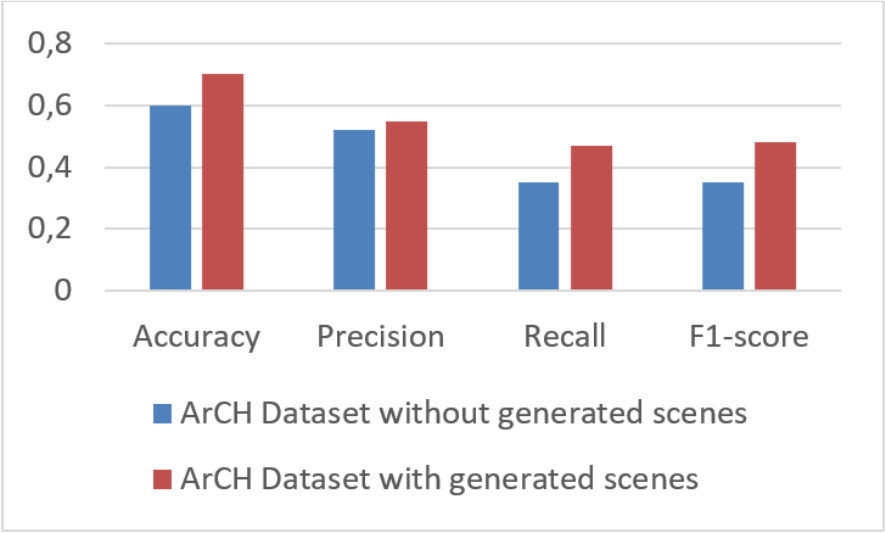


Figure 4.16: Comparison of metrics considering ArCH dataset without and with generated.

4.2 Change detection on a dynamic space

In this section, the author of this thesis reports the results of the experiments conducted on SMART Dataset, using a novel Change Detection framework, described in Section 3.2.1. The performance of the fusion classifier is presented, with the performance of the visual and textual category classifiers (based on the visual and textual feature extractors) being the key indicators to the overall classification. For the experimental analysis, the labelled dataset has been split into a training set and a test set. Each classifier was trained solely through the training set, while the test set was used for all test purposes. The considered dataset is split into two randomly selected sub-sets: 80% for training images and 20% for test images, accounting for all permutations of overall, visual, and textual annotations.

The performance of the visual classification is reported in Table 4.15.

As shown above, high values of precision and recall can be achieved, especially for pictures with Normal and Promotion visual content. The visual recognition of SOOS pictures is more difficult, due to the smaller amount of training data available and the fact that retailers pay more attention to the SOOS situation, trying to solve this problem immediately when it occurs.

In Table 4.16, the precision, recall, and F1-score of the textual classification is presented. The textual classification performance is mainly good, but lower than the visual classification performance. While the classification of visual and textual image content is equally difficult for humans, the classification of the text in the picture is

Table 4.15: Performance evaluation of the visual model

DCNNs	Category	Precision	Recall	F1-Score
VGG-16 [90]	**SOOS**	0.815	0.817	0.816
	Normal	0.863	0.854	0.858
	Promotion	0.891	0.899	0.895
	MEAN	***0.856***	***0.856***	***0.856***
AlexNet [108]	**SOOS**	0.796	0.799	0.798
	Normal	0.821	0.833	0.827
	Promotion	0.889	0.874	0.881
	MEAN	***0.835***	***0.835***	***0.835***
CaffeNet [117]	**SOOS**	0.809	0.771	0.789
	Normal	0.808	0.859	0.833
	Promotion	0.896	0.881	0.888
	MEAN	***0.837***	***0.837***	***0.837***
GoogleNet [118]	**SOOS**	0.820	0.814	0.817
	Normal	0.849	0.859	0.854
	Promotion	0.893	0.889	0.891
	MEAN	***0.854***	***0.854***	***0.854***
ResNet-50 [119]	**SOOS**	0.780	0.822	0.801
	Normal	0.850	0.845	0.847
	Promotion	0.918	0.874	0.895
	MEAN	***0.849***	***0.847***	***0.848***
ResNet-101 [119]	**SOOS**	0.848	0.792	0.814
	Normal	0.831	0.888	0.859
	Promotion	0.905	0.893	0.899
	MEAN	***0.858***	***0.858***	***0.857***

much more challenging for machines, as it needs to be detected and recognised before it can be classified. Comparing the different classes reveals that Promotion texts can be better recognised than the other two categories analysed.

Table 4.16: Performance of the textual DCNN model, predicting textual content based only on textual features.

DCNNs	Category	Precision	Recall	F1-Score
Character-level DCNN [197]	**SOOS**	0.523	0.509	0.516
	Normal	0.528	0.601	0.563
	Promotion	0.796	0.707	0.749
	MEAN	***0.616***	***0.606***	***0.609***
LSTM [207]	**SOOS**	0.457	0.404	0.437
	Normal	0.504	0.614	0.554
	Promotion	0.767	0.714	0.739
	MEAN	***0.582***	***0.578***	***0.577***

The performance of all classifiers is shown in Figure 4.17, in terms of F1-Score. The average for both kNN and SVM reached the best performance for all possible visual and textual feature extractors (see Figure 4.18).

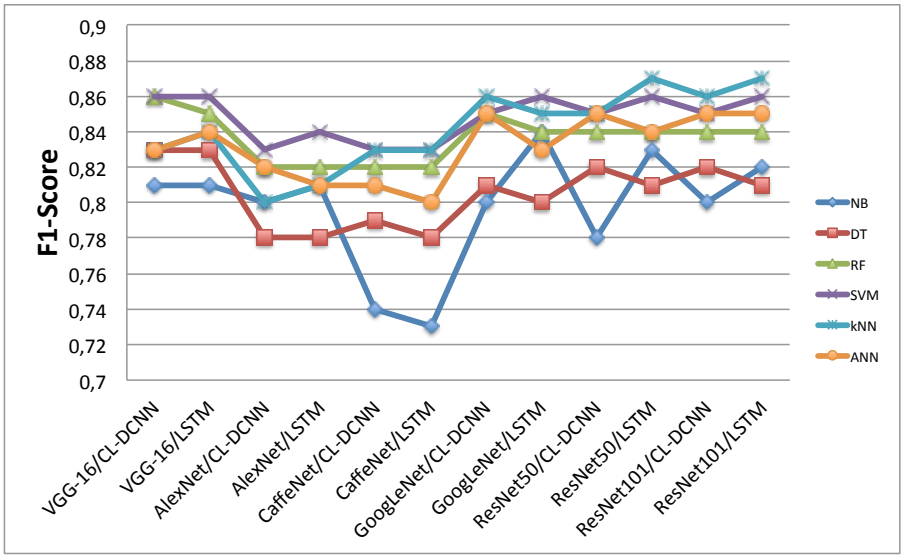


Figure 4.17: Performance in terms of F1-score for all classifiers

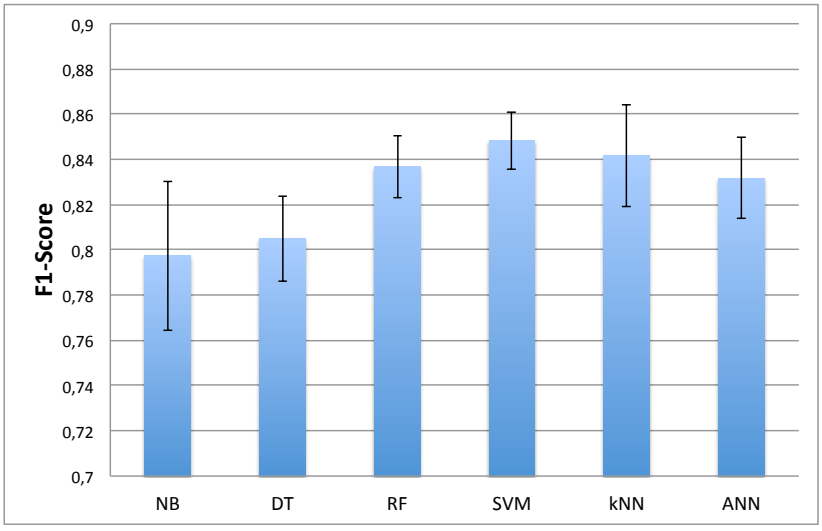


Figure 4.18: Performance in terms of F1-score for all classifiers averaged over all visual/textual feature extractors

The application of kNN as a fusion classifier for modelling the visual features extracted from the ResNet with 50 and 101 layers and the textual features extracted from LSTM achieved the best performance (i.e., F1-Score = 0.87). The proposed ROCKy navigation system, based on customer behaviour heat-maps, was able to reduce the surveying time by 45%. During the tests in two different stores, ROCKy was pro-

grammed to only visit locations that more than 10 carts and/or baskets passed on a six hours time frame for the heat-map data accumulation and a potential field navigation method. ROCKy was able to navigate a 1,500 sqm store in 18 minutes and 52 seconds with a total navigation time of about 34 minutes. The navigation system and the resulting survey feedback are suitable for deployment in real conditions and can be applied to larger retail environments with efficient results in terms of fast surveying, high accuracy, and return of investments due to low cost solutions, all with respect to human workers and inspections.

4.3 Person Re-Identification on a dynamic space

This section shows the results of experiments carried out with the Person Re-Identication methods proposed by the author. Section 4.3.1 shows the results of the TVOW approach, trained with the TVPR2 dataset. Then, Section 4.3.2 describes the performance of the temporal approach presented in Section 3.3.2, first finetuned with the TVPR2 dataset and then inserted in the museum SeSAME system.

4.3.1 TVOW framework

In this section, the results of the experiments conducted using the TVPR2 dataset are presented. Experiments are separated in two phases. First, the new TVPR2 dataset is used to find the best combination of hyperparameters and backbone network. Second, the best configuration is tested on a state-of-the-art dataset, namely the TVPR dataset [151].

The TVOW approach is tested using the backbones of several state-of-the-art networks, pretrained using the ImageNet dataset, then fine-tuned on the TVPR2 dataset. The chosen networks are:

- ResNet-50 [91], fine-tuned on Layer4, with 68,9 Millions of parameters;
- ResNext-50 [209], fine-tuned on Layer4, with 25 Millions of parameters;
- DenseNet-161 [210], fine-tuned on DenseBlock4 Layer, with 28.7 Millions of parameters;
- GoogleNet [211], fine-tuned on Inception5b Layer, with 6.7 Millions of parameters.

The dataset used for testing is TVPR2. During network training, data augmentation techniques, such as flipping, rotation, random crop and padding, are used.

The tests are carried out initially using a close-set configuration through the mean average precision (mAP) metric and CMC curves. In the second phase, an open-set environment is tested using various combinations of TTR and FTR values. The tests

are repeated with a variable number of *id*. Networks are trained sequentially with 100, 300 and 1000 people from the dataset. The metrics used in this particular configuration are the *mAP* and the *CMC*. Table 4.17 shows the mAP for each tested network. By contrast, Figure 4.19 compares the *CMC* curves of every backbone for three different ranges of a person's IDS.

Model	**100 persons**	**300 persons**	**1000 persons**
ResNet-50 [91]	92,10	81,09	76,86
ResNeXt-50 [209]	93,30	84,23	78,74
DenseNet-161 [210]	91,56	79,20	69,08
GoogleNet [211]	75,54	54,58	41,00

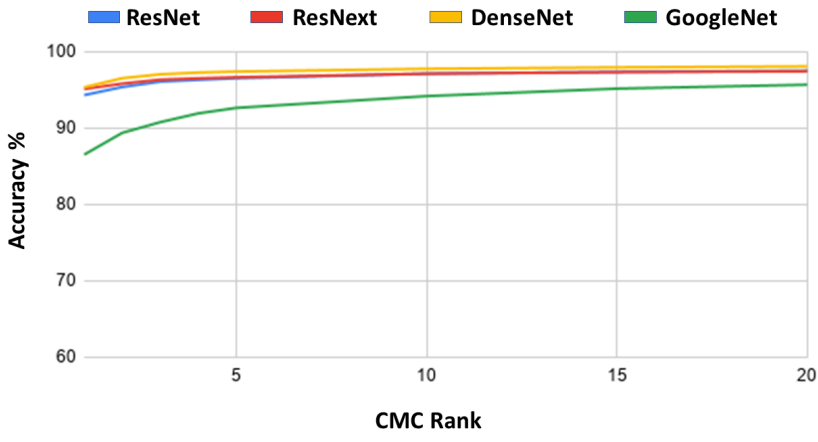
Table 4.17: Mean average precision (%) for close-set configuration.

These results indicate increasing the number of people in the phase of training results in performance deterioration. In particular, the *GoogleNet* exhibited more difficulty when learning and the largest deterioration with more people. The most stable networks are *ResNet* and *ResNext*.

