Università degli Studi di Milano-Bicocca

Department of Informatics, Systems and Communication

Master's degree in Data Science

Predicting Census Area Typologies Using Deep Learning solutions:

A Satellite Image-Based Approach to Urban Classification

Antonio Mastroianni

898723

Supervisor: Prof. **Simone Bianco**

Co-supervisor: Dott. **Stefano Biondi**

Academic Year

2023-2024

Contents

Chapter 1

Introduction

When working with territorial data, especially in the contest of the city, it is crucial to rely on the most fine-grained and standardized territorial unit to ensure accuracy in analysis and decision-making. In Italy, this unit consists of census areas provided by the Istituto Nazionale di Statistica (ISTAT), which serve as the official divisions for urban and regional analysis throughout all Italian territory. These census areas allow businesses, policymakers, and researchers to systematically study various aspects of city development, land use, and economic activity. During my internship at Generali Real Estate as Data Scientist, ISTAT census areas played a central role in analytical processes, providing the foundation for location-based assessments and market segmentation. These areas, which are classified for their land use, are only updated over a 10 years time span meaning that they often become outdated in the face of rapid urban transformation. This limitation poses a significant challenge for businesses relying on accurate and up-to-date data, particularly in the real estate sector.

To address this issue, a workflow enabling real-time or on-demand updates of census area labels was sought, ensuring they accurately reflect present urban conditions. This research focuses on developing a data-driven solution capable of verifying and, when necessary, reclassifying census areas using modern computational techniques. By integrating deep learning with publicly available mapping data, a methodology is proposed to improve the ac-

curacy and relevance of census classifications while minimizing the manual effort required for verification.

To achieve this, a dual approach was implemented that categorized census areas into two distinct groups based on the availability of external validation sources. The first category consists of classes which represent places that can be identified by direct points of interest. These classes correspond to locations with fixed, well-documented landmarks that can be cross-checked with geospatial data. The second category includes land-use classifications that cannot be directly verified due to absence of points of interest that allow their direct identification necessitating an image-based classification approach using satellite imagery and deep learning models. For the first category, the Google Maps Places API was leveraged to validate census area classifications based on the presence or absence of expected points of interest. This approach provides an efficient means of automating the validation process for a subset of census classes that correspond to specific physical structures. For the second category, where classifications cannot be validated through external databases, a deep learning-based classification model trained on satellite imagery was developed. The challenge in this approach lay in obtaining reliable labeled training data. Since prior knowledge indicated that some census areas had incorrect labels, an innovative approach was employed that combined automated and human-verified labeling. A GPT-based model was used to generate initial classification labels for satellite images of Milan's census areas. These labels were then cross-checked against ISTAT's official classifications, and any discrepancies were manually reviewed to ensure accuracy. The confirmed labels were used to fine-tune a deep neural network model capable of identifying land-use patterns from satellite imagery. The business implications of this work are substantial. In the real estate industry, access to up-to-date urban classifications is critical for numerous applications, including property valuation, risk assessment, and investment decision-making. An outdated classification system can lead to misinformed investment strategies, inaccurate pricing models, and missed opportunities in emerging neighborhoods. By establishing a semi-automated workflow for updating census area

classifications, a company like Generali Real Estate can improve their market intelligence, gain a competitive advantage, and optimize resource allocation. Moreover, city planners and policymakers can benefit from this system, as it offers a potentially scalable and cost-effective solution for monitoring urban change without waiting for decadal census updates. While this research focuses on the city of Milan as a case study, the proposed methodology has the potential to be extended to the entire Italian territory. Given that ISTAT census areas follow the same classification principles nationwide, the approach developed in this work can be adapted to other cities and regions across Italy. By applying this workflow at a broader scale, a nationwide system for continuously updating census classifications could be established, benefiting multiple industries, urban planning authorities, and policymakers. The scalability of this approach underscores its significance beyond Milan, demonstrating its applicability to urban classification challenges throughout Italy.

1.1 Outline

This thesis is structured into five main chapters including introduction. The second chapter focuses on the background and theoretical framework, reviewing relevant literature on census classification, remote sensing, and deep learning applications in urban analysis. It also explains ISTAT's census area codification system and its impact on urban planning and real estate. The third chapter discusses dataset construction and preprocessing, beginning with an overview of census data and satellite imagery sources. It then defines census area typologies and the validation approach, detailing the data collection and preprocessing steps. Additionally, this section explains the use of GPT models for automatic labeling of satellite images and the process for validation and dataset refinement. The fourth chapter covers methodology and model development, addressing data imbalance strategies, deep learning model selection, and the training and optimization process. It also presents the results and evaluation of the classification models, analyzing their performance and accuracy. The final chapter concludes the thesis by summarizing key findings and their implications for real estate and urban planning. It discusses potential improvements and future applications of the methodology, with a consideration of extending the approach to other Italian cities and regions.

Chapter 2

Background and Theoretical Framework

This chapter provides an overview of the foundational concepts relevant to this research, outlining ISTAT's census area coding system and examining pertinent literature within the context of urban classification, remote sensing, and deep learning. The problem statement is also outlined, delineating the key issues motivating this research and placing it within the larger academic literature context.

2.1 ISTAT Census Area Codification System

ISTAT's census areas – typically referred to as sezioni di censimento -are the smallest territorial units used for census data collection and dissemination[8]. They are defined as contiguous geographic units bounded by clearly identifiable features. Each census section forms a closed polygon covering part of a municipality without overlapping others, ensuring complete territorial partitioning. By design, a census section never straddles different localities or administrative subdivisions; instead, it is wholly contained within a single inhabited locality and respects any sub-municipal boundaries. ISTAT assigns every section a unique code within its municipality, ensuring distinct identification and facilitating statistical aggregation. The delineation of census

areas is grounded in both demographic and geographic criteria. Historically, sections were drawn to be manageable areas for enumerators—covering a practical number of households—using local maps and on-the-ground knowledge. In modern updates, ISTAT relies on digital cartography and administrative data to refine these boundaries. For example, in preparation for the 2021 census, ISTAT significantly updated the territorial basemap, increasing the number of census sections from approximately 403,000 in 2011 to over 756,000 in 2021, an 88% increase [9]. ISTAT also categorizes census sections based on land use, ensuring that residential, commercial, industrial, and rural areas are appropriately distinguished. In particular, it defines a total of 11 macro-categories and 53 specific land-use categories. This classification facilitates more precise statistical analyses and urban planning efforts. The use of high-resolution base maps [12] and local administrative boundaries ensures that section perimeters align with real-world features and administrative realities. As confirmed by image 2.1, ISTAT census sections are carefully crafted units: contiguous, covering all territory without gaps, delineated using physical features, and updated administrative/geospatial data to reflect new developments while grouping areas of similar land use and function. This rigorous methodology provides a stable yet up-to-date framework for census operations and statistical reporting.

By subdividing municipalities into these small areas, ISTAT and local authorities can efficiently assign work to enumerators and ensure every part of the territory is accounted for. Since 1991 the digitized census sections have been the tool that helped municipalities prepare detailed topographic plans for conducting the general censuses. In this way, the sections guarantee complete coverage and avoid omissions or double counting during data collection. Historically, one enumerator would be responsible for one or a few sections, making the massive census operation manageable at a local scale. Census sections allow ISTAT to publish population, housing, and economic data for very small areas, capturing neighborhood-level variations. They serve as the building blocks for all larger geographic aggregations. Census areas have important administrative and legal roles as they serve both ad-

Figure 2.1: Italian census areas by macro-category. Source: ISTAT.

ministrative and statistical purposes in Italy. A primary goal of the decennial census is to determine the legal population of each municipality, which influences governance, resource allocation, and planning (e.g., hospitals, schools, pharmacies). At a sub-municipal level, census sections provide standardized geographic units used for electoral districts, health zones, and urban planning. These sections help local authorities analyze population trends while ensuring privacy protection by releasing only aggregated data. Overall, ISTAT census areas support data-driven decision-making, administrative efficiency, and confidentiality.

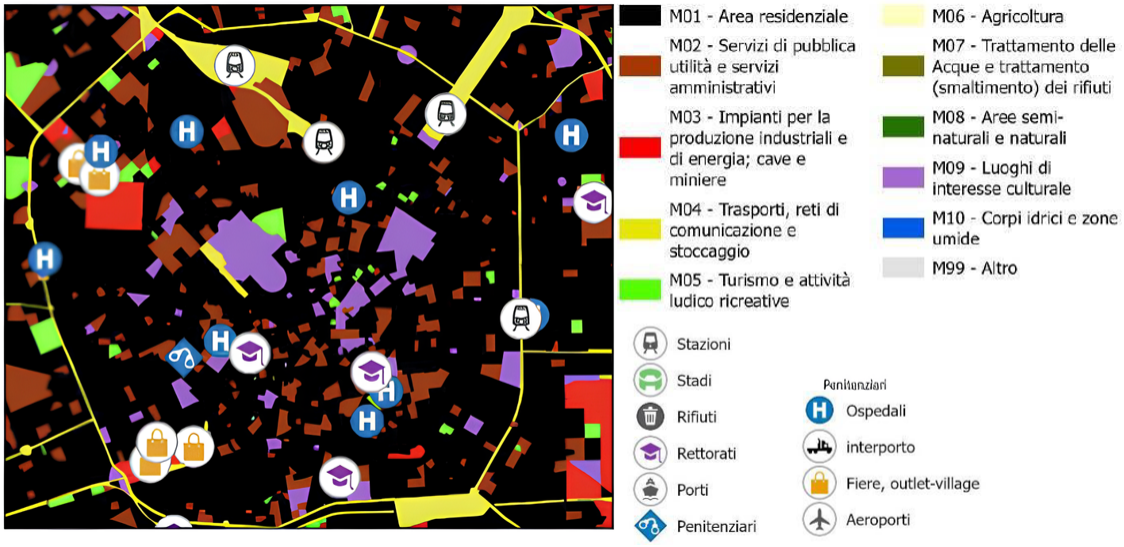


Figure 2.2: Milan census areas by macro-category and points of interest. Source: ISTAT.

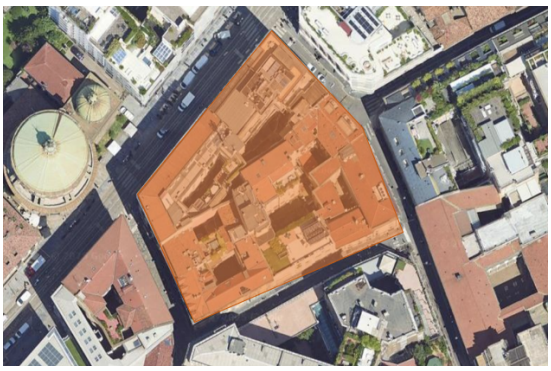
2.2 Problem Statement

ISTAT census areas, or sezioni di censimento, serve as the foundational territorial units for statistical and administrative purposes in Italy. These sections provide a structured approach to population data collection, urban planning, and resource allocation. However, while ISTAT has continuously improved the accuracy of census area boundaries through advances in digital cartography and administrative data integration, the classification of land-use typologies within these areas remains an ongoing challenge due to the infrequency of updates. The official census area classifications, published approximately every ten years, struggle to keep pace with the rapid transformations occurring in urban environments.

Urban development in cities like Milan is highly dynamic, with formerly industrial zones being repurposed for commercial or residential use, new green spaces emerging, and large-scale infrastructure projects reshaping the landscape. Despite these changes, census area classifications often fail to reflect real-world conditions, leading to potential discrepancies between recorded land-use categories and their actual function. Such misclassifications can be observed in Figure 2.3, where an urban park is classified as residential, and

a cluster of civil-use buildings is mistakenly labeled as religious.

(a) Urban park incorrectly labeled as residential buildings



(b) Civil-use buildings misclassified as a religious site

Figure 2.3: Examples of incorrect census type assignments

They have direct consequences, particularly for businesses reliant on accurate geospatial data, such as the real estate sector, where outdated classifications can distort property valuation, market trend analysis, and investment risk assessments.