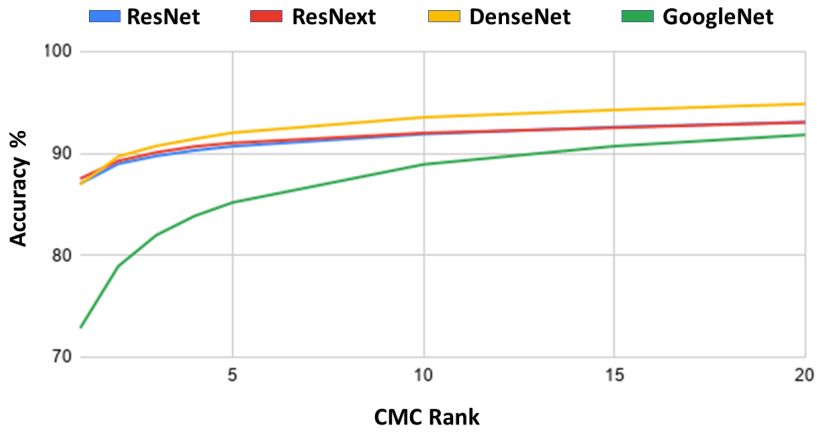
As shown by the *CMC* curves (Figure 4.19), the performances by *ResNet*, *ResNeXt* and *DenseNet* are very similar. The *rank1* exceeded 90%, indicating no particular increase with higher *rank*. *GoogleNet* remains worst performing network.

As the the best results (on average) are obtained using ResNext-50 as the backbone network (with retraining via fine-tuning on Layer4), this configuration is used for the next comparison. The following tests are then performed to evaluate performance in an open-set environment. In this situation, there are an unknown number of people that the network has not used for the training. The objective of these tests is to judge the ability of the networks to correctly identify already known *targets* while discerning unknown non-targets. To do this, new state-of-the-art metrics have been chosen, *TTR* and *FTR* values, already introduced in previous sections. The experiments are performed by varying the number of *targets*, using 100, 300 and 500 people per trial. For each of these cases, the author increases the number of non-targets by a percentage value of 10%, 50% and 100% compared to the number of *targets*, as shown in Figure 4.20.

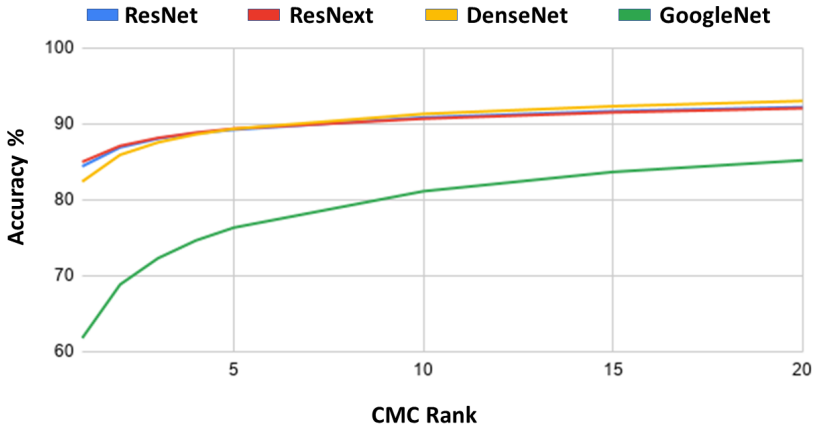
The first graph (Figure 4.20a) compares TTR and FTR values by using 100 targets and 10 non-targets for testing. This graph indicates how obtaining a low FTR value generally leads to a decreased TTR value. For an FTR value of 30%, the TTR is closer to 100%, a situation caused by the positioning of non-targets in the feature space. For a threshold of 30%, most of the non-targets' features are far from those of the targets, while going below 30% non-targets caused the targets to start moving closer and creating confusion. The TTR value decreased because the matching algorithm applied a threshold on the distance, which considered all features beyond it as belonging to non-targets. Regarding network performance, the results are clear: ResNeXt is more



(a) 100 people.



(b) 300 people.



(c) 1000 people.

Figure 4.19: CMC curves for close-set configuration. The tests were repeated with a variable number of ids: (a) 100 people, (b) 300 people and (c) 1000 people.

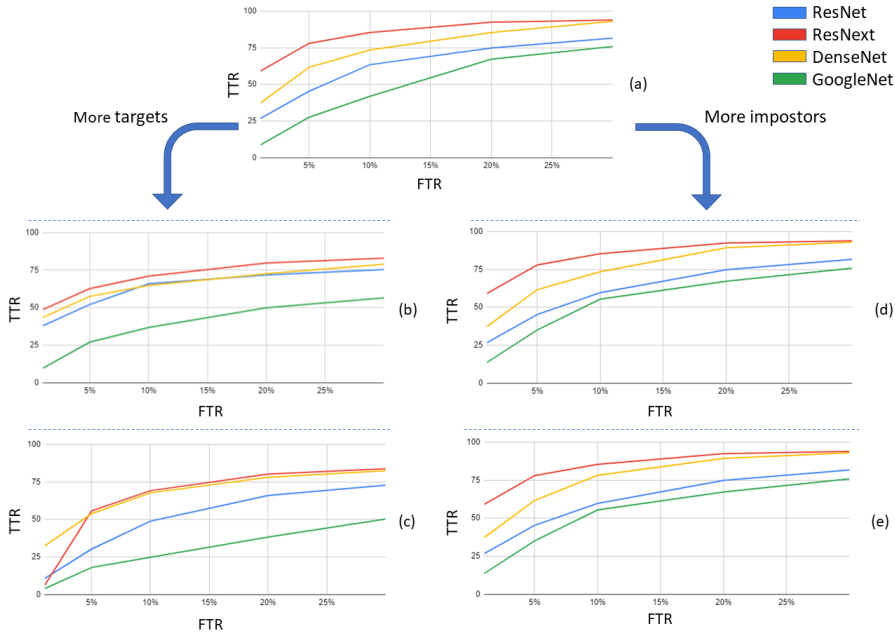


Figure 4.20: TTR and FTR value results for an open-set environment. (a) Test on 100 targets and 10 non-targets. (b) Test on 300 targets and 30 non-targets. (c) Test on 500 targets and 50 non-targets. (d) Test on 100 targets and 50 non-targets. (e) Test on 100 targets and 100 non-targets.

robust than the others.

The graphs on the left of Figure 4.20 show the metrics variation when the number of targets is increased while maintaining the percentage of non-targets. Figure 4.20b compares TTR and FTR values by using 300 targets and 30 non-targets and Figure 4.20c compares TTR and FTR values given input of 500 targets and 50 non-targets. The shape of the curves is similar to the case with 100 targets, but the values decreased generally as the number of targets increased. This result is predictable as other metrics deteriorated for the same reasons. In the 500 target case, however, there is an anomalous worsening of ResNext.

The graphs on the right of Figure 4.20 show the effect of increasing the number of non-targets from 10% (Figure 4.20a) to 50% (Figure 4.20d) and 100% (Figure 4.20e). The difference between the 10% and 50% cases is a slight deterioration in the TTR value given an FTR value range between 10% and 20%. This difference is caused by additional non-targets being positioned around the threshold values and being interpreted as targets. Increasing non-targets from 50% to 100% produced practically identical graphs, indicating the new non-targets were distributed in the same positions as the old ones. This behaviour is due to the operating principle of the triplet networks; in fact, they learn to cluster classes and this effect extends beyond learned classes to

newly encountered classes. This important result demonstrates the strength of triplet networks in an open-set environment.

Table 4.18 provides a comparison of the TVOW approach with other state-of-the-art methods concerning person re-ID from a top-view perspective. TVDH is the method of Liciotti [151] to extract anthropometric features through image processing techniques, then training machine-learning algorithms for re-ID tasks. In RGB-D-CNN, Lejbolle [153] started with a two-flow Convolutional Neural Network (CNN) (one for RGB and one for depth) and a final fusion layer. Then they improved this approach with a multimodal attention network called MAT [154], adding an attention module to extract local and discriminative features that were fused with globally extracted features. In another work, Lejbolle [155] presented a SLATT network with two types of attention modules (one spatial and one layer-wise). Table 4.18 shows the results in terms of CMC curves (rank-1, rank-5, rank-10 and rank-20).

Method/Rank	**r = 1**	**r = 5**	**r = 10**	**r = 20**
TVDH [151]	75,50	87,50	89,20	91,20
RGB-D-CNN [153]	92,55	97,87	97,87	100,00
MAT [154]	94,68	97,87	97,87	100,00
SLATT [155]	93,62	96,81	97,87	100,00
TVOW [158]	**95,13**	**98,03**	**98,75**	**100,00**

Table 4.18: Test using the TVPR dataset to compare TVOW with other state-of-the-art methods

As a final step, the TVOW approach are compared with the most recent state-of-the-art approach SLATT [155], using the TVPR2 dataset. The test was made in a closed-set environment with 100 targets. Table 4.19 shows results of this comparison in terms of mAP and CMC curves (rank-1, rank-5 rank-10 and rank-20).

Method	**mAP**	**r = 1**	**r = 5**	**r = 10**	**r = 20**
SLATT [155]	90,18	91,30	93,77	94,08	95,30
TVOW [158]	**93,30**	**94,02**	**96,68**	**97,12**	**98,55**

Table 4.19: Testing using the TVPR2 dataset comparing TVOW with SLATT. Results are based on mAP and CMC curves.

The proposed method, based on depth information, allows a person detection with the minimum possible error, because the RGB and Depth frames are automatically synchronized by the camera, both spatially and temporally. In this way, the author can also remove the noise due to the background around the person (through the background removal phase) and then learn only the main features of the person itself. Moreover, the use of triplet loss with hard batch allows to train the network in an efficient way, because it increases the distance between the features of the sample frame (anchor) and the frames of different people (negative), while it decreases the distance

with the frames of the same person but in different poses (positive).

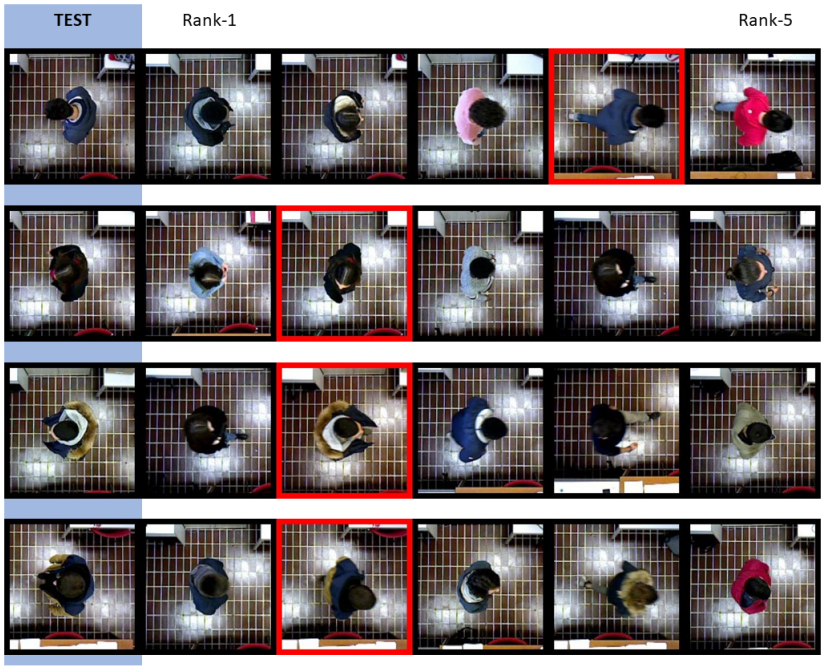


Figure 4.21: Examples of missmatched IDs for a visual analysis of the results. The first column shows the RGB frame of the person in the test set (obviously the relative depth frame is also given as input). The others show representative images for the first 5 predicted IDs. The red box figures the ground truth.

From these results, it is possible to evaluate which people were not recognised more frequently and with whom they were confused. Figure 4.21 highlights four examples of mismatched targets.

Finally, the author added a further test to compare the proposed depth-based method of person detection with a state-of-the-art one: the Region of Interest (ROI) was extracted from the original frame by using a You Only Look Once (YOLO) detector, trained only on the RGB frames. In TVOW approach, the phase of ROI detection has been improved by using a threshold on the depth channel. From Figure 4.22, it is possible to infer the improvements respect the two used methods. Using the YOLO-based method, several errors are committed in person detection: the person could be partially picked up, or shot differently on continuous frames; or even confused with some objects (Figure 4.22.a). TVOW approach ensures that a height threshold is set on the depth channel, so as to remove all the lower objects than people. Furthermore, the detection will be done only on objects that have passed a certain limit area, to be

sure that they are really people. Finally, using the depth information, the background will be removed and only the person's information will remain (Figure 4.22.b).

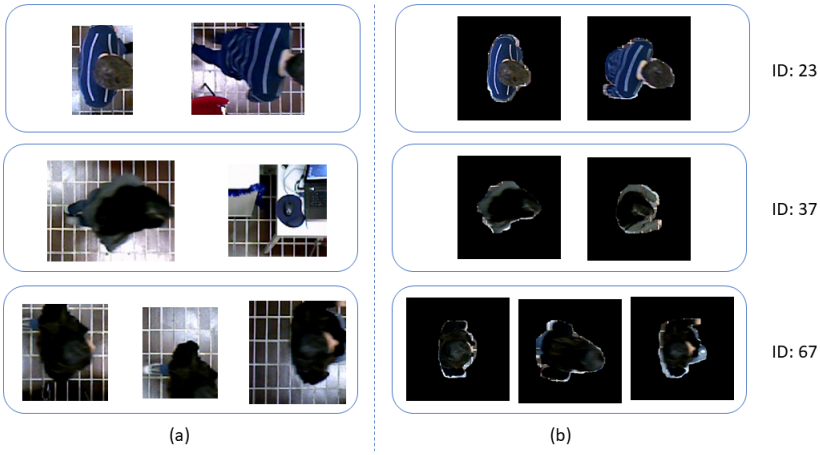


Figure 4.22: Preprocessing comparison for person detection. (a) Some incorrect detections using YOLO. (b) Correct detections of identical targets using our approach based on depth information.