Traditionally, the process of verifying census area classifications has been manual, requiring analysts to inspect each area individually using Google Maps, aerial photography, and local surveys. This approach is time-consuming and impractical for large metropolitan areas like Milan, where thousands of census areas would require verification. As a result, misclassified census areas persist for extended periods, affecting decision-making processes in urban development, infrastructure planning, and business investments.

Although ISTAT has transitioned to a Censimento Permanente (rolling permanent census) model for population statistics, the classification of census areas has not followed the same modernization process. Italy's census data publication schedule has been historically governed by legal mandates, including [20, 13], which require a full census at least every ten years. Despite advancements in data collection methodologies, census area typologies continue to be published on a long-term cycle, causing classification discrepancies to persist long after urban transformations occur.

To address this issue, this research proposes a two-step validation ap-

proach, as shown in Fig.2.4, that combines geospatial data retrieval with satellite image classification:

1. The first approach leverages Google Places API to validate structured land-use categories such as schools, hospitals, and religious sites by querying real-time geolocation databases. If a census area is labeled as a school zone but lacks an actual educational institution, this discrepancy can be flagged for correction.

2. The second approach applies deep learning models to classify census areas based on satellite imagery, allowing for the identification of broader land-use types such as residential neighborhoods, industrial zones, green spaces, and road networks. By training neural networks to recognize spatial patterns and land-use characteristics, this method enables automated classification updates without requiring exhaustive manual verification. However, Since ISTAT classifications are often outdated or incorrect, the dataset is re-labeled with the assistance of the GPT-4o model [24], which is provided with the ISTAT category definitions and tasked with predicting the most probable classification based on spatial attributes, contextual data, and satellite imagery features.

This AI-assisted validation process ensures that census areas receive a more accurate and updated classification before being used in deep learning model training. When GPT-4o's predicted label matches ISTAT's classification, the label is retained; however, when discrepancies arise, manual validation is performed to confirm whether GPT-4o's prediction is more accurate.

By integrating these two validation strategies, this study aims to provide a scalable and automated framework for continuous census classification updates. The proposed methodology ensures that territorial data remains aligned with real-world urban developments, reducing reliance on infrequent ISTAT updates. The expected outcome is a system that allows for the dynamic verification of census classifications at any point in time, significantly

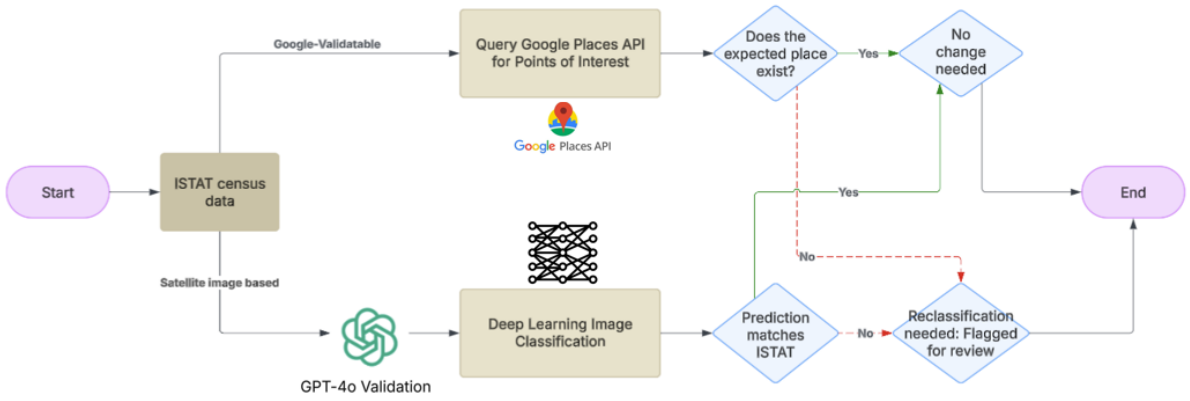


Figure 2.4: Workflow of the proposed approach

improving the accuracy of urban analysis, real estate decision-making, and municipal planning processes.

2.3 Literature Review

2.3.1 GPT-4o and data validation

Within the framework of the validation and updating of land-use classifications at ISTAT, the integration of GPT-4o into the data labeling process is immensely beneficial. GPT-4o is a later iteration of OpenAI's Generative Pre-trained Transformer line of multimodal models that is highly capable of handling and producing text, audio, and image data in real time. This feature renders GPT-4o an essential tool for improving the accuracy and efficiency of data annotation processes. GPT-4o exhibits several key features that make it particularly suited for data labeling and validation:

- **Multimodal Processing**: Unlike its predecessors, GPT-4o can interpret and generate content across multiple modalities, including text, images, and audio. This enables the model to analyze complex datasets that combine various data types, facilitating more comprehensive understanding and annotation.

- **Enhanced Contextual Understanding**:With a very large context window of 128K tokens, GPT-4o can process and retain lengthy strings of information. This allows for improved handling of context, which is necessary for proper labeling, especially in those datasets where context carries classifying power.

- **Rapid Response and Real-Time Processing**:Its design allows for swift processing and generation of outputs, making it suitable for real-time applications and efficient large-scale data annotation projects

Application of GPT-4o for validating and enhancing ISTAT's land-use categories offers a number of advantages that improve the precision, efficiency, and scalability of the classification task. One significant advantage is the potential for automated pre-labeling, where the model applies first-level classifications to data points, greatly lessening the burden for human annotators and speeding up the process of overall annotation. With its advanced contextual understanding and pattern learning, GPT-4o achieves very high consistency and accuracy in annotation, minimizing human-side errors and providing a more reliable dataset. This matters when working with ISTAT classificiations where erroneous and obsolete labels need correction on a systematic basis. With the rapidly changing urban setting, the feature guarantees classification updates and validities can be undertaken dynamically in tandem with real-world changes without needing expensive manual verification. GPT-4o also brings enormous cost-effectiveness to the process through the automation of much of the labeling process. This lowers the reliance on human annotators, bringing enormous cost savings for data annotation tasks on a large scale, without compromising or even enhancing the quality of classification.

The inclusion of GPT-4o in ISTAT land-use class validation and reclassification procedures makes it possible to have a complete, AI-supported methodology that transforms the precision, uniformity, and operational efficiency of the data. The innovation ultimately elevates the quality of geospatial data utilized for city planning, property valuation, and policy-making.

2.3.2 Use of Deep Learning approaches in land use tasks

The use of deep learning techniques in land use and land cover (LULC) classification has greatly advanced in the recent past, making it a strong complement to the use of conventional remote sensing approaches[30]. The growth in the availability of high-resolution satellite imagery from satellites like Landsat and Sentinel has promoted the creation of deep learning-based classification models, and as such, facilitated automatic, large-scale experimentation of urban and environmental changes [35]. Compared to classical approaches such as Support Vector Machines (SVMs)[6] and decision trees, deep learning eliminates the need for extensive manual feature extraction and domain-specific engineering, making it an attractive choice for land-use monitoring. Traditionally, LULC classification has depended on statistical learning algorithms and rule-based models, where classification accuracy is highly dependent on manually engineered features, for example, vegetation indices and spectral signatures[**?**]. While these approaches have yielded strong results, their generalizability to new geographic regions is still limited, and they need manual adjustments and expert knowledge. The transition towards Convolutional Neural Networks (CNNs) [22] developed a comprehensive feature learning model, which greatly enhanced classification accuracy and facilitated automation. CNN-based architectures, including U-Net[28] and SegNet[3], have been extensively used in semantic segmentation to achieve pixel-wise land classification at high resolutions. One of the most significant advancements in deep learning for LULC classification has been the introduction of Vision Transformers (ViTs), which replace traditional convolutional layers with self-attention[31] mechanisms to model long-range dependencies in imagery data[10]. Unlike CNNs, which learn local spatial features, ViTs learn images holistically and are therefore able to learn relationships between distant pixels in an image. A comparative analysis[17] has been presented for the performance of CNN-based models versus Transformer-based models for the classification of high-resolution satellite imagery, which indicated that Swin-Unet yielded the highest accuracy (96%), demonstrating attention-based architec-

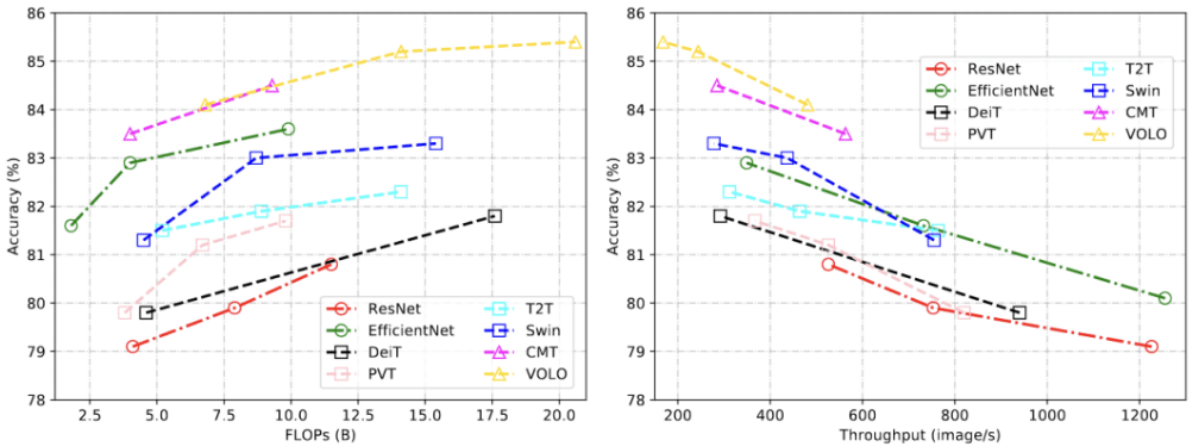


Figure 2.5: CNN vs. ViT: FLOPs and throughput comparison of CNN and Vision Transformer Models. Source: [16]

tures are superior in remote sensing. This architectural difference imparts several advantages to ViTs:

- **Global Contextual Awareness**: ViTs can inherently model global interactions within an image, which is particularly beneficial for LULC tasks where understanding the spatial arrangement and relationships between different land cover types is crucial[7].

- **Flexibility in Input Resolution**: ViTs can process images of varying resolutions without the need for significant architectural adjustments, offering greater flexibility in handling diverse datasets.

- **Robustness to Image Distortions**: Studies have shown that ViTs exhibit robustness to various image perturbations, including adversarial attacks and occlusions, making them reliable for real-world applications [27].

All of these strengths translate, as shown in fig.2.5 in improved performances for ViTs with respect to CNNs.

Despite these advances, deep learning-based LULC classification is not immune to challenges. Label scarcity and inconsistencies in official statistics undermine model performance, and external validation strategies such

as self-supervised learning or transfer learning are needed to enhance generalizability. The second challenge is the computationally intensive nature of Transformer-based models, which demands efficient training strategies such as fine-tuning on pre-trained models. The following section talks about the successful implementation of these models into practical applications, specifically in land-use mapping, census verification, and large-scale geospatial modeling.

2.3.3 Applications of Deep Learning in Land Use Mapping and Census Validation

Application of deep learning algorithms in large-scale land-use mapping has been a game-changer, enabling improved frequency, accuracy, and automatization of operations in classification a reality. Deep learning algorithms have been utilized widely by organizations and studies alike, either improving or substituting conventional methods of classification. A good example is the **OECD**'s urban land-use mapping research, which employed a U-Net-based deep learning model to delineate residential, commercial, and industrial areas from Sentinel-2 satellite imagery in 687 urban agglomerations[4]. The research demonstrated that land-use monitoring could be conducted in near real-time using deep learning-enabled methods, supplementing conventional census processes. One other example is that of the **European Space Agency** (ESA), in which global land cover datasets at a 10m resolution have been developed using deep learning-based segmentation models trained on Sentinel-2 data [1]. Maps have also been used in environmental monitoring, deforestation monitoring, and agricultural land-use classification, providing high-resolution geospatial data. In Italy, researchers at ISTAT have developed a CNN based land cover estimation system, integrating deep learning-based segmentation models applied to Sentinel-2 and EuroSAT datasets[5]. By automating land-cover-classification, this system provides a cost-effective alternative to traditional LC updates, offering a framework for more dynamic land-use classification strategies. Despite these advances, de-

ployment of deep learning models at scale is still difficult, particularly in official census validation. The majority of national datasets, like ISTAT's land-use classes, contain outdated or incorrectly labeled records, and external validation procedures are required. Additional research has tried to improve dataset quality through AI-driven re-labeling approaches, where deep learning models, in conjunction with auxiliary geospatial information, automatically improve labels[29, 11]. To overcome these difficulties, this research incorporates GPT-4o as an AI-based re-labeling mechanism to ensure the validation of ISTAT land-use classifications prior to inputting them into deep learning models. Because ISTAT census classifications are renewed just once every decade, they often do not correspond to actual-world urban change, particularly in the quickly changing metropolitan region of Milan. The utilization of GPT-4o enables a systematic cross-validation of the misclassified census areas against satellite image-based features, geospatial metadata, and Points of Interest (POI) data obtained from the Google Places API. The method enables greater robustness and accuracy of labeled data, allowing for the identification and correction of erroneous ISTAT classifications prior to Vision Transformer model training. The viability of this strategy is reinforced by a body of research on large-scale AI-supported label refinement. The effectiveness of this approach is supported by studies on large-scale AI-driven label refinement. Research comparing GPT-based labeling with traditional human annotation methods found that GPT-4-based classification achieved accuracy scores similar to human performances[33]. After label correction, the enhanced dataset is used for fine-tuning Vision Transformer models for final classification outcomes. Research has proven that Vision Transformers consistently perform better than CNN-based models in geo-spatial classification tasks, especially in cases involving global context awareness. A comparative study, which was published in Remote Sensing [23], compared the performance of ViT-based models with CNNs for vineyard classification and concluded that ViTs produced higher accuracy and F1 scores because they have the capacity to visualize intricate spatial relationships. A study [21] proposes an innovative framework utilizing transformer-based models for

Land Use and Land Cover (LULC) classification using optical satellite imagery. The research aims to balance computational efficiency and accuracy in LULC analysis by employing transfer learning and fine-tuning strategies to optimize resource utilization of transformer-based models. applied transformer models to LULC tasks, highlighting their ability to effectively analyze complex spatial patterns in satellite imagery.

The integration of AI-driven re-labeling with Vision Transformer-based classification presents a modernized, scalable alternative to traditional census validation. This methodology ensures that land-use classifications remain dynamic and reflective of real-world conditions, improving their applicability in urban planning, real estate valuation, and policymaking. As deep learning continues to evolve, the combination of self-attention architectures and AI-enhanced label validation will likely define the next generation of land-use mapping systems, reinforcing their role in data-driven urban development strategies.

Chapter 3

Dataset Construction and Preprocessing

The performance of a machine learning model relies heavily on the characteristics of the input data. For land-use mapping, dataset construction and preprocessing are important to enable the model to capture relevant geospatial and remotely sensed patterns. The effectiveness of deep learning models depends strongly on how well the input dataset reflects the real-world land-use distribution and how uniformly its format is captured across sources.

This chapter details the processes of data acquisition, preprocessing, and validation necessary for the use of the Google Maps Places API and training deep learning models on Milan's census areas. The dataset comprises ISTAT census data and satellite imagery from Google Maps. Since ISTAT's land-use classifications are often outdated or misclassified, an AI-driven relabeling process using GPT-4o was employed to correct inconsistencies before model training. Additionally, various augmentation, balancing, and geospatial transformation techniques were applied to ensure that the dataset meets the necessary conditions for robust deep learning-based classification.

3.1 Census Data

The statistics produced by ISTAT serve as the point of reference for Italy's territorial classification with detailed statistics on geographic, demographic, and land-use conditions. Milan census sections are considered in the research because it suits the business interest of the real estate firm where this research was carried out in the context of an internship program. Milan is a highly dynamic urban area where land-use classes frequently become obsolete, so it is a great case study to try out an AI-driven classification update method.

The ISTAT census dataset is the primary geospatial and statistical source for this research, offering a broad territorial classification of Italy appropriate for administrative as well as analytical purposes. The dataset contains exhaustive details regarding census sections, the smallest territorial divisions utilized for information gathering, and offers the primary attributes concerning land use, population, and geographic structure. The data is presented in the format of a GeoJSON file, covering a total of 39 attributes that represent both non-spatial and spatial features of each census section.

This study focuses on Milan's census sections, aligning with the business interests of the real estate company where this research was conducted as an internship project. The dataset was filtered to retain only census sections within Milan, reducing the dataset to 7,383 census areas, which are displayed in 3.1 . The ISTAT dataset is structured as a georeferenced shapefile containing both spatial boundaries and non-spatial attributes. Below is a breakdown of its most essential components.

To ensure consistency across all geospatial analyses, the dataset uses World Geodetic System 1984[19] (WGS 84, EPSG:4326) as its Coordinate Reference System (CRS). This system is widely used in geospatial applications and is particularly compatible with satellite imagery and mapping APIs such as Google Maps. Each census section is represented as a polygonal geometry, outlining the exact boundaries of the census area. The geometry field contains:

- **Type**:Always "Polygon," indicating that the region is stored as a closed

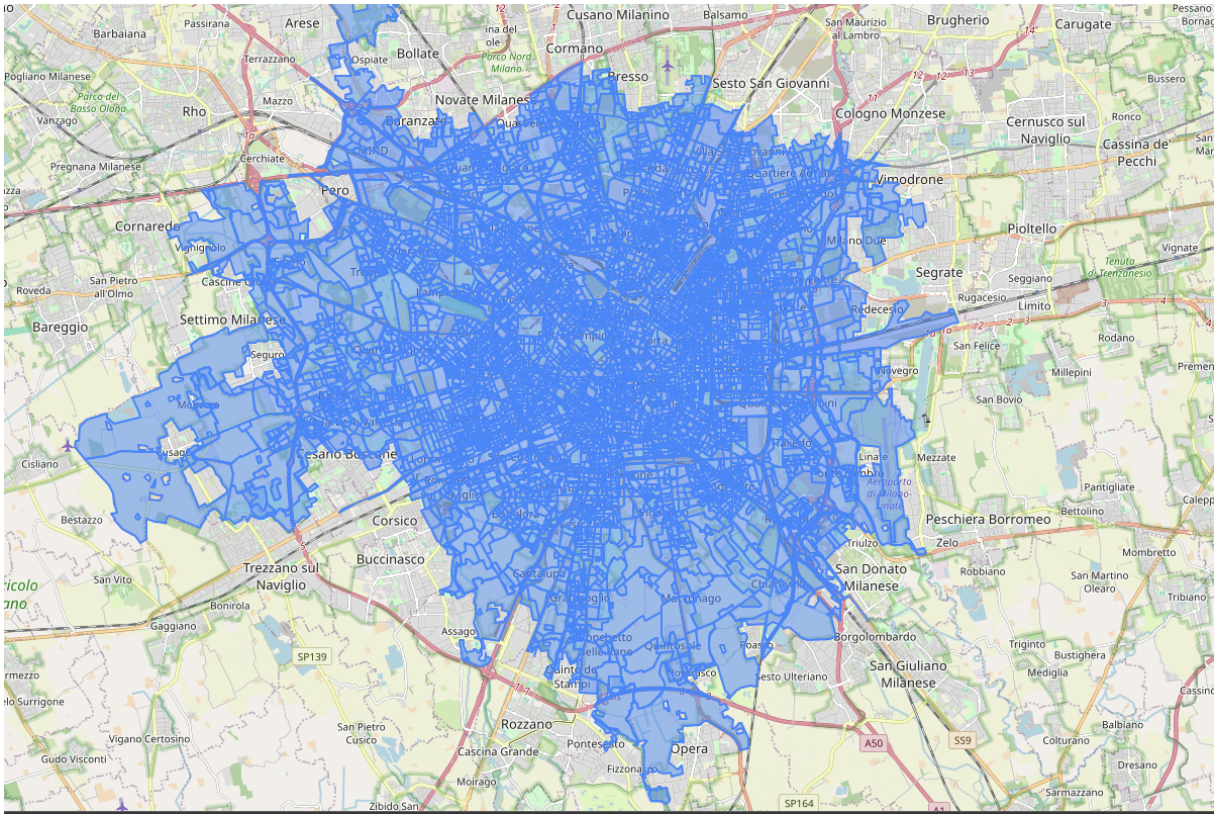


Figure 3.1: Milan census areas

shape

- **Coordinates**: A nested list of latitude-longitude pairs defining the perimeter of the census area.

These polygonal representations are crucial for overlaying census data with satellite images and point-of-interest (POI) data from Google Places API, ensuring spatial alignment in downstream analyses.

The dataset contains 39 attributes, of which the most relevant to this study include geographic identifiers uniquely identify each census section and its location within the national administrative hierarchy:

- **COD_REG**: Regional code
- **COD_UTS**: UTS (Unione Territoriale Sostenibile) code
- **PRO_COM**: Municipality code
- **SEZ21**: Census section identifier for the 2021 census
- **SEZ21_ID**: Unique identifier for the census section

- **COD_RIP**: Higher administrative division code (Ripartizione)
- **COD_PROV**: Province code
- **COD_CM**: Metropolitan city code
- **PRO_COM_T**: Extended municipal code
- **COMUNE**: Name of the municipality (set to `"Milano"` for all records)

Moreover, to account for population density and socioeconomic factors, the dataset also includes:

- **POP21**: Population within the census section in 2021
- **FAM21**: Number of families within the census section in 2021

These attributes provide additional context for understanding land-use distribution, particularly in distinguishing between residential, commercial, and public-service zones

However, the most important variable in our dataset is the target variable **COD_TIPO_S** which represents the prevalent land use of the census area. The original ISTAT classification for Milan contains 37 distinct land-use categories, but for this study, the dataset was filtered to retain only 11 categories, covering 96% of the census sections. The rationale behind this filtering process is as follows:

- **Ensuring Sufficient Training Data**: Many land-use categories had very few examples, making it difficult for deep learning models to generalize effectively. By selecting the most frequent categories, the dataset remains balanced and statistically representative.
- **Prioritizing Major Land-Use Types**: The selected 11 categories correspond to the most critical urban land uses, such as residential areas, commercial zones, transportation infrastructure, and green spaces.
- **Improving Model Generalization**: By focusing on the most prevalent land-use types, the classification model can be trained more effectively while still maintaining the possibility of extending the methodology to the remaining 4% of data in future research.

Table 3.1 provides an overview of the selected land-use categories and their corresponding ISTAT codes.

3.2 Satellite Image data overview and pre-processing

For acquiring high-resolution satellite images of the census zones of Milan, the Google Maps satellite layer was retrieved using the `leafmap` Python library. It is an advanced geospatial tool that allows for the extraction, visualization, and analysis of different map layers, including satellite images sourced from different platforms[32]. By leveraging the Google Maps Satellite imagery, which provides seamless access to high-resolution satellite images, the study ensured comprehensive coverage of the city's administrative divisions. It supplies satellite imagery with a spatial resolution of approximately 30 cm per pixel, calculated in the following way:

$$\text{meters per pixel} = \frac{156543.03392 \times \cos(\text{latitude})}{2^{\text{zoom level}}} \tag{3.1}$$

For Milan (latitude $\approx$ 45.46N), this results in a resolution of approximately 0.212 meters per pixel. Using this offering a high level of detail

Class	**Description**
2	Religious Sites (Church, Mosque, Synagogue)
9	Healthcare Facilities (Hospitals, Clinics, ASL Offices)
18	Educational Institutions (Universities, Research Centers)
37	Schools and Public Services (Libraries, Post Offices)
16	Sports Facilities (Stadiums, Pools, Athletic Fields)
1	Residential Areas
12	Industrial and Production Areas
10	Train Stations and Railway Infrastructure
36	Roads and Highways
26	Agricultural Areas
5	Green Areas and Parks

Table 3.1: Selected Census Area Categories for the work

suitable for differentiating between various land-use categories, such as residential, commercial, industrial, and green spaces [14]. This ensures that the extracted images closely align with the most recent ISTAT census classification, minimizing discrepancies between census data and real-world land use. Satellite image and census data synchronization enhances the credibility of spatial analysis, thereby enabling more precise observation of infrastructural growth and environmental changes throughout Milan[35].