4.3.2 A multimodal Person Re-Identification framework.

In this section, experiments carried out on the SeSAME framework will be described. The first experiment concerns the choice of the best person re-ID system, exploiting all the information available: the RGB colour channels, the Deph channel, the temporal information given by the acquisition of videos with respect to the single images. As stated in the previous Sections, the best approach is used for the multi-camera system in museum environment. In particular, the proposed approach has to recognise people detected by 6 different cameras, and then evaluate statistics regarding the behaviour of the visitors inside the museum.

The first experiments concern 3D-CNN and 2D-CNN (with related temporal modelling methods) tested on TVPR2 dataset. The performance of these networks are evaluated in terms of mAP and CMC. The mean average precision (mAP) metric is defined as the mean of the maximum precision as a function of recall values. mAP can be used to measure accuracy in classification problems of ordered sequences. The cumulative matching characteristic (CMC) curves indicate the probability of finding the right match in the first n most-expected matches. The CMC curve metric is a common metric in the evaluation of re-ID methods.

The implemented networks are the following:

- ResNet-50 standard, pre-trained on ImageNet (2D-CNN);

- 3D ResNet-50, pre-trained on Kinectics (3D-CNN).

The networks are trained by using an Adam optimiser and a batch size of 32. The test phase needed gallery and query sets for each person. The gallery was created using the videos which contain first passage of the person under the camera (left to right), while the query set was based on videos in which it is show the return to the initial position (right to left). Each video was divided into many clips of fixed lengths (T frames) before the shorter-length clips were increased by duplicating the frames.

First, performance was evaluated by comparing two different ways of fusing the features extracted from the two networks – concatenating and summing them – after changing the encoding of the depth frames into false RGB colours by using JET colormap.

Table 4.20 reports the results using the JET colormap for the depth channel and concatenating the features extracted from both networks. Table 4.21 summarises the results using the JET colormap for the depth channel and summing the features extracted from both networks.

Table 4.20: Results using the JET colormap for the depth channel and concatenating the features extracted from both networks.

Model	*mAP*	CMC_1	CMC_5	CMC_{10}	CMC_{20}
3D-CNN	72.6	82.6	88.3	91.7	94.4
Temporal Pooling	74.2	82.7	90.3	95.2	96.6
Temporal Attention	**77.9**	**84.2**	**94.6**	**97.1**	**98.8**
RNN	73.5	80.4	91.5	93.9	95.1

Table 4.21: Results using the JET colormap for the depth channel and summing the features extracted from both networks.

Model	*mAP*	CMC_1	CMC_5	CMC_{10}	CMC_{20}
3D-CNN	70.8	80.4	86.7	90.1	93.3
Temporal Pooling	72.1	80.6	88.1	90.2	94.3
Temporal Attention	**75.9**	**83.1**	**90.8**	**93.9**	**95.0**
RNN	70.5	81.3	85.5	89.3	93.8

This first part of experiments demonstrated the effectiveness of the feature concatenation respect the summarization one, both in terms of mAP and CMC values.

The second part regarded tests on the four temporal approaches to understand the benefits of the multimodal approach, where both RGB and Depth information are merged.

The experiments performed were based on the following idea. Initially, the temporal approaches were tested using only the RGB stream without the depth information. Then, the results obtained were compared using both streams, RGB and depth, through the feature-fusion method. These two tests were carried out by training the networks with videos of 100 people and then validating them on another 100 people never seen

by the network. This way allowed to verify whether the addition of the depth information improved the performances of the networks, which use only the depth ones. Finally, these trained networks have been tested on another 800 people never seen by the network. This test prove the generalisation of the network on larger datasets.

Table 4.22 shows the results a dataset of 100 people for training and another 100 people for the test with only the RGB features. It can be infered that the Temporal Pooling approach achieves good results both in terms of mAP and CMC curves.

Model	*mAP*	CMC_1	CMC_5	CMC_{10}	CMC_{20}
3D-CNN	27.4	24.8	37.0	48.4	62.0
Temporal Pooling	**58.4**	**55.7**	**68.1**	**73.7**	**86.8**
Temporal Attention	11.9	8.9	20.3	31.1	42.0
RNN	16.5	10.9	27.1	34.9	46.1

Table 4.22: Experimental results using 100 people for training and another 100 people for the test. The approach only took the RGB stream as input without using the depth information.

Table 4.23 shows the results of using both the RGB and depth features with a dataset of 100 people for training and another 100 people for the validation. As it can be noticed, depth information significantly improve the performances of the temporal approaches. Figure 4.23 shows the validation accuracy of all the temporal modelling approaches on 1000 epochs of training. All the methods tended to converge to a specific value after 400 epochs. The results on the validation set show that the Temporal Attention method obtains good performance in terms of mAP. However, it does not performs well on accuracy. The Temporal Pooling achieves the best performance in terms of CMC curves yet.

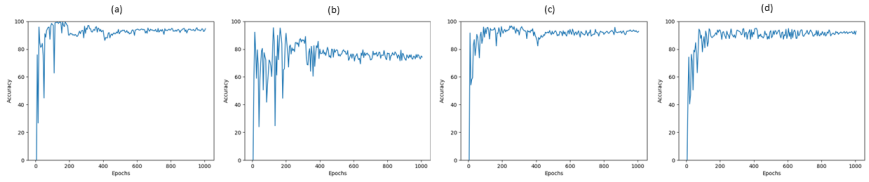


Figure 4.23: Validation accuracy of all the Temporal Modeling Approaches on the training phase. (a) Temporal Pooling. (b) Temporal Attention. (c) RNN. (d) 3D-CNN

Finally, the trained networks are tested on the other 800 person recorded in the TVPR2 dataset, which have never been seen from the networks. The 100 validation people were chosen to be very heterogeneous among them, to have an exhaustive sample of the entire population, and thus discriminating different people carefully. Table 4.23 shows that the Temporal Pooling approach reaches performances in terms of mAP and CMC curves, hence demonstrating that this network generalised well on a larger dataset.

Table 4.23: Experimental results using 100 people for training and another 100 people for the Validation. The approach took both the RGB and depth streams as inputs. Finally these trained networks were tested on another 800 people.

	Validation (100 People)					**Test (800 People)**				
Model	*mAP*	CMC_1	CMC_5	CMC_{10}	CMC_{20}	*mAP*	CMC_1	CMC_5	CMC_{10}	CMC_{20}
3D-CNN	67.0	91.3	94.1	96.0	97.9	77.2	93.0	96.5	98.1	99.0
Temporal Pooling	68.0	**94.7**	96.7	97.2	**99.2**	**81.6**	**98.8**	**99.4**	**99.7**	**99.7**
Temporal Attention	**68.3**	74.7	82.5	85.6	89.6	78.6	93.1	96.4	97.8	98.6
RNN	66.7	90.9	**97.5**	**98.5**	98.7	75.6	98.7	99.3	99.4	99.6

Experiments on the Museum Dataset

The Temporal Pooling approach proved to be the best method for re-ID in a Top-View configuration, with a multi-modal approach of RGB-D and Temporal data. This network, pre-trained on the TVPR2 dataset, was adapted for SeSAME.

The network allows to extract the features of the people detected at the entrance of the museum, generating a gallery of the incoming people. The camera system, thanks to the same network, was able to detect the same people at the various entrances and exits of the building floors. All this information was constantly sent to the final module of the framework, which generated very useful statistics about the visitors.

Each day, the system generates information for each user, concerning: day of acquisition; unique person id; all visited floors; entry timestamp; floor 1 entry timestamp; minutes spent on floor 1; floor 2 entry timestamp; minutes spent on floor 2; floor entry timestamp -1; minutes spent on floor -1; total minutes spent inside the museum.

From these information, SeSAMe can generate the following daily statistics: total persons during the day; total persons who visited floor 1; total persons who visited floor 2; total persons who visited floor -1; total persons who visited floor -1 and floor 1; total persons who visited floor -1 and floor 2; total persons having visited floor 1 and floor 2; total persons who visited floor -1, floor 1 and floor 2; average minutes spent inside the museum; average minutes spent inside floor 1; average minutes spent inside floor 2; average minutes spent in floor -1. This is all important information that allows the museum to study users' behaviour on the different floors, and to make decisions in order to improve visitors' satisfaction.

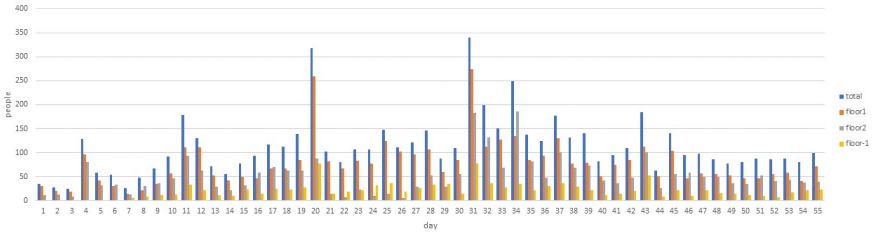


Figure 4.24: People detected for each day of the Museum dataset.

Figure 4.24 represents people detected for each day: SeSAME evaluates the total people and the people who visited each floor. The results show that the floor that generated the most interest was floor 1, while the least interesting floor was floor -1.

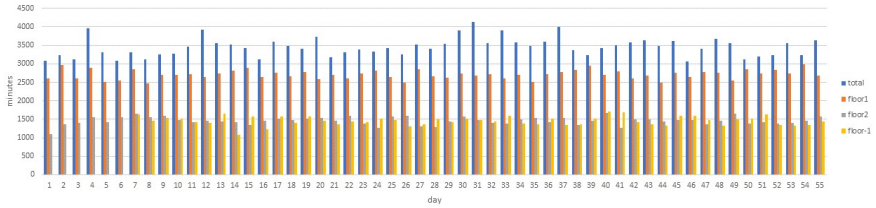


Figure 4.25: Minutes spent by people for each day of the Museum dataset.

Figure 4.25 shows the graph of minutes spent by detected people for each day of the dataset: the system calculates the total minutes and the minutes spent on each floor. Even in this case, the results show that the floor that generated the most interest was floor 1, while the least interesting floor was floor -1.

Finally, Table 4.24 shows the average values of the statistics concerning the persons detected and the minutes spent, both in the museum and on each floor.

Table 4.24: Mean values of the statistics processed on the entire Museum dataset.

Statistic	*Museum*	*Floor1*	*Floor2*	*Floor-1*
People	112,27	77,4	51,2	20,9
Minutes	3453,9	2710	1458	1298

Chapter 5

Discussions

5.1 Thesis Contributions

In this Section the author discusses and answers the research questions provided in Section 1.2.

How can Computer Vision algorithms be applied to the different subtasks regarding Space Understanding?

As described above, there are a number of machine learning methodologies that can be useful in solving various problems related to spatial understanding. The choice of the right approach may depend on various factors, including time limits for data processing, quality and quantity of data collected, and expectations to be met.

Starting from the simplest definition of space, i.e. seen as a static concept, the author of this thesis proposed in Section 3.1 an approach based on semantic segmentation as it is perfect for locating and defining all the physical entities within a scene. In addition, the choice of data typology has fallen on Point Clouds because, although compared to RGB images and videos they are slower to acquire and more expensive to save, they have much more useful information than them. Therefore, in the static understanding of space as a static concept, the time limitation of the acquisition will not be a problem.

Subsequently, the study of the evolution of a dynamic space was solved by moving to different approaches, obviously not based on Point Clouds due to the limitations described above. The author of the thesis proposed in Section 3.2 a new mixed approach between ML and DL algorithms for the Change Detection task. In this way it is possible to save and identify over time the changes occurred within a space. Since systems of this type need to save data over long periods of time, it was decided to use simple RGB images as the type of data to be acquired.

Finally, for the study of the temporal evolution in the short term of the entities enclosed within a space, the author proposed in Section 3.3 a new Re-Identification approach useful both in closed world and in more realistic open world environments. The choice of the type of entities to be recognised has focused on people, since they are the most common and dynamic ones. In this case, since the time limit for saving data is shorter than in the previous case, the author has chosen RGB-D videos as the type of data to be acquired. They are a good compromise between advantages and

disadvantages obtained from Point Clouds compared to simple RGB images, as they are an intermediate type between them.

The integration of data of a different nature could help the understanding of a Space, or the entities contained within it?

In Section 3.1.1, the author showed that in the Point Cloud Semantic Segmentation task, the addition of further features than those usually used in the state of the art allowed to improve the performance of the Deep Learning approaches. In fact, the acquisition of point clouds, in addition to simple coordinates and colour information, allows to obtain additional features such as normal vectors and other data related to the materials of the acquired objects. These features proved to be very discerning and therefore useful to segment a space into its various parts.

Furthermore, the author has verified in Section 3.1.2, that also handcrafted features can be useful to better guide the learning of the neural networks. In Section 3.2, the author has also shown that useful information can be extracted but of a different nature with respect to the starting data, that is the extraction of textual features from rgb images. Extracting these features and using them in conjunction with visual features improved the performance of the proposed approach.

Finally, the most comprehensive experiment on intregation between data of different nature was carried out in Section 3.3.2. In fact, the author proposed a new Person Re-Identification method with good performance due to the integration between RGB data, the depth D channel and temporal features extracted from the acquired videos.

Is it worth following the trend of using pure deep learning methods respect the machine learning ones? Or could a mixed approach bring improvements?

As described in Section 4.1.2 regarding the Point Cloud Semantic Segmentation, the author think that there is still no winning solution between the ML and DL approaches. The OA of the best ML method and the DL one differs slightly. However, contrasting results are highlighted if the classes are analysed individually, where approaches could be chosen according to the needs. Both techniques have strengths and weaknesses. In the case of ML, there is a customisation of the training set according to the scene to be predicted, very useful in the CH domain, while for the DL there is the possibility of cutting out the manual annotation, further automating the process. Another element to take into consideration when comparing machine and deep learning approaches is the processing time. If the ML pipeline is well defined, within the DL framework, it is necessary to make a distinction between two possible scenarios which considerably differ in times. In the first scenario, when an annotated training set is not available, it is necessary to manually label as many scenes as possible (a very time-consuming task), pre-process the data (e.g., subsampling, normals computation, centering on the [0,0,0] point, block creation, etc.), then wait for the training phase from a few hours to a few days. In the second scenario, it is possible to start from saved weights of a network which had been pre-trained on a released benchmark (ArCH in this case), and directly proceed to the preparation and test of the new scene, without any manual

annotation phase. So, depending on whether one compares the RF with the first or second scenario, the balance needle can tip in favor of one or the other technique. In Figure 5.1, a comparison between the times required for the tests carried out in Section 4.1.2 is shown. It must be considered that ML tests were run on an Nvidia GTX 1050 TI 8 GB, 32 GB RAM, processor Intel(R) Xeon(R) CPU E5-1650 0 @ 3.20 GHz, while for the DL an Nvidia RTX 2080 TI 11 GB, 128 GB RAM, processor Intel(R) Xeon(R) Silver 4214 CPU @ 2.20 GHz was used.

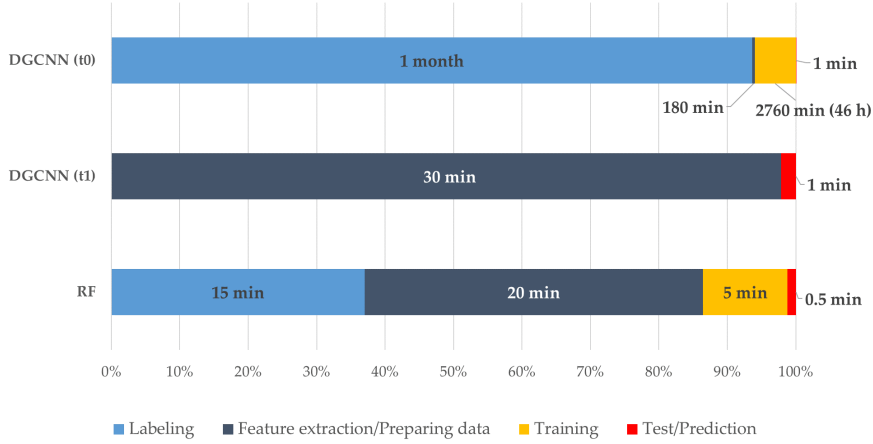


Figure 5.1: Normalised comparison of times required for the different scenarios test. NN (t0) represents the first scenario in which the whole dataset has been manually labeled and the DGCNN-based methods have been trained on all the scenes. NN (t1), on the other hand, represents the next scenario in which it is possible to use the weights from the pre-trained neural network and conduct directly the data preparation (feature extraction, scaling, blocks creation, subsampling, etc.) and the final test for the prediction.

Finally, it is fair to state that the main drawback in the comparison between different algorithms is the limited similarity of their pipeline. In fact, a proper comparison between algorithms would necessarily require the same input and/or output. As regards the input, considering the different nature of the algorithms, this would mean giving to the ML classifiers a huge amount of annotated data which would compromise its performances, or viceversa training the neural network with a few data compared to that required. For this reason, in order to analyse the best classification approaches for heritage scenarios, the author preferred to use different training scenes for the ML and DL input. Concerning the output, for the DL approach an interpolation with the initial scene should be conducted for a comparison with the same number of points, leading to a likely OA decrease. However, as the subsampling operation is mainly due to computational reasons, easily solved in the near future with more and more performing machines, the usefulness of the interpolation would certainly be reduced and

become even pointless. Moreover, using different interpolation algorithms would introduce a further element of error making the pipeline less objective and reproducible.

In general terms, the training time of classical ML techniques can be up to one order of magnitude smaller; conversely, a small but noteworthy improvement in performance could be witnessed for DL techniques over classical ML techniques, considering the whole benchmark dataset. In ML, hyperparameter optimization or tuning is the problem of choosing a set of optimal hyperparameters for a learning algorithm. Its value is used to control the process of learning. Instead, DL techniques have the advantage of allowing more additional experimentation with the model setup. Using DL techniques on a dataset of this size and for this type of problem therefore shows promise, especially in performance critical applications. On the other side, the DL model is largely influenced by the processes of tuning the structural parameters both in computational cost and operational time. However, given that state-of-the-science large-scale inventories are moving towards deep learning-based classifications, it can be expected that in the upcoming future the growing availability of training dataset will overcome such limitation. The feature engineering and feature extraction are key, and time consuming parts of the ML workflow, since these phases transforming training data and augmenting it with additional features in order to make ML algorithms more effective. DL has been changing this process and deep neural networks have been explored as black-box modelling strategies.

Finally, a further comparison between these methodologies is described in Table 5.1, where pros and cons of both ML/DL methods are summarised. It was aimed at opening a positive debate among the different involved domain experts.

Table 5.1: Comparative overview table with the key differences between the two proposed frameworks in the CH domain. From low (*) to high (***).

	Machine Learning	**Deep Learning**
Training Set Size Dependencies	*	***
Programming Skills	*	***
Feature Engineering	***	*
Algorithm Structure	*	***
Interpretability	***	*
Training Time	*	**
Hyperparameter Tuning	***	***
Processing Power and Expensive hardware (GPUs)	**	***

Are the proposed algorithms better than the standard algorithms widely used in the literature? To answer this question it is necessary to consider the various tasks into which the problem of Space Understanding has been divided.

Starting from Space as a static concept, the author initially proposed in Section 3.1.1 a new Deep Learning method for Semantic Point Cloud Segmentation. This approach was tested in the Cultural Heritage domain using the ARCH dataset and

compared with several state of the art approaches ([34, 65, 67, 68]). The comparison showed that the DGCNN-Mod network improved the average Segmentation Accuracy by 8.5% on the test scene, compared mainly to the original network on which the approach is based. The average Precision, Recall and F1-Score metrics also improved by 4.1%, 8.5% and 7.6% respectively. The proposed approach also guaranteed a good degree of generability. In fact, it was tested in Section 4.1.1 on the indoor S3DIS dataset, obtaining an improvement in accuracy of 1.8% compared to state-of-the-art approaches.

In Section 3.1.2, the author improved the proposed method by adding handcrafted features, a loss function suitable for unbalanced datasets and data normalization techniques. Here again, an exhaustive comparison with state-of-the-art techniques, both Machine Learning ([178, 179, 180, 181, 182, 183, 184, 185, 186, 187].) and Deep Learning ([68, 69]), was made. The tests conducted and the results described in Section 4.1.2 show that the introduction of 3D features has led to an increase in OA, if compared to the simple use of radiometric components and normals. This increase is about 10% in the tests on the symmetric scene (Trompone church), while it is lower (approximately 2%) in the tests run with different scenes as training and SMV or SMG as tests. In the latter case, however, the introduction of the 3D features, associated with the use of the normals and the RGB features, has improved the recognition of the classes with fewer points and which, previously, resulted with lower metrics (for example Column, Door-Window and Stair). As it is possible to notice in [176], for all the approaches, the worst recognised classes are Arch, Door-Window and, alternatively, Molding or Stair. This result is likely due to the fact that these are the classes with the lowest number of points within the scenes. A similar conclusion can be made for the introduction of the focal loss, which, with the same hyperparameters configuration, has led to an increase of the performance for the Molding, Door-Window, and Stair classes.

Then, generative methods were used as a data augmentation technique to counterbalance classes with few points. The author selected 3 of the best state-of-the-art networks for point cloud generation ([80, 81, 82]) and then compared the proposed approach using first the original dataset and then the augmented dataset. The results in Figure 4.16 show that the accuracy of semantic segmentation has improved thanks to the inclusion of these new generated scenes. It is fairly straightforward to deduce the motivation behind this increase in accuracy; data generation has been performed for those classes with a lower number of points available, that is to say unbalanced classes. Consequently, by balancing the dataset, the overall performances yield better results. To the sake of completeness, the latest discussion is devoted to the generative model.

In Section 3.2, the author propose a novel mixed Change Detection framework to study the evolution of a dynamic Retail Space on long-period time. The approach is part of ROCKy, a mobile robot system able to detect and map SOOS and PA

in a retail environment. In addition to the identification of SOOS, PA, and misplaced items, ROCKy can provide information on promotions and discounts to the customers. The proposed system could also be used to monitor warehouses and run surveillance at night. The approach is able to learn a high level representation of both visual and textual content and achieve high classification performance. The visual feature extractor was chosen by comparing different state-of-the-art image classifier ([90, 108, 117, 118, 119]), while the textual feature extractor was chosen from the comparison between the two most widely used methodologies in text classification (DCNN [197] and RNN [207]). The experiments conducted on SMART Dataset have a rather high accuracy and demonstrate the effectiveness and suitability of the approach used. The performance of the overall content classification is higher (i.e., F1-score=0.87) than the performance of the single visual (i.e., F1-score=0.86) and textual classification (i.e., F1-score=0.61). The overall classifier was chosen from the comparison between the most widely used machine learning classifiers ([198, 199, 200, 201, 202, 203, 204]).

Finally, in Section 3.3 the author described a new approach for the re-identification of people, i.e. to study the short-term temporal evolution of the most common entities in a space. The backbone of the TVOW approach has been choosen by comparing different state-of-the-art methods ([91, 209, 210, 211]). The final approach was compared with several state-of-the-art methods regarding RGB-D videos, in a Top-View configuration ([151, 153, 154, 155]). The results in Section 4.3.1 show that the TVOW method achieves an improvement in Accuracy of at least 0.45% on the TVPR dataset and 2.72% on the larger TVPR2 dataset. The mAP is also increased by 3.12% compared to the state of the art. Subsequently, the approach was also tested in the open word environment, using the TTR and FTR metrics. Although it was not possible to make a comparison with the state of the art methods, as this is a very recent issue, the results show that even increasing the number of intruders, the approach maintains a high TTR rate.

5.2 Challenges and Limitations

Exploiting AI for the Space Understangind, in all its aspects, whether static or dynamic, comes with many challenges, including the variability of the data source, the management of heterogeneous data, the purpose of data processing. However, the more pronounced challenges related to application can be categorised as follows:

- **Lack of available dataset**: Regardless of the topic and/or the kind of data in the training phase (given the assumption that DL models can be arranged to fit a specific task), there is sometimes a lack of available datasets in the literature to be used as benchmarks. It is well known that ML and DL are data-driven technique that perform better as the number of input samples increases. Some attempts to

solve this problem have involved the generation of the datasets proposed in this thesis: ArCH dataset [83], SMART dataset [120] and TVPR2 dataset [158]. Recently, generative models have proven to be effective for this task. Generative adversarial networks are an appealing DL approach developed in 2014 by Goodfellow [222]. Generative adversarial networks are an unsupervised deep learning approach in which two neural networks challenge each other, and each of the two networks improves at its given task with each iteration. For the image generation issue, the generator begins with Gaussian noise to generate images, and the discriminator determines how valuable the generated images are. This process proceeds until the generator development outputs. Generative networks have been used to generate artificial images and videos as well as to generate point clouds, such as those used by the author of this thesis (PointGrow [80], PointFlow [81], and PointGMM [82]), in Section 3.1.3.

- **Domain dependant models**: When there is no all-in-one solution for every task, each AI-based model should be chosen according to the task one is attempting to solve. In other words, as AI improves, the need has emerged to understand how to make such models effective, choosing them according to the kind of data for which they have been designed. Integrating the knowledge of domain expert into AI models increases the reliability and the robustness of algorithms, making decisions more accurate. Some tests have been done in Section 3.1.2 by the author, using hadcrafted features designed by experts in the CH field. Moreover, the knowledge acquired for one task can be used to solve related ones by transfer learning strategies. Transfer learning allows to leverage knowledge (such as features, weights etc) from previously trained AI models for training newer models and even tackle problems like having less data for the newer task.

- **Data preprocessing**: Broadly speaking, spatial data have intrinsic features that make them very challenging for DL, especially convolutional neural networks. The reason for this is that AI is intended to use data that are ordered, regular, and on a structured grid. This means that data should be ordered, and pre-processing operations are still time consuming. Actually, this represents one of the main bottlnecks, as it requires the presence of an expert for every single application domain.

- **Hardware limitations**: despite the growing computational capabilities of better-performing CPUs and the advances in distributed and parallel high performance computing (HPC), the computational costs of the above-mentioned tasks remain high. We are not still at a stage where the ratio between time/gained and resources/spent is in balance, making the use of AI-based methods unhelpful at times compared with time-consuming but more affordable manual solutions.

5.3 Lesson Learnt

The AI-based solutions described in this thesis deserve some comments. DL techniques are delivering a promising solution to develop smart systems and to make innovation at a rapid pace. The combination of ICT technologies offers a framework for the understanding of any space and all that it contains, relying on different types of data (spatial, visual, temporal, etc.) gathered from a complex infrastructure of sensors and smart devices. Numerous challenges exist in implementing such a framework, one of them is to meet the data and services requirements on AI-based applications in terms of energy efficiency, sensing data quality, network resource consumption, and latency. For these reasons, several innovative but independent approaches have been described in this thesis, for solving the understanding of various aspects of a space.