The extraction process was carried out using the built-in function `leafmap.map_tiles_to_geotiff()`, a method in the leafmap library designed to capture tiled satellite imagery from online map sources such as Google Maps, ESRI, OpenStreetMap, Nasa Blue Marble and many others. This function requires the specification of bounding coordinates for each census area, allowing for precise retrieval of image tiles corresponding to the designated regions. As part of the preprocessing pipeline, once the satellite images are retrieved, they undergo a cropping process to match the exact shape of the census areas. The downloaded images are stored as GeoTIFF (Georeferenced Tagged Image File Format), a widely used raster format that embeds spatial metadata within the file. Unlike standard image formats like PNG or JPEG, GeoTIFF files contain coordinate reference system (CRS) information, georeferencing data, and transformation parameters, allowing them to be accurately aligned with other geographic datasets.

This makes GeoTIFF essential for geospatial analysis, as it ensures that

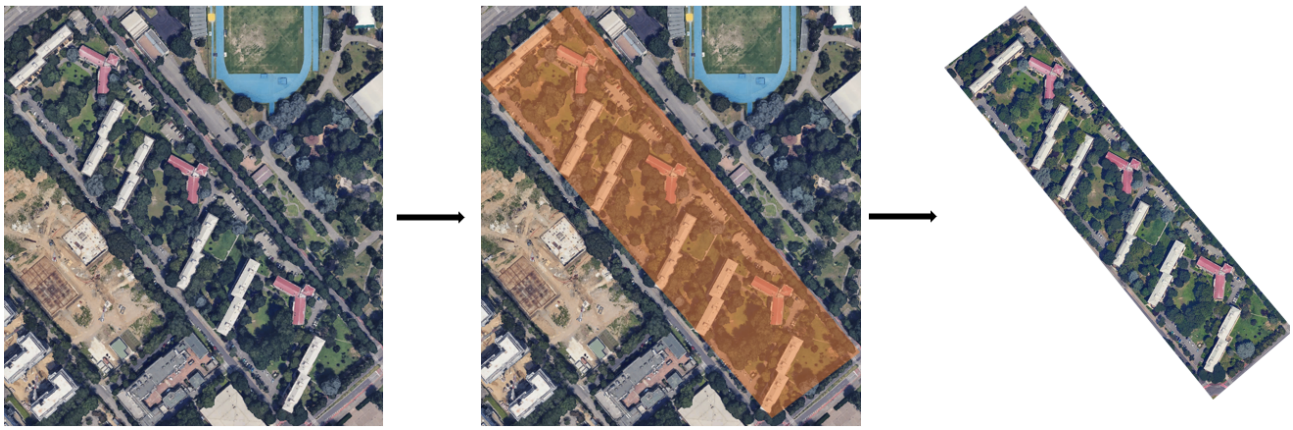


Figure 3.2: Milan census areas image preprocessing

each pixel corresponds to a specific location on Earth. Using the census dataset's geometries, the function applies a masking operation with the `rasterio.mask` function, ensuring that only pixels within the census boundary are retained. This step, which is displayed in Figure 3.2 eliminates extraneous background information outside the official census borders, making the dataset more spatially precise and reducing noise in subsequent analyses. Furthermore, to enhance computational efficiency and storage capacity, the images are first padded to ensure a square aspect ratio and then resized to a standardized 224×224 resolution. Additionally, as part of the preprocessing pipeline, the images undergo normalization, where each pixel value, originally ranging from 0 to 255, is first converted into a tensor representation and then scaled to a range between -1 and 1 using the following transformation:

$$I_{\text{norm}} = \frac{I - \mu}{\sigma}$$

where: Where:

- I is the original image (pixel values).
- μ is the mean of the image or dataset.
- σ is the standard deviation of the image or dataset.

This normalization process is applied independently to each color channel (red, green, and blue) to ensure that the model receives inputs with a consistent pixel distribution. By standardizing the pixel values, normalization reduces variations in brightness, contrast, and dynamic range across different images. Additionally, by centering the data around zero, it improves the stability of gradient descent, leading to faster convergence and better generalization in machine learning tasks such as image classification, object detection, or land-use classification.

Through the use of this structured process—bounding box extraction, accurate masking, intelligent padding, resizing, and normalization—the dataset ensures both high spatial precision and computational efficiency, yielding an

enhanced and fully preprocessed satellite imagery dataset suitable for urban and demographic analysis. The end of the pre-processing step ends with saving the images into PNG format.

3.3 Maps Places API aimed data pre-processing

As already mentioned, the Google Places API, an advanced geospatial tool, was employed in this study to corroborate the land-use classification of census areas as established by ISTAT for some classifications. By leveraging this API, it became possible to engage in a dynamic, scalable, and automated process of cross-checking census-based classifications with empirical observations, and thus improve the accuracy, reliability, and contemporary relevance of urban land-use analysis. Google Maps API [14] is a collection of web services provided by Google to embed geospatial data, mapping functionality, and location-based services into applications. Within its feature-rich sets, it exposes detailed road maps, satellite images, panoramic street-level imagery, real-time traffic conditions, and points of interest (POIs). Within this context, the Google Places API is a specialized service that gives real-time access to geospatial information, including businesses, points of interest, and landmarks across the globe. The API supports a range of functionalities, including:

- **Place Search**: Allows querying for specific places based on location, keyword, category, or ranking.
- **Place Details**: Provides additional metadata for a given place, such as address, contact details, and user-generated reviews.
- **Place Photos**: Returns photos of locations from the Google database.
- **Place Autocomplete**: Suggests locations dynamically as users type in a search query.

- **Geocoding and Reverse Geocoding**: Converts addresses into geographic coordinates and vice versa.

For this study, the `NearbySearch` query function of the Google Places API was employed to validate the accuracy of ISTAT's land-use classifications. This function enables the retrieval of place information based on geographic proximity, making it ideal for systematically verifying whether specific types of locations—such as schools, hospitals, religious sites, and commercial areas—are actually present within a given census zone.

In the ISTAT dataset, every census area is represented by a polygonal boundary. For enabling an efficient API-based search, it was necessary to identify a representative point for every area. Rather than considering the centroid, which is not always contained within the polygon, a `geopandas.representative_point()` point was set as the best available reference point for querying Google Places. The use of a representative point over the centroid of the census area has been chosen because the last is found as the geometric center by computing the mean of its vertex coordinates. The proof is carried out using the formula by:

$$C_x = \frac{\sum_{i=1}^{n} x_i}{n}, \quad C_y = \frac{\sum_{i=1}^{n} y_i}{n}$$

where:

- x_i, y_i are the coordinates of the i^{th} vertex of the polygon,
- n is the total number of vertices

But this method has a severe limitation: In the case of concave or highly irregular polygons, the centroid may lie outside the physical boundaries of the polygon. This is particularly undesirable for non-uniformly shaped areas, disjoint areas, or holes in the polygons as it leads to incorrect placement in spatial queries. To overcome this issue, a representative point is used instead. The representative point is always inside the polygon, ensuring a reliable reference for location-based searches. Its calculation follows an algorithmic

approach rather than a strict mathematical formula. The process can be described as follows:

1. **Compute the Polygon's Centroid**: First, the centroid is calculated using the above standard geometric formula.

2. **Check if the Centroid is Inside the Polygon**: If the centroid lies within the polygon, it is used as the representative point.

3. **If the Centroid is Outside, Find an Interior Point Heuristically**: The algorithm selects a point within the polygon using a method that ensures it is well-positioned inside, often preferring a location near the geometric center while avoiding edges or holes.

4. **Using the Largest Interior Triangle**: A technique known as Delaunay triangulation or Voronoi-based selection can be applied to find a well-centered point inside the polygon. The largest interior triangle from the polygon's Constrained Delaunay Triangulation (CDT) is often a good candidate. The incenter (center of the incircle) of this triangle can serve as the representative point.

By using this approach, the representative point avoids placement errors and ensures that API queries are performed from a valid location within the census area. Using the representative point instead of the centroid enhances accuracy and reliability in spatial analysis because:

- It guarantees that the reference point remains within the actual census area.

- It prevents incorrect placements in concave or irregular regions.

- It ensures consistency when querying geographic services like Google Places, where an external centroid could lead to misrepresentations or errors in POI retrieval.

Thus, the representative point is the preferred choice for defining a meaningful central location in geographic data processing.

As mentioned previously, it was also required to establish a dynamic search radius for each polygonal zone with the aim of restricting the search space without jeopardizing the portrayal of each census zone. This radius was set in a manner that would achieve a balance between ensuring full coverage of the census area, minimizing unnecessary overlap with adjacent regions, adapting dynamically to the varying sizes and shapes of different census areas, all trying to minimize the costs of the API calls.

The search radius was determined based on the maximum diagonal length of the census area polygon. The diagonal represents the longest possible straight-line distance between any two vertices of the polygon.

The maximum diagonal length d_{max} was computed using the Euclidean distance formula:

$$d_{\max} = \max\left(\sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}\right), \quad \forall i, j \in \text{vertices}$$

where:

- (x_i, y_i) and (x_j, y_j) are the coordinates of any two vertices within the polygon.
- $d_{\max}$ is the longest possible straight-line distance within the census area.

This step ensures that the largest spatial extent of the polygon is considered when determining the search area.

Once the maximum diagonal length d_{max} was determined, the query radius was defined as:

$$r_{\text{query}} = \frac{d_{max}}{2}$$

This formulation sets the search radius as half the longest diagonal of the census area polygon. This decision was guided by the following considerations:

- **Coverage Efficiency**: A radius of $d_{max}/2$ ensures that a centrally placed search query fully encompasses the polygon, minimizing gaps in coverage. This approach helps to reduce redundant queries, preventing unnecessary overlap with adjacent census areas.

- **Dynamic Adaptation**: Larger census areas automatically receive a wider search coverage, ensuring proportional query expansion. Smaller census areas retain localized queries, preventing unnecessary inclusion of results from neighboring regions.

The adaptive radius approach, based on half the maximum diagonal length, provides an efficient and scalable method for defining search queries in the Google Places API. The land-use categories that will be validated through the Google Maps places API are listed in 3.2 alongside the frequency of each category.

Table 3.2: Google Places API Query Keywords

Class	**Frequency**	**Search Keywords**
2	199	Church, Mosque, Synagogue, Oratory
9	54	Hospital, Clinic, Assisted Living, ASL Offices
18	30	University, Research Institute, Library
37	341	School, Post Office, Library
16	118	Arena, Sports Complex, Stadium

3.4 Automatic Labeling of Satellite Images Using AI

The assurance of precision and consistency of labeled data is a key phase in the deep learning model training process intended for land-use classification. Although the ISTAT census dataset contains an official land-use class classification, existing literature has established that such labels tend to be afflicted with obsolescence or misclassification because of the lengthy update

cycle associated with census data. In the present research, GPT-4o was utilized to confirm and enhance census labels prior to training the deep learning model so that the training data is as precise and recent as possible.

ISTAT's land-use classification system is updated approximately every ten years, meaning that many census areas in Milan may have undergone significant transformations that are not reflected in the official dataset. To address these discrepancies, an automated labeling approach was implemented using GPT-4o, which was tasked with predicting the most probable land-use classification for each census area based on satellite imagery and auxiliary metadata.

GPT-4o was used to revalidate land-use classifications in a systematic pipeline, which involved the following steps:

1. **Extracting Land-Use Context from ISTAT Descriptions**: ISTAT official category definitions were provided as input to GPT-4o in order for the AI model to operate within the ISTAT classification system's boundaries. This allowed the model to not generate ambiguous or out-of-scope classifications.