For each of the aspects studied in this work, there were lessons learned. Firstly, the proposed approach to understanding space as a static concept defined the point that combining features of various kinds, whether raw or handcrafted, related to point cloud acquisition can lead to improved performance of any Deep Learning approach. The way the author designed the DGCNN-Mod network from DGCNN can be a starting point to modify any more recent method for Point Cloud Semantic Segmentation, regardless of the domain of work [223, 224].

In addition, the problem of unbalanced Point Cloud datasets in the state of the art remains unchanged, especially in the CH domain. The approach proposed by the author of the thesis in Section 3.1.3 may be the right way to solve this issue. The automatic generation of 3D class object can revolutionize the Cultural Heritage domain in several aspects. (H)BIM, given the uniqueness of CH objects, requires time-consuming parametrization that are nowadays completely entrusted to human operation; the proposed method opens up to great opportunities in terms of automatization. Moreover, as demonstrated by the literature, one of the main bottlenecks towards the full exploitation of DL methods for semantic segmentation is the lack of available and annotated datasets. Again, the proposed method proves to be a robust alternative to manual and time consuming manual labelling.

Secondly, the proposed Change Detection approach also leads the way to a new methodology that is not common in the state of the art, in which visual and textual features of the same acquisition are combined in order to obtain a final overall result. Recently, more innovative approaches have been created concerning both visual and textual feature extraction, including the OCR phase. It would be interesting to use more recent image classifiers, such as EfficientNet [225], and the latest text classifiers based on a recent concept called Transformer, like the T5 [226] developed by Google or the innovate GPT-3 [227] proposed by OpenAI. Therefore, the approach is certainly improvable in its single parts, but it remains an excellent global methodology to be used for the Change Detection task with RGB images.

Finally, with regard to the Person Re-Identification task, the combination of differ-

ent types of features (temporal, spatial and visual) led to better results than the state of the art methods. Again, the same methodology which uses all this information can be adapted to the latest state of the art approaches. Person Re-Identification in Top-View configuration has several advantages, such as the possibility to avoid occlusions between entities in the space, but unfortunately it is not very developed in the state of the art. For this reason, both the proposed approach by the author and the new acquired TVPR2 dataset, will certainly bring several advantages to the scientific community that would like to further investigate this field.

Chapter 6

Conclusions and Future Works

This thesis focused on the use of AI methods for Space Understanding in all its aspects. In fact, a space is not easy to describe and therefore should be studied by breaking it down into sub problems. Even if each use case considered presents different features and needs, it has been possible to outline a common path for its global comprehension. The studies were oriented towards understanding potentials and weak points, from a multidisciplinary perspective.

Certainly, AI is thoroughly changing several application domains. Different types of AI approaches were presented, covering the domain of Cultural Heritage, the world of Retail and finally that of Museums. This thesis provides insight on new trends, techniques, and methods for the global Space Understanding.

Chapter 1 portrays the concept of Space Understanding, starting from its definition as a static concept, to its dynamic aspects and all that follows. Indeed, the temporal evolution of both the space itself and the entities within it are aspects to be taken into account.

With Chapter 2, a review the state of art about methods, technology and applications has been provided with a specific focus on the state of the art for Point Clouds Semantic Segmentation task. In particular, for this field was analysed the methods and techniques, also main paths that most approaches follow are summarized and their contributions are pointed out. The reviewed approaches were categorized and compared from multiple perspectives, including methodology, function and analyse the pros and cons of each category. Then, a thorough survey of the literature related to the use of AI for the detection of changes inside a space, and its methods has been presented, with a particular focus on ML and DL approaches. Finally, specific emphasis has been given to re-identification approaches. In fact, this type of task allows the temporal evolution of entities within a space to be defined and tracked. In particular, approaches for person re-identification from a top-view perspective have been studied in order to avoid occlusions due to crowded environments. The reviewed approaches have been categorised and compared from multiple perspectives, pointing out their advantages and disadvantages.

In Chapter 3, the author describes all the methodologies presented to solve the various tasks concerning all aspects of Space Understanding. The first approach deals

with the semantic segmentation of point clouds, which is described emphasizing the innovations compared to the state of the art, the new dataset acquired for the task, and finally the metrics and experiments that have been carried out. Then, an innovative method for the Change Detection task is described, based on the use of multimodal information extracted from RGB images. Furthermore, emphasis is given to the various parts of this framework, as it is a mixed approach between Machine Learning methods and Neural Networks. Also for this task, a new dataset was acquired. Finally, the last remaining aspect to be considered concerns the temporal evolution of entities in a space. To solve this task, the author proposed a new temporal and multimodal method for the Person Re-Identification. In this section it is emphasised that the proposed approach performs well both in a closed world and in a more realistic open world environment. A further novelty concerns the publication of a new dataset of 1027 persons captured by RGB-D cameras.

Chapter 4 extensively describes all the experiments carried out, including the settings of the hyper-parameters of the approaches and the tools used.

Chapter 5 outlines needs, bottlenecks and weakness points for each case study of the thesis.

To conclude, it has been done the first steps in the description of a very complex concept such as Space Undestanding. Furthermore, the introduction and release of public datasets attract more colleagues from the AI community towards advancing the state of the art.

It has been interesting to investigate the integration of different types of data to solve the same problem, regardless of scope. Integration was tested in both semantic segmentation of point clouds and person re-identification, from the use of handcrafted features to the combination of spatial/temporal features. In addition, it was interesting to test generative approaches as a data augmentation technique for unbalanced datasets. The results are a starting point for the scientific community, which will certainly attract special attention.

This thesis thus paves the way for further research, in every considered aspect. With regard to sematic segmentation, future research could concern the implementation of the same methodology designed for the DGCNN-Mod on even more recent approaches. Back to generative approaches, it would be interesting to implemente novel methods which could generate additional features, such as colours, to make more realistic objects. Or, alternative approaches could be explored to counterbalance the datasets, such as the generation of synthetic data [63, 64]. The proposed Change Detection approach can be improved in its fundamental building blocks. Indeed, it would be interesting to use more recent image classifiers, such as EfficientNet [225], and the latest text classifiers based on a recent concept called Transformer [226, 227]. Finally, regarding the Person Re-Identification task, future works are focused on improving the performance of the proposed approach by integrating more complex CNN architectures. Incremental learning methods should also be investigated because could

allows to enhance the online performance of the proposed re-ID approach.

Bibliography

[1] C. C. Aggarwal and C. Zhai, "A survey of text classification algorithms," in Mining text data. Springer, 2012, pp. 163–222.

[2] K. Kowsari, K. Jafari Meimandi, M. Heidarysafa, S. Mendu, L. Barnes, and D. Brown, "Text classification algorithms: A survey," Information, vol. 10, no. 4, p. 150, 2019.

[3] D. L. Abd AL-Nabi and S. S. Ahmed, "Survey on classification algorithms for data mining:(comparison and evaluation)," International Journal of Computer Engineering and Intelligent Systems, vol. 4, no. 8, pp. 18–27, 2013.

[4] M. A. Fischler and R. A. Elschlager, "The representation and matching of pictorial structures," IEEE Transactions on computers, vol. 100, no. 1, pp. 67–92, 1973.

[5] L. Liu, W. Ouyang, X. Wang, P. Fieguth, J. Chen, X. Liu, and M. Pietikäinen, "Deep learning for generic object detection: A survey," International journal of computer vision, vol. 128, no. 2, pp. 261–318, 2020.

[6] H. Yu, Z. Yang, L. Tan, Y. Wang, W. Sun, M. Sun, and Y. Tang, "Methods and datasets on semantic segmentation: A review," Neurocomputing, vol. 304, pp. 82–103, 2018.

[7] K. Zhang, M. Hao, J. Wang, C. W. de Silva, and C. Fu, "Linked dynamic graph cnn: Learning on point cloud via linking hierarchical features," arXiv preprint arXiv:1904.10014, 2019.

[8] P. Tang, D. Huber, B. Akinci, R. Lipman, and A. Lytle, "Automatic reconstruction of as-built building information models from laser-scanned point clouds: A review of related techniques," Automation in construction, vol. 19, no. 7, pp. 829–843, 2010.

[9] M. Tamke, H. L. Evers, M. Zwierzycki, R. Wessel, S. Ochmann, R. Vock, and R. Klein, "An automated approach to the generation of structured building information models from unstructured 3d point cloud scans," in Proceedings of IASS Annual Symposia, vol. 2016, no. 17. International Association for Shell and Spatial Structures (IASS), 2016, pp. 1–10.

[10] H. Macher, T. Landes, and P. Grussenmeyer, "From point clouds to building information models: 3d semi-automatic reconstruction of indoors of existing buildings," Applied Sciences, vol. 7, no. 10, p. 1030, 2017.

[11] C. Thomson and J. Boehm, "Automatic geometry generation from point clouds for bim," Remote Sensing, vol. 7, no. 9, pp. 11 753–11 775, 2015.

[12] S. Shi, X. Wang, and H. Li, "Pointrcnn: 3d object proposal generation and detection from point cloud," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 770–779.

[13] Y. Zhou and O. Tuzel, "Voxelnet: End-to-end learning for point cloud based 3d object detection," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 4490–4499.

[14] Y. Xie, J. Tian, and X. Zhu, "A review of point cloud semantic segmentation," IEEE Geoscience and Remote Sensing Magazine (GRSM), 2020.

[15] S. Makuti, F. Nex, and M. Y. Yang, "Multi-temporal classification and change detection using uav images." International Archives of the Photogrammetry, Remote Sensing & Spatial Information Sciences, vol. 42, no. 2, 2018.

[16] E. Guo, X. Fu, J. Zhu, M. Deng, Y. Liu, Q. Zhu, and H. Li, "Learning to measure change: Fully convolutional siamese metric networks for scene change detection," arXiv preprint arXiv:1810.09111, 2018.

[17] K. Sakurada and T. Okatani, "Change detection from a street image pair using cnn features and superpixel segmentation." in BMVC, vol. 61, 2015, pp. 1–12.

[18] Y. Zhan, K. Fu, M. Yan, X. Sun, H. Wang, and X. Qiu, "Change detection based on deep siamese convolutional network for optical aerial images," IEEE Geoscience and Remote Sensing Letters, vol. 14, no. 10, pp. 1845–1849, 2017.

[19] J. Shi, J. Wang, and Y. Xu, "Object-based change detection using georeferenced uav images," Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci, vol. 38, no. 1/C22, pp. 177–182, 2011.

[20] S. D. Khan and H. Ullah, "A survey of advances in vision-based vehicle re-identification," Computer Vision and Image Understanding, vol. 182, pp. 50–63, 2019.

[21] P. C. Ravoor and T. Sudarshan, "Deep learning methods for multi-species animal re-identification and tracking–a survey," Computer Science Review, vol. 38, p. 100289, 2020.

[22] Q. Leng, M. Ye, and Q. Tian, "A survey of open-world person re-identification," IEEE Transactions on Circuits and Systems for Video Technology, 2019.

[23] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," International journal of computer vision, vol. 88, no. 2, pp. 303–338, 2010.

[24] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein et al., "Imagenet large scale visual recognition challenge," International journal of computer vision, vol. 115, no. 3, pp. 211–252, 2015.

[25] Z. Liang, Y. Guo, Y. Feng, W. Chen, L. Qiao, L. Zhou, J. Zhang, and H. Liu, "Stereo matching using multi-level cost volume and multi-scale feature constancy," IEEE transactions on pattern analysis and machine intelligence, vol. 43, no. 1, pp. 300–315, 2019.

[26] Y. Guo, F. Sohel, M. Bennamoun, M. Lu, and J. Wan, "Rotational projection statistics for 3d local surface description and object recognition," International journal of computer vision, vol. 105, no. 1, pp. 63–86, 2013.

[27] Y. Guo, M. Bennamoun, F. Sohel, M. Lu, and J. Wan, "3d object recognition in cluttered scenes with local surface features: A survey," IEEE transactions on pattern analysis and machine intelligence, vol. 36, no. 11, pp. 2270–2287, 2014.

[28] X. Chen, H. Ma, J. Wan, B. Li, and T. Xia, "Multi-view 3d object detection network for autonomous driving," in Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, 2017, pp. 1907–1915.

[29] Y. Xie, J. Tian, and X. X. Zhu, "Linking points with labels in 3d: A review of point cloud semantic segmentation," IEEE Geoscience and Remote Sensing Magazine, vol. 8, no. 4, pp. 38–59, 2020.

[30] E. M. Mikhail, J. S. Bethel, and J. C. McGlone, "Introduction to modern photogrammetry," New York, vol. 19, 2001.

[31] J. Shan and C. K. Toth, Topographic laser ranging and scanning: principles and processing. CRC press, 2018.

[32] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Nießner, "Scannet: Richly-annotated 3d reconstructions of indoor scenes," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 5828–5839.

[33] M. Shahzad and X. X. Zhu, "Robust reconstruction of building facades for large areas using spaceborne tomosar point clouds," IEEE Transactions on Geoscience and Remote Sensing, vol. 53, no. 2, pp. 752–769, 2014.

[34] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "Pointnet: Deep learning on point sets for 3d classification and segmentation," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 652–660.

[35] Z. Wu, S. Song, A. Khosla, F. Yu, L. Zhang, X. Tang, and J. Xiao, "3d shapenets: A deep representation for volumetric shapes," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 1912–1920.

[36] M. A. Uy, Q.-H. Pham, B.-S. Hua, T. Nguyen, and S.-K. Yeung, "Revisiting point cloud classification: A new benchmark dataset and classification model on real-world data," in Proceedings of the IEEE/CVF international conference on computer vision, 2019, pp. 1588–1597.

[37] A. X. Chang, T. Funkhouser, L. Guibas, P. Hanrahan, Q. Huang, Z. Li, S. Savarese, M. Savva, S. Song, H. Su et al., "Shapenet: An information-rich 3d model repository," arXiv preprint arXiv:1512.03012, 2015.

[38] K. Mo, S. Zhu, A. X. Chang, L. Yi, S. Tripathi, L. J. Guibas, and H. Su, "Partnet: A large-scale benchmark for fine-grained and hierarchical part-level 3d object understanding," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 909–918.

[39] I. Armeni, O. Sener, A. R. Zamir, H. Jiang, I. Brilakis, M. Fischer, and S. Savarese, "3d semantic parsing of large-scale indoor spaces," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 1534–1543.

[40] T. Hackel, N. Savinov, L. Ladicky, J. D. Wegner, K. Schindler, and M. Pollefeys, "Semantic3d. net: A new large-scale point cloud classification benchmark," arXiv preprint arXiv:1704.03847, 2017.

[41] X. Song, P. Wang, D. Zhou, R. Zhu, C. Guan, Y. Dai, H. Su, H. Li, and R. Yang, "Apollocar3d: A large 3d car instance understanding benchmark for autonomous driving," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 5452–5462.

[42] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in 2012 IEEE conference on computer vision and pattern recognition. IEEE, 2012, pp. 3354–3361.

[43] J. Behley, M. Garbade, A. Milioto, J. Quenzel, S. Behnke, C. Stachniss, and J. Gall, "Semantickitti: A dataset for semantic scene understanding of lidar sequences," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 9297–9307.

[44] G. Elbaz, T. Avraham, and A. Fischer, "3d point cloud registration for localization using a deep neural network auto-encoder," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 4631–4640.

[45] A. Zeng, K.-T. Yu, S. Song, D. Suo, E. Walker, A. Rodriguez, and J. Xiao, "Multi-view self-supervised deep learning for 6d pose estimation in the amazon picking challenge," in 2017 IEEE international conference on robotics and automation (ICRA). IEEE, 2017, pp. 1386–1383.

[46] X.-F. Han, H. Laga, and M. Bennamoun, "Image-based 3d object reconstruction: State-of-the-art and trends in the deep learning era," IEEE transactions on pattern analysis and machine intelligence, vol. 43, no. 5, pp. 1578–1604, 2019.

[47] A. Ioannidou, E. Chatzilari, S. Nikolopoulos, and I. Kompatsiaris, "Deep learning advances in computer vision with 3d data: A survey," ACM Computing Surveys (CSUR), vol. 50, no. 2, pp. 1–38, 2017.

[48] E. Ahmed, A. Saint, A. E. R. Shabayek, K. Cherenkova, R. Das, G. Gusev, D. Aouada, and B. Ottersten, "A survey on deep learning advances on different 3d data representations," arXiv preprint arXiv:1808.01462, 2018.

[49] M. M. Rahman, Y. Tan, J. Xue, and K. Lu, "Notice of violation of ieee publication principles: Recent advances in 3d object detection in the era of deep neural networks: A survey," IEEE Transactions on image processing, vol. 29, pp. 2947–2962, 2019.

[50] Y. Guo, H. Wang, Q. Hu, H. Liu, L. Liu, and M. Bennamoun, "Deep learning for 3d point clouds: A survey," IEEE transactions on pattern analysis and machine intelligence, vol. 43, no. 12, pp. 4338–4364, 2020.

[51] J. Llamas, P. M Lerones, R. Medina, E. Zalama, and J. Gómez-García-Bermejo, "Classification of architectural heritage images using deep learning techniques," Applied Sciences, vol. 7, no. 10, p. 992, 2017.

[52] E. Grilli, E. Özdemir, and F. Remondino, "Application of machine and deep learning strategies for the classification of heritage point clouds," The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences, vol. 42, pp. 447–454, 2019.

[53] E. Grilli and F. Remondino, "Classification of 3d digital heritage," Remote Sensing, vol. 11, no. 7, p. 847, 2019.

[54] N. Oses, F. Dornaika, and A. Moujahid, "Image-based delineation and classification of built heritage masonry," Remote Sensing, vol. 6, no. 3, pp. 1863–1889, 2014.

[55] B. Riveiro, P. B. Lourenço, D. V. Oliveira, H. González-Jorge, and P. Arias, "Automatic morphologic analysis of quasi-periodic masonry walls from lidar," Computer-Aided Civil and Infrastructure Engineering, vol. 31, no. 4, pp. 305–319, 2016.

[56] S. G. Barsanti, G. Guidi, and L. De Luca, "Segmentation of 3d models for cultural heritage structural analysis–some critical issues," ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, vol. 4, p. 115, 2017.

[57] F. Poux, R. Neuville, P. Hallot, and R. Billen, "Point cloud classification of tesserae from terrestrial laser data combined with dense image matching for archaeological information extraction," International Journal on Advances in Life Sciences, vol. 4, pp. 203–211, 2017.

[58] E. Grilli, D. Dininno, L. Marsicano, G. Petrucci, and F. Remondino, "Supervised segmentation of 3d cultural heritage," in 2018 3rd Digital Heritage International Congress (DigitalHERITAGE) held jointly with 2018 24th International Conference on Virtual Systems & Multimedia (VSMM 2018). IEEE, 2018, pp. 1–8.

[59] E. Grilli, E. Farella, A. Torresani, and F. Remondino, "Geometric features analysis for the classification of cultural heritage point clouds." International Archives of the Photogrammetry, Remote Sensing & Spatial Information Sciences, 2019.

[60] E. Grilli and F. Remondino, "Machine learning generalisation across different 3d architectural heritage," ISPRS International Journal of Geo-Information, vol. 9, no. 6, p. 379, 2020.

[61] A. Murtiyoso and P. Grussenmeyer, "Virtual disassembling of historical edifices: Experiments and assessments of an automatic approach for classifying multi-scalar point clouds into architectural elements," Sensors, vol. 20, no. 8, p. 2161, 2020.

[62] J. Zhang, X. Zhao, Z. Chen, and Z. Lu, "A review of deep learning-based semantic segmentation for point cloud (november 2019)," IEEE Access, 2019.

[63] R. Pierdicca, M. Mameli, E. S. Malinverni, M. Paolanti, and E. Frontoni, "Automatic generation of point cloud synthetic dataset for historical building representation," in International Conference on Augmented Reality, Virtual Reality and Computer Graphics. Springer, 2019, pp. 203–219.

[64] D. Griffiths and J. Boehm, "Synthcity: A large scale synthetic point cloud," arXiv preprint arXiv:1907.04758, 2019.

[65] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "Pointnet++: Deep hierarchical feature learning on point sets in a metric space," in Advances in neural information processing systems, 2017, pp. 5099–5108.

[66] E. Malinverni, R. Pierdicca, M. Paolanti, M. Martini, C. Morbidoni, F. Matrone, and A. Lingua, "Deep learning for semantic segmentation of 3d point cloud," International Archives of the Photogrammetry, Remote Sensing & Spatial Information Sciences, 2019.

[67] M. Atzmon, H. Maron, and Y. Lipman, "Point convolutional neural networks by extension operators," arXiv preprint arXiv:1803.10091, 2018.

[68] Y. Wang, Y. Sun, Z. Liu, S. E. Sarma, M. M. Bronstein, and J. M. Solomon, "Dynamic graph cnn for learning on point clouds," ACM Transactions on Graphics (TOG), vol. 38, no. 5, p. 146, 2019.

[69] R. Pierdicca, M. Paolanti, F. Matrone, M. Martini, C. Morbidoni, E. S. Malinverni, E. Frontoni, and A. M. Lingua, "Point cloud semantic segmentation using a deep learning framework for cultural heritage," Remote Sensing, vol. 12, no. 6, p. 1005, 2020.

[70] J. Llamas and M. Lerones, "P.; medina, r.; zalama, e.; gómez-garcía-bermejo," J. Classification of architectural heritage images using deep learning techniques. Appl. Sci, vol. 7, p. 992, 2017.

[71] M. S. Arshad and W. J. Beksi, "A progressive conditional generative adversarial network for generating dense and colored 3d point clouds," in 2020 International Conference on 3D Vision (3DV). IEEE, 2020, pp. 712–722.

[72] G. Myers, "A dataset generator for whole genome shotgun sequencing." in ISMB, 1999, pp. 202–210.

[73] H. Fan, H. Su, and L. J. Guibas, "A point set generation network for 3d object reconstruction from a single image," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 605–613.

[74] Z. Li, L. Zhang, R. Zhong, T. Fang, L. Zhang, and Z. Zhang, "Classification of urban point clouds: A robust supervised approach with automatically generating training data," IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 10, no. 3, pp. 1207–1220, 2016.

[75] B. Wu, X. Zhou, S. Zhao, X. Yue, and K. Keutzer, "Squeezesegv2: Improved model structure and unsupervised domain adaptation for road-object segmentation from a lidar point cloud," in 2019 International Conference on Robotics and Automation (ICRA). IEEE, 2019, pp. 4376–4382.

[76] F. Wang, Y. Zhuang, H. Gu, and H. Hu, "Automatic generation of synthetic lidar point clouds for 3-d data analysis," IEEE Transactions on Instrumentation and Measurement, vol. 68, no. 7, pp. 2671–2673, 2019.

[77] P. M. Chu, Y. Sung, and K. Cho, "Generative adversarial network-based method for transforming single rgb image into 3d point cloud," IEEE Access, vol. 7, pp. 1021–1029, 2018.

[78] V. Egiazarian, S. Ignatyev, A. Artemov, O. Voynov, A. Kravchenko, Y. Zheng, L. Velho, and E. Burnaev, "Latent-space laplacian pyramids for adversarial representation learning with 3d point clouds," arXiv preprint arXiv:1912.06466, 2019.

[79] J. Martínek, L. Lenc, and P. Král, "Training strategies for ocr systems for historical documents," in IFIP International Conference on Artificial Intelligence Applications and Innovations. Springer, 2019, pp. 362–373.

[80] Y. Sun, Y. Wang, Z. Liu, J. Siegel, and S. Sarma, "Pointgrow: Autoregressively learned point cloud generation with self-attention," in Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2020, pp. 61–70.

[81] G. Yang, X. Huang, Z. Hao, M.-Y. Liu, S. Belongie, and B. Hariharan, "Pointflow: 3d point cloud generation with continuous normalizing flows," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 4541–4550.

[82] A. Hertz, R. Hanocka, R. Giryes, and D. Cohen-Or, "Pointgmm: A neural gmm network for point clouds," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 12 054–12 063.

[83] F. Matrone, A. Lingua, R. Pierdicca, E. Malinverni, M. Paolanti, E. Grilli, F. Remondino, A. Murtiyoso, and T. Landes, "A benchmark for large-scale heritage point cloud semantic segmentation," 2020.

[84] R. J. Radke, S. Andra, O. Al-Kofahi, and B. Roysam, "Image change detection algorithms: a systematic survey," IEEE transactions on image processing, vol. 14, no. 3, pp. 294–307, 2005.

[85] R. C. Daudt, B. Le Saux, and A. Boulch, "Fully convolutional siamese networks for change detection," in 2018 25th IEEE International Conference on Image Processing (ICIP). IEEE, 2018, pp. 4063–4067.

[86] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in International Conference on Medical

image computing and computer-assisted intervention. Springer, 2015, pp. 234–241.

[87] P. F. Alcantarilla, S. Stent, G. Ros, R. Arroyo, and R. Gherardi, "Street-view change detection with deconvolutional networks," Autonomous Robots, vol. 42, no. 7, pp. 1301–1322, 2018.

[88] K. Sakurada, W. Wang, N. Kawaguchi, and R. Nakamura, "Dense optical flow based change detection network robust to difference of camera viewpoints," arXiv preprint arXiv:1712.02941, 2017.

[89] K. Sakurada, M. Shibuya, and W. Wang, "Weakly supervised silhouette-based semantic scene change detection," in 2020 IEEE International conference on robotics and automation (ICRA). IEEE, 2020, pp. 6861–6867.

[90] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv preprint arXiv:1409.1556, 2014.

[91] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[92] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in IEEE conference on computer vision and pattern recognition, 2015, pp. 3431–3440.

[93] Y. Lei, D. Peng, P. Zhang, Q. Ke, and H. Li, "Hierarchical paired channel fusion network for street scene change detection," IEEE Transactions on Image Processing, vol. 30, pp. 55–67, 2020.

[94] K. R. Prabhakar, A. Ramaswamy, S. Bhambri, J. Gubbi, R. V. Babu, and B. Purushothaman, "Cdnet++: Improved change detection with deep neural network feature correlation," in 2020 International Joint Conference on Neural Networks (IJCNN). IEEE, 2020, pp. 1–8.