2. **Generating AI-Predicted Land-Use Labels**: GPT-4o was prompted with:

    - The satellite image corresponding to a census section
    - A pdf document containing the explanation assigned by ISTAT itself to each land-use category.

3. **Comparing GPT-4o Predictions with ISTAT Labels**: If the GPT-4o prediction matched the ISTAT classification, the label was considered valid and retained for training. However, if a discrepancy was detected, the prediction was flagged for manual verification.

4. **Manual Verification of Discrepancies**: The primary focus of this verification process was on mismatched classifications that could not be validated through direct querying of external databases such as Google

Places API. Certain land-use types, such as hospitals, schools, or train stations, can be verified through POI-based methods since they are well-documented in online databases. However, other categories, such as residential, industrial, or agricultural areas, lack directly searchable markers, making them unsuitable for API-based validation. Instead, these labels require assessment through neural models trained on satellite imagery to confirm their correctness.

The results of the GPT4-o label validation is shown in table 3.3 outline the validation process for ISTAT land-use labels using GPT-4o, including an analysis of discrepancies and manual verification outcomes. The key observations from the results are as follows:

GPT-4o's predictions aligned well with ISTAT classifications for most census sections. The largest agreement was observed in Residential Areas (Class 1), where GPT-4o correctly matched ISTAT's labels in 81% of cases. Roads & Highways (Class 36) also showed a strong match, with 93% accuracy. The highest mismatch rates were found in agricultural areas (Class 26) and industrial areas (Class 12), with 42.6% and 42.2% mismatches, respectively. Since these categories do not have directly searchable markers, they required further manual verification or deep learning-based classification. For census sections where GPT-4o's classification differed from ISTAT's but could not be verified using API data, manual validation was conducted. Out of the 303 non-verifiable mismatches in Residential Areas, 30($\sim$10%) were confirmed as correct by GPT-4o. These findings highlight the necessity of integrating AI-assisted label refinement techniques into land-use classification. By combining GPT-4o with API verification and manual checks, this approach provides a high-confidence dataset, improving the accuracy of machine learning models trained for automated land-use classification. Hence, the dataset used in this study consists of ISTAT census sections that were validated through GPT-4o label verification and further refined via manual validation. This ensures that the land-use classifications used for training the deep learning model are as accurate and up-to-date as possible, reducing

the risk of learning from outdated or incorrect ISTAT labels.

Class	API	Total	Mismatches	No API Validation	Actual
1	No	5,533	1,057 (19.1%)	303	30
5	No	151	45 (29.8%)	14	7
10	No	117	33 (28.2%)	28	-
12	No	223	94 (42.2%)	24	11
26	No	141	60 (42.6%)	10	2
36	No	210	14 (6.7%)	3	-

Table 3.3: Validation of ISTAT Census Labels Using GPT-4o

The final dataset consists of census sections where GPT-4o's predictions matched ISTAT's classifications, plus the 30 manually validated census areas where GPT-4o corrected ISTAT's misclassifications. This refined dataset represents a high-confidence ground truth, forming the basis for training deep learning models for land-use classification. To ensure that the model generalizes well to unseen data, the dataset was divided into training(80%) and test(20%) sets using a stratified sampling approach.

Looking at the total number of census areas used for model training and testing after label validation in table 3.4, the dataset exhibits class imbalance, where residential land-use type is significantly overrepresented compared to the others. This imbalance poses a risk to model training, as deep learning algorithms tend to favor dominant classes, leading to biased predictions where the model may disproportionately classify new samples as the most frequent category. To mitigate the effects of class imbalance, several strategies were carefully selected and implemented at different stages of the dataset preparation and model training process. These strategies aim to improve model generalization, ensure fair representation of all land-use types, and prevent bias toward majority classes.

To ensure that the model performs well across all land-use categories, the following strategies were employed on a scale that will be discussed in the following chapter:

Class	Training Samples	Testing Samples
Residential Areas (1)	3,302	901
Green Areas (5)	81	23
Train Stations (10)	50	17
Industrial Areas (12)	101	28
Agricultural Areas (26)	64	17
Roads and Highways (36)	203	40

Table 3.4: Dataset Sample Distribution

- **Oversampling of Minority Classes**: Oversampling involves artificially increasing the frequency of minority classes by copying or synthetically generating new samples. It ensures that the model receives increased exposure to rare classes so it does not disregard them during training. Unbalanced classes were oversampled by duplicating existing images and applying data augmentation techniques. Augmented transformations included random rotations, flips, brightness variations, and affine transformations to ensure that oversampled instances were not exact duplicates of existing samples. By oversampling, the proportion of underrepresented land-use categories was increased, ensuring that the model learns to distinguish them effectively.

- **Undersampling of Majority Classes**: Undersampling involves reducing the number of samples from the dominant class (residential areas) to create a more balanced dataset. Instead of duplicating minority-class samples, this approach removes a subset of highly redundant examples from the dominant class. A random subset of residential area images was removed to ensure that its proportion did not exceed a certain threshold of the dataset after balancing. This approach helped prevent the model from being biased towards residential classifications while maintaining a sufficient number of samples for effective feature learning. Undersampling was particularly useful in reducing redundancy in the training dataset, ensuring that a diverse set of residential images remained for model training.

- **Weighted Cross-Entropy Loss**: Even after applying oversampling and undersampling, some degree of class imbalance remained. To further address this, a Weighted Cross-Entropy Loss function, which will be discussed in the following chapter, was implemented during model training.

  Overall, the dataset construction and preprocessing workflow presented in this chapter lays a solid foundation for the subsequent deep learning classification models. By integrating AI-assisted labeling, geospatial validation, and preprocessing techniques, this study ensures that the dataset is not only statistically representative but also dynamically adaptable to the evolving nature of urban land use. The following chapter will explore how this dataset is utilized in deep learning model training, optimization, and performance evaluation, leveraging state-of-the-art neural architectures for automated land-use classification.

Chapter 4

Strategy, Models Development and Results

This chapter presents the methodologies and results of two complementary approaches employed for the classification and validation of ISTAT census sections based on their land-use categories. The first approach that will be disucssed is the Google Maps Places API as an external validation tool to cross-check ISTAT's census classifications. Later, the deep neural network models that have been fine-tuned for the classification task will be discussed.

4.1 API Census Validation

To evaluate the effectiveness of the Google Maps Places API in confirming ISTAT's land-use classifications, a dataset of census areas was queried using predefined search terms (visible in Table4.1) corresponding to each land-use category.

The validation process classified census areas into two groups. Census sections where Google Places API returned at least one relevant POI matching the ISTAT-designated land-use category, thereby confirming the official classification and census sections where Google Places API did not return a corresponding POI for the assigned ISTAT category, indicating a possible classification mismatch or missing data in Google's geospatial records.

Category	Place Types
2	Church, Mosque, Synagogue, Oratorio, Convento
9	Hospital, Nursing home, Clinic, ASST, ATS
18	University, Research Institute, University Campus, Library
37	Preschool, Primary school, Secondary school, School, Library, Post office
16	Arena, Athletic field, Sports activity location, Stadium, Sports club, Sports complex, Swimming pool

Table 4.1: Categories and Corresponding Maps Places API keywords

The quantitative results for each validated category, presented in Table 4.2, show that schools and public services exhibited the highest validation rate (99.7%), with only one census section failing validation. This is likely due to the high density and well-documented nature of educational institutions and public offices in Google Places data. All unvalidated points were manually checked, confirming that the areas did not actually contain the institutions or facilities listed by ISTAT. This suggests that the discrepancies were due to outdated or incorrect census records rather than limitations in Google Places data. Therefore, we can trust this validation method as a reliable approach.

Hence, given the thorough manual verification of unvalidated points, we can conclude that this validation method is robust and trustworthy

Class	**POI Validated**	**POI Not Validated**	**Validation Rate (%)**
2	166	33	83.4%
9	46	7	86.8%
16	111	7	94.0%
18	26	4	86.7%
37	340	1	99.7%

Table 4.2: POI Validation Statistics

4.2 Used Architectures

The section explores two deep learning architectures used for this work:

- **Convolutional Neural Networks (CNNs)**: A widely used approach in computer vision that extracts local spatial features from images
- **Vision Transformers (ViTs)**: A more recent advancement in deep learning that models global spatial relationships through self-attention mechanisms.

By evaluating these two architectures, the study aims to determine which approach is more effective in modeling land-use characteristics from satellite imagery.

The method described in this chapter involves defining the model architectures, selecting the appropriate hyperparameters, and employing training techniques that trade off between performance and prevention of overfitting. Evaluation is done on the basis of accuracy, precision, recall, and F1-score, but with a specific focus on evaluating how well each of the models generalizes to the different classes of land-use. In addition, misclassification case studies are conducted to ascertain typical errors and improvement opportunities for the classification workflow.

The model architecture is one of the essential considerations in the success of land-use classification from satellite imagery. Deep models must be able to identify spatial patterns, tell apart lookalike land-use types, and generalize across various geographic locations. With these considerations in mind, the current study uses two different architectures: Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs). CNNs are the baseline model, while ViTs are a more recent option that uses self-attention mechanisms to capture spatial relationships. Neither model is trained from scratch but is fine-tuned on the dataset, benefiting from transfer learning in adapting pre-trained representations to the specific land-use classification problem.

Convolutional Neural Networks (CNNs) have emerged as the most popular approach in deep learning for image classification and remote sensing applications. Their capacity to learn spatial hierarchies, identify features like edges, textures, and shapes, and identify patterns makes them extremely well-suited for land-use classification applications. CNNs have been extensively applied in remote sensing because of their capability to effectively process high-resolution satellite imagery and identify localized spatial features that are paramount in differentiating between various land-use categories.

Within land-use classification of a census section, CNNs preserve high-resolution spatial details, i.e., road textures, vegetation structures, and man-made structures, allowing them to discover representative patterns for different land-use classes. CNNs do have their limitations, however, in terms of modeling long-distance dependencies and global spatial relationships that might be dominant in land-use classification when context information exceeds the local spatial receptive field of convolutional kernels.

For this study, ResNet-50[18], a widely used deep CNN architecture, was selected as the baseline model. Specifically, the Microsoft ResNet-50 model with 25M parameters from Hugging Face was employed, leveraging a pre-trained version to facilitate transfer learning (see Figure 4.1for the model structure).

The key reasons for choosing ResNet-50 are:

- Proven Performance in Image Classification: ResNet-50 has demonstrated strong performance across various image recognition tasks, including remote sensing and aerial imagery classification.

- Efficient Feature Extraction: The residual learning framework introduced in ResNet models helps mitigate the vanishing gradient problem, allowing for deeper networks that can capture complex spatial patterns.

- Computational Efficiency: Compared to deeper variants (e.g., ResNet-101), ResNet-50 provides a balance between model complexity and computational efficiency

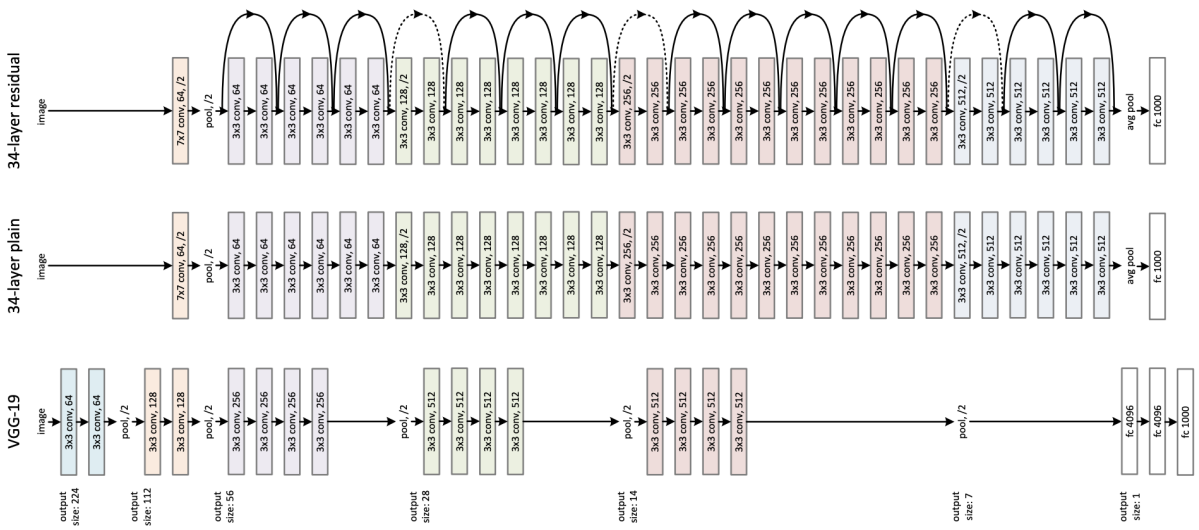


Figure 4.1: ResNet-50 architecture. Source [18]

Instead of training ResNet-50 from scratch, the model was fine-tuned our satellite data, leveraging the pre-trained feature representations while adapting the model to the specific characteristics of census section imagery. By leveraging ResNet-50 as a baseline, the study establishes a strong reference point for evaluating the performance improvements introduced by Vision Transformers.

While CNNs have been the dominant architecture for image classification tasks, recent advances in deep learning have introduced Vision Transformers (ViTs) as a promising alternative. Unlike CNNs, which use local convolutional filters, ViTs divide an image into fixed-size patches and process them using self-attention mechanisms, allowing them to capture long-range dependencies between different parts of the image. In particular, `ViT-Base-Patch16-224` with 86M parameters has been chosen. (see Fig.4.2 for the model structure).

The key advantage of ViTs for land-use classification is that they can see global spatial relationships. Land-use patterns are typically wider than local pixel structures, and the model must be able to learn contextual information at a larger scale. For example, discriminating between industrial and residential regions may require looking at wider city structures, which ViTs do

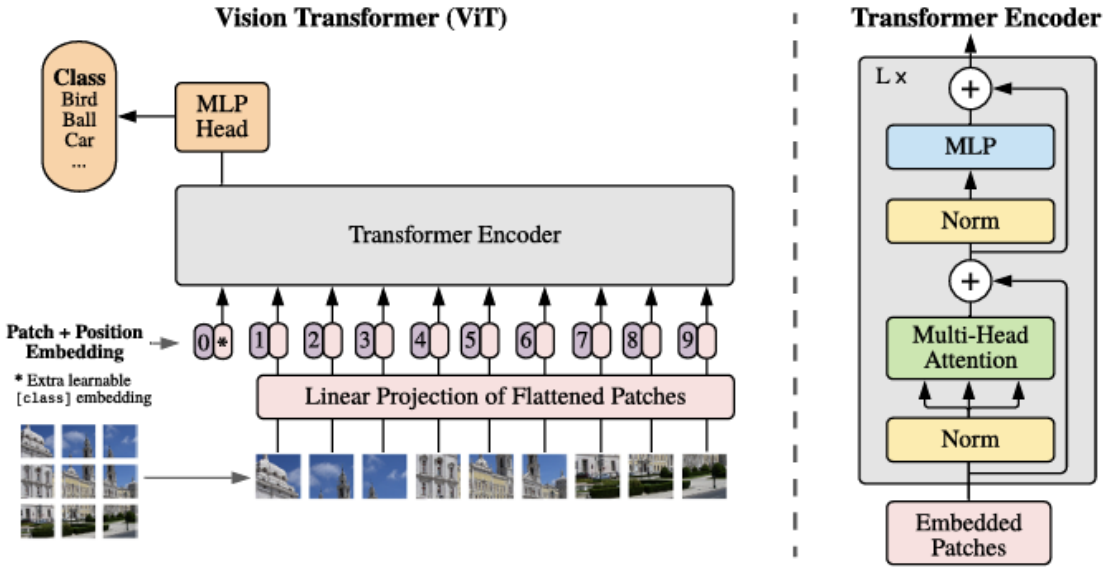


Figure 4.2: Vision Transformer model architecture: source [10]

more effectively than CNNs.

The key reasons for choosing Google's `ViT-Base-Patch16-224` are:

- Efficient Patch-Based Processing: The model divides an input image into 16×16 pixel patches, reducing computational complexity while preserving global spatial information.
- Pretrained Weights on Large-Scale Datasets: The model has been pretrained on ImageNet-21k, allowing for efficient transfer learning and faster convergence when fine-tuning for land-use classification.
- Superior Performance in Remote Sensing Tasks: Studies have shown that ViTs outperform CNNs in various geospatial applications, particularly in tasks that require understanding large-scale spatial structures.

Rather than being trained from scratch, both the ResNet-50 and the ViT models were were fine-tuned on two main tasks:

1. **Binary classification**: an initial approach to the problem in which residential areas (1) correspond to one class and all the others (5, 10,

12, 26, 36) are merged into a single category. A correctly classifying model could already be useful for Generali Real Estate as their core works are mainly focused on residential areas. Thus, being able to correctly classify residential areas all over the others could represent an already important checkpoint.

2. **Classification of the six classes**: represents the most challenging task of this work. Being able to correctly classify all categories would represent the success of the scope of this work.

Fine-tuning involved replacing the final classification head with a task-specific layer for the 2 or 6 land-use categories, optimizing the transformer layers while keeping some pretrained attention mechanisms frozen in the early stages of training. As already mentioned, data augmentation was performed for all presented fine-tuned models to improve generalization, ensuring the model learns spatial variations in census sections. Additionally, augmentation was applied in a way that ensured none of the unbalanced class frequencies exceeded a ratio of 1:7 with respect to the most frequent class. Fine-tuning these models allows to retain general feature representations learned from large-scale datasets while adapting to the specific characteristics of census section images.

While both ResNet-50 and ViT-Base-Patch16-224 are designed for image classification, their architectural differences lead to distinct advantages and challenges. As showed by table 4.3, one of the primary advantages of ResNet-50 is its computational efficiency and ability to perform well with limited training data. CNNs leverage spatial priors—the assumption that nearby pixels are more related than distant ones—allowing them to learn effectively from relatively small datasets. ViTs, however, lack these inherent priors and must learn spatial relationships from data, making them more data-intensive and computationally demanding. As a result, ViTs require large-scale pre-training (such as ImageNet-21k) to perform optimally, whereas CNNs can generalize effectively even with smaller datasets. Scalability-wise, ViTs have an advantage when dealing with satellite images. Their patch-based pro-

cessing structure allows them to maintain fine details in processing global structures and are thus better suited for large-scale geospatial processing. CNNs, on the other hand, require larger and larger convolutional filters to capture broader spatial patterns, which can become computationally costly.

Feature	**ResNet-50 (CNN)**	**ViT-Base-Patch16-224 (Transformer)**
Local vs. Global Features	Extracts local features through convolutional filters	Captures global relationships using self-attention
Spatial Awareness	Primarily focuses on nearby pixels	Can model long-range dependencies across images
Training Complexity	More efficient and requires less data	Requires more data but generalizes better on large-scale tasks
Pretraining Influence	Pretrained CNNs provide robust feature maps for satellite imagery	ViTs benefit greatly from large-scale pretraining, improving performance on complex patterns

Table 4.3: Comparison of ResNet-50 (CNN) and ViT-Base-Patch16-224 (Transformer)

The comparison of ViTs and CNNs for land-use classification brings to the forefront the compromises between effectiveness, spatial comprehension, and scalability. CNNs continue to be an effective baseline model owing to their ability to extract features efficiently and their eased training regimes, whereas ViTs are a newer, more potent model that is ideally suited to capture intricate spatial patterns at the global scale. The subsequent sections will review the training methods and evaluation metrics more closely in order to establish which model is more appropriate for the classification of Milan's census sections from satellite images.

4.3 Training Methodology

The land-use classification training protocol using CNNs (ResNet-50) and Vision Transformers (ViT-Base-Patch16-224) was developed to maximize model performance while addressing main concerns of class imbalance, limited labeled data, and computational constraints. The ISTAT-GPT4o validated dataset was used as the foundation for training, utilizing fine-tuning to adapt pre-trained models to the land-use classification task. This chapter outlines

the dataset splitting, input preparation, training configuration, optimization methods, and regularization techniques applied to achieve model generalization and performance.

The dataset was split into training, validation, and test sets to ensure robust model evaluation. The initial split consisted of 80% training data and 20% testing data, with 50% of the test set aside as validation set. This division ensures that the model is tested on unseen data, reducing the risk of overfitting while providing a reliable measure of generalization performance.

The dataset was loaded into the model using PyTorch[25]'s `DataLoader` to ensure efficient processing. A `Datasets AugmentedBalancedDataset` class was implemented to:

1. Load images from both majority and minority classes in a balanced manner.

2. Apply transformations on-the-fly to prevent overfitting.

3. Use batch-wise parallel processing with optimized data loading parameters.

For efficient training, data loading was optimized with parallel workers and pinned memory, reducing CPU overhead. The final data loaders were configured with a batch size of 64 for both training and validation to balance computational efficiency and gradient stability.

Given the class imbalance, **weighted cross-entropy loss** was used as the primary loss function. This approach modifies the standard cross-entropy loss by assigning higher weights to underrepresented classes, ensuring that the model does not favor majority classes disproportionately. The weighted cross-entropy loss is defined as:

$$L = -\sum_{i=1}^{N} w_{y_i} \cdot \log p(y_i)$$

where:

- N is the total number of samples,
- y_i is the true class label for the i-th sample,
- $p(y_i)$ is the predicted probability for the true class y_i,
- w_{y_i} is the weight assigned to class y_i.

The class weights were computed based on the inverse class frequency to counteract the imbalance, using the formula:

$$w_c = \frac{1}{f_c}$$

where:

- f_c is the frequency of class c in the dataset.

By applying these weights, the loss function emphasizes the contribution of underrepresented classes, thereby improving model performance across all categories.

The Adam optimizer was selected due to its adaptive learning rate adjustment and efficiency in handling sparse gradients, which are often encountered in satellite imagery classification tasks. Adam is particularly beneficial for fine-tuning deep pretrained models, as it:

- Combines momentum (like SGD) with adaptive learning rates, leading to stable convergence.
- Reduces the need for extensive hyperparameter tuning, making it effective for limited datasets.
- Performs well even in non-convex loss landscapes, common in deep CNNs and ViTs.

The models were trained for 30 epochs, a number chosen to allow sufficient learning without overfitting to training data. To further prevent overfitting, early stopping with a patience of 3 epochs was applied. This means that if

the validation loss did not improve for 3 consecutive epochs, training was halted. This strategy has been chosen because it helps prevent unnecessary computations while maintaining optimal generalization.

The training process was conducted using Google Colab[15]'s NVIDIA T4 GPU, which offers a practical balance between computational power and accessibility. This platform was chosen due to its ease of use, cloud-based infrastructure, and ability to accelerate deep learning workloads without requiring dedicated high-performance hardware. Given that Vision Transformers (ViTs) are inherently more computationally demanding than traditional Convolutional Neural Networks (CNNs), leveraging Google Colab's GPU resources significantly improved training efficiency. Additionally, the availability of hardware acceleration, including support for Tensor Processing Units (TPUs) and GPUs, facilitated faster model convergence while mitigating the computational overhead associated with transformer-based architectures.

Since the dataset was relatively small for training a deep neural network from scratch, only the classification head was fine-tuned, while the lower layers were kept frozen. Training all layers from scratch would require significantly more data with respect to the available. Moreover, early layers in both CNNs and ViTs learn generic spatial features[34, 2] (edges, textures), which are useful across different datasets. Fine-tuning only the top classification layers allows the model to adapt its final decision-making process while leveraging pretrained feature extractors. In particular:

- The original fully connected (FC) layer was replaced with a new classification head, while the lower convolutional layers remained frozen. Only the final classification layer was fine-tuned to prevent overfitting while leveraging the pretrained feature representations. The total number of trainable parameters is 4098.