[95] A. Puig-Pey, Y. Bolea, A. Grau, and J. Casanovas, "Public entities driven robotic innovation in urban areas," Robotics and Autonomous Systems, vol. 92, pp. 162 – 172, 2017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0921889016302792

[96] S. Kumar, G. Sharma, N. Kejriwal, S. Jain, M. Kamra, B. Singh, and V. K. Chauhan, "Remote retail monitoring and stock assessment using mobile robots," in International Conference on Technologies for Practical Robot Applications (TePRA). IEEE, 2014, pp. 1–6.

[97] E. Frontoni, P. Raspa, A. Mancini, P. Zingaretti, and V. Placidi, "Customers activity recognition in intelligent retail environments," in International Conference on Image Analysis and Processing. Springer, 2013, pp. 509–516.

[98] K. Kamei and et al., "Effectiveness of cooperative customer navigation from robots around a retail shop," in Third International Conference on Privacy, Security, Risk and Trust and IEEE Third International Conference on Social Computing, Boston, MA. IEEE, 2011, pp. 235–241.

[99] N. Matsuhira, F. Ozaki, S. Tokura, T. Sonoura, T. Tasaki, H. Ogawa, M. Sano, A. Numata, N. Hashimoto, and K. Komoriya, "Development of robotic transportation system-shopping support system collaborating with environmental cameras and mobile robots," in 41st International Symposium on Robotics (ISR) and 6th German Conference on Robotics (ROBOTIK). VDE, 2010, pp. 1–6.

[100] V. Kulyukin, C. Gharpure, and J. Nicholson, "Robocart: Toward robot-assisted navigation of grocery stores by the visually impaired," in IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2005, pp. 2845–2850.

[101] R. Limosani, R. Esposito, A. Manzi, G. Teti, F. Cavallo, and P. Dario, "Robotic delivery service in combined outdoor–indoor environments: technical analysis and user evaluation," Robotics and Autonomous Systems, vol. 103, pp. 56 – 67, 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0921889016302457

[102] A. Bancroft and C. W. Ward, "Methods for facilitating a retail environment," Apr. 17 2007, uS Patent 7,206,753.

[103] A. Purohit, Z. Sun, S. Pan, and P. Zhang, "Sugartrail: Indoor navigation in retail environments without surveys and maps," in 10th Annual Communications Society Conference on Sensor, Mesh and Ad Hoc Communications and Networks (SECON). IEEE, 2013, pp. 300–308.

[104] M. Paolanti, D. Liciotti, R. Pietrini, A. Mancini, and E. Frontoni, "Modelling and forecasting customer navigation in intelligent retail environments," Journal of Intelligent & Robotic Systems, 2017.

[105] A. Mancini, E. Frontoni, P. Zingaretti, and V. Placidi, "Smart vision system for shelf analysis in intelligent retail environments," in ASME 2013 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, American Society of Mechanical Engineers, Portland, OR, USA, vol. 47, 2013.

[106] D. Liciotti, M. Contigiani, E. Frontoni, A. Mancini, P. Zingaretti, and V. Placidi, "Shopper analytics: A customer activity recognition system using a distributed rgb-d camera network," in International workshop on video analytics for audience measurement in retail and digital signage. Springer, 2014, pp. 146–157.

[107] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, "Backpropagation applied to handwritten zip code recognition," Neural computation, vol. 1, no. 4, pp. 541–551, 1989.

[108] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in Advances in neural information processing systems, 2012, pp. 1097–1105.

[109] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in European conference on computer vision. Springer, 2014, pp. 818–833.

[110] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun, "Overfeat: Integrated recognition, localization and detection using convolutional networks," in International Conference on Learning Representations, 2014.

[111] J. Dean, G. Corrado, R. Monga, K. Chen, M. Devin, M. Mao, A. Senior, P. Tucker, K. Yang, Q. V. Le et al., "Large scale distributed deep networks," in Advances in neural information processing systems, 2012, pp. 1223–1231.

[112] P. Sermanet, S. Chintala, and Y. LeCun, "Convolutional neural networks applied to house numbers digit classifications," in International Conference on Pattern Recognition, 2012.

[113] P. Sermanet and Y. LeCun, "Traffic sign recognition with multi-scale convolutional networks," in International Joint Conference on Neural Networks. IEEE, 2011, pp. 2809–2813.

[114] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in IEEE conference on computer vision and pattern recognition, 2014, pp. 580–587.

[115] A. Toshev and C. Szegedy, "Deeppose: Human pose estimation via deep neural networks," in IEEE conference on computer vision and pattern recognition, 2014, pp. 1653–1660.

[116] G. J. Scott, R. A. Marcum, C. H. Davis, and T. W. Nivin, "Fusion of deep convolutional neural networks for land cover classification of high-resolution imagery," Geoscience and Remote Sensing Letters, vol. 14, no. 9, pp. 1638–1642, 2017.

[117] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," in 22nd international conference on Multimedia. ACM, 2014, pp. 675–678.

[118] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich et al., "Going deeper with convolutions." Conference on Computer Vision and Pattern Recognition, 2015.

[119] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[120] M. Paolanti, M. Sturari, A. Mancini, P. Zingaretti, and E. Frontoni, "Mobile robot for retail surveying and inventory using visual and textual analysis of monocular pictures based on deep learning," in European Conference on Mobile Robots (ECMR). IEEE, 2017, pp. 1–6.

[121] G. Adaimi, S. Kreiss, and A. Alahi, "Deep visual re-identification with confidence," Transportation research part C: emerging technologies, vol. 126, p. 103067, 2021.

[122] A. Wu, W.-S. Zheng, and J.-H. Lai, "Robust depth-based person re-identification," IEEE Transactions on Image Processing, vol. 26, no. 6, pp. 2588–2603, 2017.

[123] N. Karianakis, Z. Liu, Y. Chen, and S. Soatto, "Person depth reid: Robust person re-identification with commodity depth sensors," arXiv preprint arXiv:1705.09882, 2017.

[124] A. Wu, W.-S. Zheng, H.-X. Yu, S. Gong, and J. Lai, "Rgb-infrared cross-modality person re-identification," in Proceedings of the IEEE international conference on computer vision, 2017, pp. 5380–5389.

[125] J. Zhao, H. Xu, H. Liu, J. Wu, Y. Zheng, and D. Wu, "Detection and tracking of pedestrians and vehicles using roadside lidar sensors," Transportation research part C: emerging technologies, vol. 100, pp. 68–87, 2019.

[126] B. Abdulhai and S. M. Tabib, "Spatio-temporal inductance-pattern recognition for vehicle re-identification," Transportation Research Part C: Emerging Technologies, vol. 11, no. 3-4, pp. 223–239, 2003.

[127] S. Gong, M. Cristani, C. C. Loy, and T. M. Hospedales, "The re-identification challenge," in Person re-identification. Springer, 2014, pp. 1–20.

[128] L. Zheng, Y. Yang, and A. G. Hauptmann, "Person re-identification: Past, present and future," arXiv preprint arXiv:1610.02984, 2016.

[129] F. Zheng, C. Deng, X. Sun, X. Jiang, X. Guo, Z. Yu, F. Huang, and R. Ji, "Pyramidal person re-identification via multi-loss dynamic training," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 8514–8522.

[130] M. Paolanti, L. Romeo, M. Martini, A. Mancini, E. Frontoni, and P. Zingaretti, "Robotic retail surveying by deep learning visual and textual data," Robotics and Autonomous Systems, vol. 118, pp. 179–188, 2019.

[131] N. Ferracuti, C. Norscini, E. Frontoni, P. Gabellini, M. Paolanti, and V. Placidi, "A business application of rtls technology in intelligent retail environment: Defining the shopper's preferred path and its segmentation," Journal of Retailing and Consumer Services, vol. 47, pp. 184–194, 2019.

[132] D. Liciotti, M. Paolanti, R. Pietrini, E. Frontoni, and P. Zingaretti, "Convolutional networks for semantic heads segmentation using top-view depth data in crowded environment," in 2018 24th International Conference on Pattern Recognition (ICPR). IEEE, 2018, pp. 1384–1389.

[133] W.-S. Zheng, S. Gong, and T. Xiang, "Towards open-world person re-identification by one-shot group-based verification," IEEE transactions on pattern analysis and machine intelligence, vol. 38, no. 3, pp. 591–606, 2015.

[134] S. Liao, Z. Mo, J. Zhu, Y. Hu, and S. Z. Li, "Open-set person re-identification," arXiv preprint arXiv:1408.0872, 2014.

[135] X. Zhu, B. Wu, D. Huang, and W.-S. Zheng, "Fast open-world person re-identification," IEEE Transactions on Image Processing, vol. 27, no. 5, pp. 2286–2300, 2017.

[136] P. Dollár, Z. Tu, H. Tao, and S. Belongie, "Feature mining for image classification," in 2007 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2007, pp. 1–8.

[137] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani, "Person re-identification by symmetry-driven accumulation of local features," in 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. IEEE, 2010, pp. 2360–2367.

[138] D. Gray and H. Tao, "Viewpoint invariant pedestrian recognition with an ensemble of localized features," in European conference on computer vision. Springer, 2008, pp. 262–275.

[139] B. Ma, Y. Su, and F. Jurie, "Local descriptors encoded by fisher vectors for person re-identification," in European Conference on Computer Vision. Springer, 2012, pp. 413–422.

[140] R. Zhao, W. Ouyang, and X. Wang, "Unsupervised salience learning for person re-identification," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 3586–3593.

[141] M. Hirzer, P. M. Roth, M. Köstinger, and H. Bischof, "Relaxed pairwise learned metric for person re-identification," in European conference on computer vision. Springer, 2012, pp. 780–793.

[142] W.-S. Zheng, S. Gong, and T. Xiang, "Reidentification by relative distance comparison," IEEE transactions on pattern analysis and machine intelligence, vol. 35, no. 3, pp. 653–668, 2012.

[143] D. Liciotti, M. Paolanti, E. Frontoni, and P. Zingaretti, "People detection and tracking from an rgb-d camera in top-view configuration: review of challenges and applications," in International Conference on Image Analysis and Processing. Springer, 2017, pp. 207–218.

[144] M. Paolanti, L. Romeo, D. Liciotti, R. Pietrini, A. Cenci, E. Frontoni, and P. Zingaretti, "Person re-identification with rgb-d camera in top-view configuration through multiple nearest neighbor classifiers and neighborhood component features selection," Sensors, vol. 18, no. 10, p. 3471, 2018.

[145] R. Vezzani, D. Baltieri, and R. Cucchiara, "People reidentification in surveillance and forensics: A survey," ACM Computing Surveys (CSUR), vol. 46, no. 2, p. 29, 2013.

[146] X. Li, A. Wu, and W.-S. Zheng, "Adversarial open-world person re-identification," in Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 280–296.

[147] A. Bedagkar-Gala and S. K. Shah, "A survey of approaches and trends in person re-identification," Image and Vision Computing, vol. 32, no. 4, pp. 270–286, 2014.

[148] W.-S. Zheng, S. Gong, and T. Xiang, "Transfer re-identification: From person to set-based verification," in 2012 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2012, pp. 2650–2657.

[149] B. Cancela, T. M. Hospedales, and S. Gong, "Open-world person re-identification by multi-label assignment inference." 2014.

[150] H. Wang, X. Zhu, T. Xiang, and S. Gong, "Towards unsupervised open-set person re-identification," in 2016 IEEE International Conference on Image Processing (ICIP). IEEE, 2016, pp. 769–773.

[151] D. Liciotti, M. Paolanti, E. Frontoni, A. Mancini, and P. Zingaretti, "Person re-identification dataset with rgb-d camera in a top-view configuration," in Video Analytics. Face and Facial Expression Recognition and Audience Measurement. Springer, 2016, pp. 1–11.

[152] A. Haque, A. Alahi, and L. Fei-Fei, "Recurrent attention models for depth-based person identification," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 1229–1238.

[153] A. R. Lejbolle, K. Nasrollahi, B. Krogh, and T. B. Moeslund, "Multimodal neural network for overhead person re-identification," in 2017 International Conference of the Biometrics Special Interest Group (BIOSIG). IEEE, 2017, pp. 1–5.

[154] A. R. Lejbolle, B. Krogh, K. Nasrollahi, and T. B. Moeslund, "Attention in multimodal neural networks for person re-identification," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2018, pp. 179–187.

[155] A. R. Lejbølle, K. Nasrollahi, B. Krogh, and T. B. Moeslund, "Person re-identification using spatial and layer-wise attention," IEEE Transactions on Information Forensics and Security, 2019.

[156] A. Hermans, L. Beyer, and B. Leibe, "In defense of the triplet loss for person re-identification," arXiv preprint arXiv:1703.07737, 2017.

[157] Y. Yuan, W. Chen, Y. Yang, and Z. Wang, "In defense of the triplet loss again: Learning robust person re-identification with fast approximated triplet loss and label distillation," arXiv preprint arXiv:1912.07863, 2019.

[158] M. Martini, M. Paolanti, and E. Frontoni, "Open-world person re-identification with rgbd camera in top-view configuration for retail applications," IEEE Access, vol. 8, pp. 67 756–67 765, 2020.

[159] R. Quattrini, R. Pierdicca, M. Paolanti, P. Clini, R. Nespeca, and E. Frontoni, "Digital interaction with 3d archaeological artefacts: evaluating user's behaviours at different representation scales," Digital Applications in Archaeology and Cultural Heritage, vol. 18, p. e00148, 2020.

[160] L. Nisiotis, L. Alboul, and M. Beer, "A prototype that fuses virtual reality, robots, and social networks to create a new cyber–physical–social eco-society system for cultural heritage," Sustainability, vol. 12, no. 2, p. 645, 2020.