- The transformer encoder layers were kept frozen, and only the multilayer perceptron (MLP) head was retrained to adapt the model to land-use classification. This approach retained the self-attention mechanisms learned during pretraining while refining the final decision-

making layer. The total number of trainable parameters is 4,614.

Regularization was essential to prevent overfitting, especially given the dataset's size. Instead of manually implementing additional techniques, this study leveraged built-in regularization mechanisms within the pretrained models:

- **Dropout Layers**: ViTs and CNNs inherently include dropout layers in their architecture, which randomly deactivate neurons to prevent overfitting.
- **Weight Decay**: The optimizer used L2 regularization[26], which penalizes large weights by adding a squared L2 norm term to the loss function, to encourage simpler models and reduce excessive complexity.

These built-in mechanisms were sufficient for ensuring model generalization without requiring additional explicit regularization techniques.

4.4 Models and Metrics Evaluation

It is critical to assess the performance of deep learning models in land-use classification to determine their reliability and suitability for application in real-world scenarios. Given that this study comprises both binary classification (residential and non-residential areas) and multi-class classification (six land-use categories), there is a need to utilize a variety of evaluation metrics that determine overall performance, per-class performance, and the model's generalizability to unseen census sections.

To assess the effectiveness of the trained models, the following metrics were calculated:

- **Accuracy**: Accuracy estimates the number of correctly predicted census blocks as a fraction of total test instances. While accuracy is a simple metric, it cannot be relied on in class imbalance scenarios because

the model might be able to predict the large categories more accurately but perform appallingly for minority classes. Thus, accuracy alone is insufficient, and others must be employed.

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

- **Precision**:evaluates how many of the predicted positive instances are actually correct. A high precision score indicates that the model makes few false-positive classifications.

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

- **Recall**: measures how well the model identifies instances of a specific class. A high recall score indicates that the model successfully detects most occurrences of a given category.

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

- **F1-score**: is the harmonic mean of precision and recall, balancing both metrics in cases where class distribution is skewed

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

These metrics were computed per class to provide insight into which land-use categories were most challenging for the model

For a deeper understanding of misclassification patterns, a confusion matrix was plotted for binary and multi-class classification. The confusion matrix is a plot of the model predictions versus true labels, showing where the model is consistently making mistakes.

In order to compare the performance of ResNet-50 (CNN) and ViT-Base-Patch16-224 (Vision Transformer) in the classification of ISTAT census sections, a comprehensive comparison of performance was undertaken. Compar-

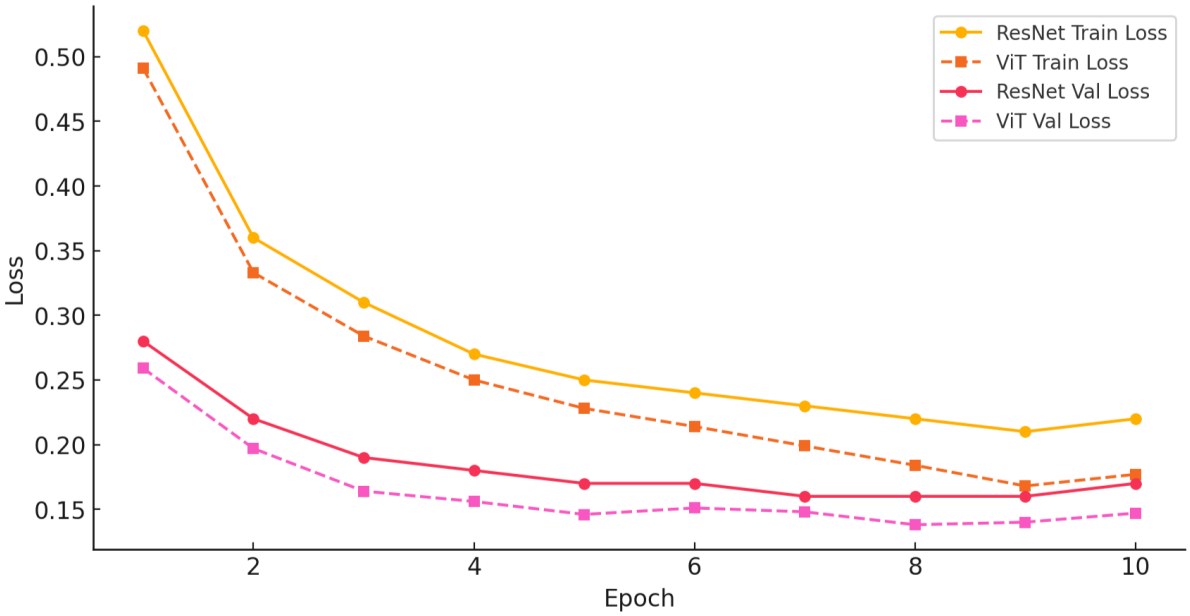


Figure 4.3: Fine-tuning for binary land-use classification: ResNET50 vs ViT

ison was done in four major aspects: overall classification accuracy, per-class precision/recall/F1-score, misclassification patterns using the confusion matrix, and computational efficiency in terms of inference speed. Due to the class imbalance issue and intricate spatial relations, this comparison tries to ascertain if CNNs are still a good baseline or if ViTs offer considerable gains in land-use classification.

4.4.1 Binary Classification Task

The binary classification task was designed to separate residential census sections (Class 1) from all other land-use types combined (Classes 5, 10, 12, 26, 36). Given its practical applications in real estate analysis, a reliable model for this task is already valuable.

The training and validation loss curves for both models, shown in Figure4.3, indicate the learning behavior over 10 epochs. It can bee seen that:

- ViT converges faster than ResNet-50, achieving lower validation loss in fewer epochs, indicating better optimization.
- ResNet-50 exhibits slightly higher validation loss, suggesting that its

feature extraction approach, which relies on convolutional operations, may not be as effective for distinguishing between residential and non-residential areas as ViT's self-attention mechanism.

- The validation loss for both models stabilizes towards the later epochs, but ViT maintains a lower overall loss, demonstrating better generalization to unseen data.

These findings suggest that ViT captures global spatial dependencies more effectively than CNNs, which primarily focus on localized patterns.

The confusion matrices for both models offer deeper insights into misclassifications trends and their implications. By looking at Table 4.4 it is possible to deduce that ViT significantly reduces false positives and false negatives compared to ResNet-50 which misclassified 38 residential sections as non-residential, suggesting difficulty in distinguishing residential structures from mixed-use or industrial areas. Overall, ViT model misclassified fewer instances , demonstrating better contextual differentiation between urban structures and residential areas.

	ResNet-50		**ViT**	
True Class \ Predicted Class	**0**	**1**	**0**	**1**
0 (Non-Residential)	56	11	60	7
1 (Residential)	38	407	19	426

Table 4.4: Comparison of Confusion Matrices for ResNet-50 and ViT for the binary classification task.

The evaluation metrics shown in Table4.5 further highlight the quantitative advantages of ViT in binary classification. ViT has nearly achieved a 5% improvement in accuracy over ResNet-50, demonstrating its effectiveness in binary classification. Higher recall for ViT suggests that it better detects residential areas, reducing false negatives, which is crucial for applications in real estate analytics. Higher precision means fewer false positives, leading to more reliable predictions. These results suggest that ViT is the superior

model for binary land-use classification, making it a strong candidate for the more complex multi-class classification task.

Metric	ResNet-50	ViT-Base-Patch16-224
Precision	0.9242	0.9625
Recall	0.9043	0.9573
F1-Score	0.9108	0.9590
Accuracy	0.9043	0.9573

Table 4.5: Comparison of Performance Metrics for ResNet-50 and ViT-Base-Patch16-224.

4.4.2 Six Categories Classification Task

The multi-class classification task required the models to delineate six land-use classes, adding complexity to the binary classification task. The classification task posed further challenges due to inter-class similarity among certain land-use classes and comparatively fewer samples for certain classes. Owing to these challenges, the ResNet-50 and ViT performance was evaluated based on various metrics such as accuracy, precision, recall, and F1-score, along with confusion matrix analysis.

As seen in Figure 4.4 the training and validation loss curves indicate that ViT consistently achieves lower validation loss than ResNet-50. This suggests superior generalization across land-use categories. ResNet-50 exhibited a smaller gap between training and validation loss, which may indicate a lower tendency to overfit but also suggests limitations in classification capacity for this more complex problem. ViT converged faster, reaching lower loss values in fewer epochs, reinforcing previous findings that transformer-based models process spatial dependencies more effectively than CNNs in land-use classification tasks.

These observations support the hypothesis that ViTs can leverage long-range dependencies in satellite imagery more effectively than CNNs, which

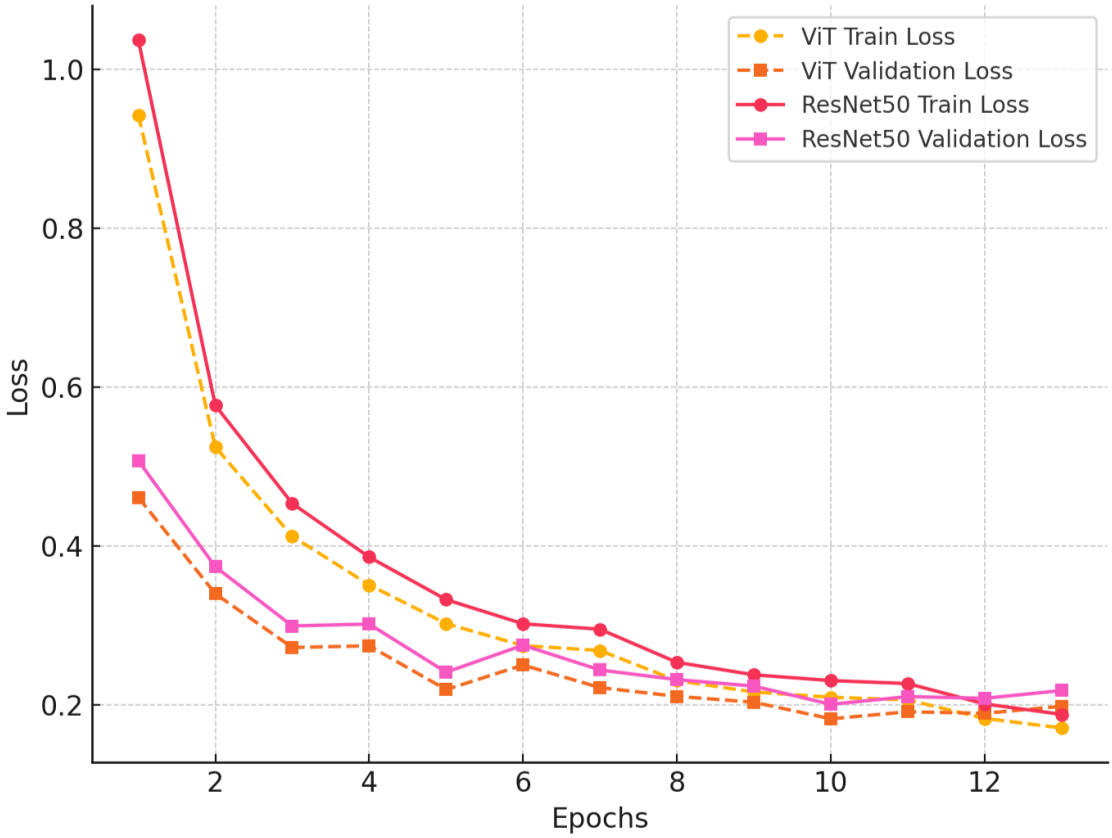


Figure 4.4: Fine-tuning for classification of six land-use categories: ResNET50 vs ViT

primarily rely on localized feature extraction. The quantitative performance evaluation of both models, shown in Table 4.6 reveals that ViT outperformed ResNet-50 across all major metrics. It achieved a 2.79% improvement in accuracy (93.96% vs. 95.22%), along with higher precision (0.952vs. 0.928), recall (0.952 vs. 0.923), and F1-score (0.947 vs. 0.927).