[161] S. Karaman, A. D. Bagdanov, L. Landucci, G. D'Amico, A. Ferracani, D. Pezzatini, and A. Del Bimbo, "Personalized multimedia content delivery on an

interactive table by passive observation of museum visitors," Multimedia Tools and Applications, vol. 75, no. 7, pp. 3787–3811, 2016.

[162] S. Alletto, R. Cucchiara, G. Del Fiore, L. Mainetti, V. Mighali, L. Patrono, and G. Serra, "An indoor location-aware system for an iot-based smart museum," IEEE Internet of Things Journal, vol. 3, no. 2, pp. 244–253, 2015.

[163] J. Lanir, T. Kuflik, J. Sheidin, N. Yavin, K. Leiderman, and M. Segal, "Visualizing museum visitors' behavior: Where do they go and what do they do there?" Personal and Ubiquitous Computing, vol. 21, no. 2, pp. 313–326, 2017.

[164] A. Chianese and F. Piccialli, "Designing a smart museum: When cultural heritage joins iot," in 2014 eighth international conference on next generation mobile apps, services and technologies. IEEE, 2014, pp. 300–306.

[165] G. Del Fiore, L. Mainetti, V. Mighali, L. Patrono, S. Alletto, R. Cucchiara, and G. Serra, "A location-aware architecture for an iot-based smart museum," International Journal of Electronic Government Research (IJEGR), vol. 12, no. 2, pp. 39–55, 2016.

[166] M. Paolanti, R. Pietrini, A. Mancini, E. Frontoni, and P. Zingaretti, "Deep understanding of shopper behaviours and interactions using rgb-d vision," Machine Vision and Applications, vol. 31, no. 7, pp. 1–21, 2020.

[167] D. Liciotti, M. Paolanti, R. Pietrini, E. Frontoni, and P. Zingaretti, "Convolutional networks for semantic heads segmentation using top-view depth data in crowded environment," in Pattern Recognition (ICPR), 2018 24rd International Conference on. IEEE, 2018.

[168] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The kitti dataset," International Journal of Robotics Research (IJRR), 2013.

[169] M. De Deuge, A. Quadros, C. Hung, and B. Douillard, "Unsupervised feature learning for classification of outdoor 3d scans," in Australasian Conference on Robitics and Automation, vol. 2, 2013, p. 1.

[170] T. Hackel, N. Savinov, L. Ladicky, J. D. Wegner, K. Schindler, and M. Pollefeys, "SEMANTIC3D.NET: A new large-scale point cloud classification benchmark," in ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, vol. IV-1-W1, 2017, pp. 91–98.

[171] D. Munoz, J. A. Bagnell, N. Vandapel, and M. Hebert, "Contextual classification with functional max-margin markov networks," in 2009 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2009, pp. 975–982.

[172] F. Matrone, A. Lingua, R. Pierdicca, E. S. Malinverni, M. Paolanti, E. Grilli, F. Remondino, A. Murtiyoso, and T. Landes, "A benchmark for large-scale heritage point cloud semantic segmentation," ISPRS Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci. 2020, XLIII-B2, vol. 2, no. W4, pp. 4558—-4567, 2020.

[173] X.-F. Hana, J. S. Jin, J. Xie, M.-J. Wang, and W. Jiang, "A comprehensive review of 3d point cloud descriptors," arXiv preprint arXiv:1802.02297, 2018.

[174] M. Weinmann, B. Jutzi, S. Hinz, and C. Mallet, "Semantic point cloud interpretation based on optimal neighborhoods, relevant features and efficient classifiers," ISPRS Journal of Photogrammetry and Remote Sensing, vol. 105, pp. 286–304, 2015.

[175] S. Sural, G. Qian, and S. Pramanik, "Segmentation and histogram generation using the hsv color space for image retrieval," in Proceedings. International Conference on Image Processing, vol. 2. IEEE, 2002, pp. II–II.

[176] F. Matrone, E. Grilli, M. Martini, M. Paolanti, R. Pierdicca, and F. Remondino, "Comparing machine and deep learning methods for large 3d heritage semantic segmentation," ISPRS International Journal of Geo-Information, vol. 9, no. 9, p. 535, 2020.

[177] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The kitti dataset," The International Journal of Robotics Research, vol. 32, no. 11, pp. 1231–1237, 2013.

[178] B. Chen, S. Shi, W. Gong, Q. Zhang, J. Yang, L. Du, J. Sun, Z. Zhang, and S. Song, "Multispectral lidar point cloud classification: A two-step approach," Remote Sensing, vol. 9, no. 4, p. 373, 2017.

[179] J. Zhang, X. Lin, and X. Ning, "Svm-based classification of segmented airborne lidar point clouds in urban areas," Remote Sensing, vol. 5, no. 8, pp. 3749–3775, 2013.

[180] P. Laube, M. O. Franz, and G. Umlauf, "Evaluation of features for svm-based classification of geometric primitives in point clouds," in 2017 Fifteenth IAPR International Conference on Machine Vision Applications (MVA). IEEE, 2017, pp. 59–62.

[181] P. Babahajiani, L. Fan, and M. Gabbouj, "Object recognition in 3d point cloud of urban street scene," in Asian Conference on Computer Vision. Springer, 2014, pp. 177–190.

[182] Z. Li, L. Zhang, X. Tong, B. Du, Y. Wang, L. Zhang, Z. Zhang, H. Liu, J. Mei, X. Xing et al., "A three-step approach for tls point cloud classification," IEEE

Transactions on Geoscience and Remote Sensing, vol. 54, no. 9, pp. 5412–5424, 2016.

[183] S. K. Lodha, D. M. Fitzpatrick, and D. P. Helmbold, "Aerial lidar data classification using adaboost," in Sixth International Conference on 3-D Digital Imaging and Modeling (3DIM 2007). IEEE, 2007, pp. 435–442.

[184] Y. Liu, M. Aleksandrov, S. Zlatanova, J. Zhang, F. Mo, and X. Chen, "Classification of power facility point clouds from unmanned aerial vehicles based on adaboost and topological constraints," Sensors, vol. 19, no. 21, p. 4717, 2019.

[185] Z. Kang, J. Yang, and R. Zhong, "A bayesian-network-based classification method integrating airborne lidar data with optical images," IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 10, no. 4, pp. 1651–1661, 2016.

[186] D. R. Thompson, E. J. Hochberg, G. P. Asner, R. O. Green, D. E. Knapp, B.-C. Gao, R. Garcia, M. Gierach, Z. Lee, S. Maritorena et al., "Airborne mapping of benthic reflectance spectra with bayesian linear mixtures," Remote Sensing of Environment, vol. 200, pp. 18–30, 2017.

[187] M. Belgiu and L. Drăguţ, "Random forest in remote sensing: A review of applications and future directions," ISPRS Journal of Photogrammetry and Remote Sensing, vol. 114, pp. 24–31, 2016.

[188] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg et al., "Scikit-learn: Machine learning in python," the Journal of machine Learning research, vol. 12, pp. 2825–2830, 2011.

[189] G. H. John and P. Langley, "Estimating continuous distributions in bayesian classifiers," arXiv preprint arXiv:1302.4964, 2013.

[190] N. Chehata, L. Guo, and C. Mallet, "Airborne lidar feature selection for urban classification using random forests," 2009.

[191] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in Proceedings of the IEEE international conference on computer vision, 2017, pp. 2980–2988.

[192] R. Pierdicca, M. Paolanti, R. Quattrini, M. Martini, E. S. Malinverni, and E. Frontoni, "Generative networks for point cloud generation in cultural heritage," in Proceedings of the of the joint international event 9th ARQUEOLÓGICA 2.0 3rd GEORES, 2021.

[193] P. Achlioptas, O. Diamanti, I. Mitliagkas, and L. Guibas, "Learning representations and generative models for 3d point clouds," in International conference on machine learning. PMLR, 2018, pp. 40–49.

[194] E. Frontoni, F. Marinelli, R. Rosetti, and P. Zingaretti, "Shelf space reallocation for out of stock reduction," Computers & Industrial Engineering, vol. 106, pp. 32–40, 2017.

[195] T. Gruen, D. Corsten, and G. M. of America, A Comprehensive Guide to Retail Out-of-stock Reduction in the Fast-moving Consumer Goods Industry. Grocery Manufacturers of America, 2007. [Online]. Available: https://books.google.it/books?id=AmK4MgAACAAJ

[196] M. Paolanti, C. Kaiser, R. Schallner, E. Frontoni, and P. Zingaretti, "Visual and textual sentiment analysis of brand-related social media pictures using deep convolutional neural networks," in 19th International Conference on Image Analysis and Processing, 2017.

[197] X. Zhang, J. Zhao, and Y. LeCun, "Character-level convolutional networks for text classification," in Advances in neural information processing systems, 2015, pp. 649–657.

[198] T. H. Bø, B. Dysvik, and I. Jonassen, "Lsimpute: accurate estimation of missing values in microarray data with least squares methods," Nucleic acids research, vol. 32, no. 3, pp. e34–e34, 2004.

[199] O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, and R. B. Altman, "Missing value estimation methods for dna microarrays," Bioinformatics, vol. 17, no. 6, pp. 520–525, 2001.

[200] C. Cortes and V. Vapnik, "Support-vector networks," Machine learning, vol. 20, no. 3, pp. 273–297, 1995.

[201] J. R. Quinlan, "Induction of decision trees," Machine learning, vol. 1, no. 1, pp. 81–106, 1986.

[202] L. Breiman, "Random forests," Machine learning, vol. 45, no. 1, pp. 5–32, 2001.

[203] I. Rish, "An empirical study of the naive bayes classifier," in IJCAI 2001 workshop on empirical methods in artificial intelligence, vol. 3, no. 22. IBM, 2001, pp. 41–46.

[204] R. Lippmann, "An introduction to computing with neural nets," Assp magazine, vol. 4, no. 2, pp. 4–22, 1987.

[205] M. Liao, B. Shi, X. Bai, X. Wang, and W. Liu, "Textboxes: A fast text detector with a single deep neural network," arXiv preprint arXiv:1611.06779, 2016.

[206] M. Jaderberg, K. Simonyan, A. Vedaldi, and A. Zisserman, "Reading text in the wild with convolutional neural networks," International Journal of Computer Vision, vol. 116, no. 1, pp. 1–20, 2016.

[207] Y. Kim, Y. Jernite, D. Sontag, and A. M. Rush, "Character-aware neural language models," in Association for the Advancement of Artificial Intelligence, 2016, pp. 2741–2749.

[208] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in 2009 IEEE conference on computer vision and pattern recognition. Ieee, 2009, pp. 248–255.

[209] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in European conference on computer vision. Springer, 2016, pp. 630–645.

[210] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," 2016.

[211] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," 2014.

[212] A. Eitel, J. T. Springenberg, L. Spinello, M. Riedmiller, and W. Burgard, "Multimodal deep learning for robust rgb-d object recognition," in Intelligent Robots and Systems (IROS), 2015 IEEE/RSJ International Conference on. IEEE, 2015, pp. 681–687.

[213] C. Hazirbas, L. Ma, C. Domokos, and D. Cremers, "Fusenet: Incorporating depth into semantic segmentation via fusion-based cnn architecture," in Asian conference on computer vision. Springer, 2016, pp. 213–228.

[214] J. Gao and R. Nevatia, "Revisiting temporal modeling for video-based person reid," arXiv preprint arXiv:1805.02104, 2018.

[215] K. Hara, H. Kataoka, and Y. Satoh, "Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet?" in Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, 2018, pp. 6546–6555.

[216] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev et al., "The kinetics human action video dataset," arXiv preprint arXiv:1705.06950, 2017.

[217] S. Gupta, R. Girshick, P. Arbeláez, and J. Malik, "Learning rich features from rgb-d images for object detection and segmentation," in European Conference on Computer Vision. Springer, 2014, pp. 345–360.

[218] Y. Liu, J. Yan, and W. Ouyang, "Quality aware network for set to set recognition," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 5790–5799.

[219] Z. Zhou, Y. Huang, W. Wang, L. Wang, and T. Tan, "See the forest for the trees: Joint spatial and temporal recurrent neural networks for video-based person re-identification," in Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on. IEEE, 2017, pp. 6776–6785.

[220] N. McLaughlin, J. Martinez del Rincon, and P. Miller, "Recurrent convolutional network for video-based person re-identification," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 1325–1334.

[221] Y. Yan, B. Ni, Z. Song, C. Ma, Y. Yan, and X. Yang, "Person re-identification via recurrent feature aggregation," in European Conference on Computer Vision. Springer, 2016, pp. 701–716.

[222] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in Advances in neural information processing systems, 2014, pp. 2672–2680.

[223] H. Thomas, C. R. Qi, J.-E. Deschaud, B. Marcotegui, F. Goulette, and L. J. Guibas, "Kpconv: Flexible and deformable convolution for point clouds," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 6411–6420.

[224] Q. Hu, B. Yang, L. Xie, S. Rosa, Y. Guo, Z. Wang, N. Trigoni, and A. Markham, "Randla-net: Efficient semantic segmentation of large-scale point clouds," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 11 108–11 117.

[225] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in International Conference on Machine Learning. PMLR, 2019, pp. 6105–6114.

[226] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," arXiv preprint arXiv:1910.10683, 2019.

[227] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell et al., "Language models are few-shot learners," arXiv preprint arXiv:2005.14165, 2020.