To further assess the strengths of the best-performing model (ViT), the confusion matrix in Table 4.7 was analyzed to identify misclassification trends

Metric	**ResNet-50**	**ViT-Base-Patch16-224**
Accuracy	92.43%	95.22%
Precision	0.928	0.952
Recall	0.923	0.952
F1-Score	0.927	0947

Table 4.6: Performance comparison of ResNet-50 and ViT-Base-Patch16-224

and their possible causes.

Class 1 (Residential) achieved a high classification accuracy of 97.39% (446 out of 458 correctly classified), with minor confusion with green areas (6 instances) and agricultural areas (1 instance). The most difficult class to classify was Class 10 (Train Stations), with only 30% accuracy (3 out of 10 correctly classified). The confusion between train stations and roads (6 instances misclassified as Class 36) likely stems from shared structural features, such as tracks and transit-related elements. Similarly, industrial areas (Class 12) were frequently misclassified, achieving only 42.86% accuracy (6 out of 14 correctly classified), with 8 instances being misclassified as residential areas. Class 26 (Agricultural Areas) achieved 100% accuracy, suggesting clear distinctions from other classes. Class 36 (Roads/Highways) performed well with 94.44% accuracy, with only one sample misclassified as train stations. On the other hand, Class 5 (Green Areas) had a high misclassification rate, with only 7 out of 11 correctly classified (63.64%), and misclassified as residential (3 instances) and roads (1 instance). This suggests that green spaces within mixed-use areas were difficult for the model to categorize correctly. Increasing sample representation for these categories could enhance classification performance. Overall, the findings indicate that ViT's self-attention mechanism provides advantages in capturing spatial relationships and long-range dependencies in satellite imagery, making it a strong candidate for land-use classification tasks. However, challenges remain, particularly concerning class imbalances and the misclassification of visually similar land-use

True Class \ Predicted Class	**1**	**10**	**12**	**26**	**36**	**5**
1 (Residential)	446	0	0	1	1	6
10 (Train Stations)	0	3	0	0	6	0
12 (Industrial Areas)	8	0	6	0	0	0
26 (Agricultural Areas)	0	0	0	8	0	0
36 (Roads/Highways)	0	1	0	0	17	0
5 (Green Areas)	3	0	0	0	1	7

Table 4.7: Confusion Matrix of the test set on the fine-tuned ViT model for the six land-use categories classification

categories. Incorporating external data sources, such as Google Maps Places API, presents a promising avenue for improving classification accuracy by complementing satellite imagery with real-world contextual information.

The next section will further investigate these misclassifications by performing feature analysis to understand how different land-use categories are represented in the model's feature space. By examining feature representations, we aim to identify patterns that contribute to misclassification and explore potential ways to refine the model's ability to distinguish visually similar classes more effectively.

Chapter 5

Conclusions & Future Works

5.1 Conclusions

This research proposed an innovative workflow for validating and classifying ISTAT census areas, addressing the limitations of outdated land-use classifications through a combination of geospatial data validation and deep learning techniques. The study explored two complementary approaches:

1. Google Maps Places API Validation, which was leveraged to cross-check ISTAT's classifications against real-world points of interest (POIs). This method was effective for structured land-use categories such as schools, healthcare facilities, and religious sites but showed limitations in validating more ambiguous categories like industrial and agricultural areas.

2. Deep Learning-Based Classification, where two state-of-the-art models -ResNet-50 (CNN) and ViT-Base-Patch16-224 (Vision Transformer)— were fine-tuned using high-resolution satellite imagery. These models were trained on a dataset that was validated and refined using GPT-4o-assisted re-labeling, ensuring that the classification models learned from a more accurate and up-to-date ground truth.

The results showed that ViT-Base-Patch16-224 outperformed ResNet-50 in both binary and multi-class classification tasks. The transformer-based architecture proved superior in capturing global spatial relationships, particularly in distinguishing land-use categories that require broader contextual understanding.

Despite these improvements, the study identified persistent challenges in classifying train stations and green areas, which were frequently misclassified due to their structural similarities with residential areas. The findings suggest that external geo-spatial datasets and additional contextual features could further improve classification accuracy.

This research demonstrated that a hybrid approach combining API-based validation, AI-assisted re-labeling, and deep learning classification offers a scalable and automated solution for continuously updating ISTAT's census classifications. The methodology developed in this study has practical applications in real estate, urban planning, and geographic analysis, providing decision-makers with more reliable and dynamic land-use data.

5.2 Future Works

While this research successfully introduced a novel framework for census area classification, there are several areas where improvements can be made to further enhance accuracy, scalability, and adaptability:

- **Expansion of Geospatial Data Sources**: the reliance on Google Places API for structured POI validation was effective but exhaustive of all the locations. Future research could incorporate additional datasets, such as OpenStreetMap, Sentinel-2 satellite imagery, and government GIS databases, to improve validation coverage for industrial, agricultural, and green areas.

- **Improving Class Balance and Training Data Augmentation**: The dataset exhibited class imbalances, particularly in underrepresented categories such as train stations and green spaces. Addressing

this issue through using data for all of Italian territory or additional labeled training data could significantly enhance model performance.

- **Integration of Multi-Modal Data for Improved Classification**: This study relied solely on satellite imagery and census metadata for classification. Future iterations of the model could integrate other data types such as street-view imagery or night-time satellite imagery.
- **Adapting the Approach for National-Scale Implementation**: While this study focused on Milan's census sections, the workflow is scalable and could be applied to the entire Italian territory.

5.3 Final Remarks

This research has demonstrated the feasibility and effectiveness of a hybrid approach to automating ISTAT census validation and classification. By integrating API-based geospatial validation, AI-assisted label correction, and deep learning models, this study provides a scalable, adaptable, and efficient methodology for addressing the limitations of static census classifications. The findings underscore the importance of continuously updating territorial data to reflect real-world changes, with direct applications in real estate market analysis, urban development planning, and policy decision-making. The methodology developed here has the potential to be extended beyond Milan, offering a framework for nationwide dynamic census classification. As deep learning technologies continue to evolve, the integration of multi-modal data sources, improved model architectures, and scalable automation pipelines will further enhance the accuracy and reliability of land-use classification. Future research in this domain holds the potential to transform geospatial analysis and urban planning, making census data more dynamic, accurate, and actionable in real time.

Bibliography

[1] European Spatial Agency. Sentinel-2. *ESA*, 2015. URL https://sentinel.esa.int/documents/247904/685211/sentinel-2_user_handbook.

[2] Shir Amir, Yossi Gandelsman, Shai Bagon, and Tali Dekel. Deep vit features as dense visual descriptors. *arXiv preprint arXiv:2112.05814*, 2021. URL https://arxiv.org/abs/2112.05814.

[3] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(12):2481–2495, 2017. doi: 10.1109/TPAMI.2016.2644615.

[4] Alexandre Banquet, Paul Delbouve, Michiel Daams, and Paolo Veneri. Monitoring land use in cities using satellite imagery and deep learning. *OECD Regional Development Papers*, 2022. URL https://www.oecd-ilibrary.org/docserver/dc8e85d5-en.pdf?expires=1670218043&id=id&accname=guest&checksum=62C4D938B5D7F408CD9DF9968107777F#:~:text=This%20study%20lays%20a%20methodological,large%20surfaces%20of%20land.

[5] Erika Cerasti, Fabrizio De Fausti, Angela Pappagallo, Francesco Pugliese, and Diego Zardetto. A deep learning approach to land cover estimation from satellite imagery. *Workshop on Methodologies for Official Statistics*, 2022. URL www.istat.it/wp-content/uploads/2022/11/abstract-3-_-2_DeFausti.pdf.

[6] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Springer*, 1995.

[7] Ashim Dahal, Saydul Akbar Murad, and Nick Rahimi. Heuristical comparison of vision transformers against convolutional neural networks for semantic segmentation on remote sensing imagery. *arXiv preprint arXiv:2411.09101*, 2024. URL `https://arxiv.org/abs/2411.09101`.

[8] Istituto Nazionale di Statistica. In ISTAT, editor, *Anagrafe della popolazione*, 1992.

[9] Istituto Nazionale di Statistica. In ISTAT, editor, *Basi Territoriali 2021 Metadati*, 2024.

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *Arxiv*, 2020. URL `https://arxiv.org/abs/2010.11929`.

[11] Vicky Espinoza, Lorenzo Ade Booth, and Joshua H. Viers. Land use misclassification results in water use, economic value, and ghg emission discrepancies in california's high-intensity agriculture region. *ResearchGate*, 2023. URL `https://www.researchgate.net/publication/370136675_Land_Use_Misclassification_Results_in_Water_Use_Economic_Value_and_GHG_Emission_Discrepancies_in_California%27s_High-Intensity_Agriculture_Region`.

[12] ESRI. In ESRI Conferenza Italiana 2022, editor, *Dalle microzone alle sezioni di censimento 2021: i numeri delle nuove BT dell'Istat*, 2022.

[13] Unione Europea. In Gazzetta ufficiale dell'Unione europea, editor, *Regolamento Comunità Europea N. 763/2008 del Parlamento Europeo e del Consiglio*, 2008.

[14] Google. Google maps platform: High-resolution satellite imagery, 2024. URL https://developers.google.com/maps. Accessed: 2024-03-08.

[15] Google. *Google Colaboratory*, n.d. URL https://colab.research.google.com/. Accessed: 2025-03-09.

[16] Kai Han, Yunhe Wang, Hanting Chen, Xinghao Chen, Jianyuan Guo, Zhenhua Liu, Yehui Tang, An Xiao, Chunjing Xu, Yixing Xu, Zhaohui Yang, Yiman Zhang, and Dacheng Tao. A survey on visual transformer. *Arxiv*.

[17] Mengmeng Hao, Xiaohan Dong, Dong Jiang, Xianwen Yu, and Fangyu Ding. Land-use classification based on high-resolution remote sensing imagery and deep learning models. *Plos One*, 2024. URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0300473.

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*, 2015. URL https://arxiv.org/abs/1512.03385.

[19] Environmental Systems Research Institute. Spatial references. URL https://developers.arcgis.com/documentation/spatial-references/.

[20] Repubblica Italiana. In Presidenza del consiglio dei ministri, editor, *Decreto Legislativo 6 Settembre 1989, N. 322*, 1989.

[21] Mehak Khan, Abdul Hanan, Meruyert Kenzhebay, Michele Gazzea, and Reza Arghandeh. Transformer-based land use and land cover classification with explainability using satellite imagery. *Nature*, 2024. URL https://www.nature.com/articles/s41598-024-67186-4.

[22] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in*

Neural Information Processing Systems (NeurIPS), volume 25, pages 1097–1105, 2012.

[23] D. Leite, I. Teixeira, R. Morais, J.J. Sousa, and A. Cunha. Comparative analysis of cnns and vision transformers for automatic classification of abandonment in douro's vineyard parcels. *MDPI*, 2024. URL `https://www.mdpi.com/2072-4292/16/23/4581`.

[24] OpenAI. Gpt-4o system card. *OpenAI*, 2024. URL `https://cdn.openai.com/gpt-4o-system-card.pdf`.

[25] PyTorch Team. *PyTorch Documentation*, 2025. URL `https://pytorch.org/docs/stable/index.html`. Accessed: 2025-03-09.

[26] PyTorch Team. *PyTorch: Regularization Techniques*, 2025. URL `https://pytorch.org/docs/stable/nn.html#torch.nn.L2Loss`. Accessed: 2025-03-09.

[27] Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy. Do vision transformers see like convolutional neural networks? *Arxiv*, 2021. URL `https://arxiv.org/abs/2108.08810`.

[28] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *Cornell Univeristy*, 2015. URL `https://arxiv.org/abs/1511.00561`.

[29] Ribana Roscher, Marc Rußwurma nd Caroline Gevaert, Michael Kampffmeyer, Jefersson A. dos Santos, Maria Vakalopoulou, Ronny Hänsch, Stine Hansen, Keiller Nogueira, Jonathan Prexl, and Devis Tuia. Better, not just more: Data-centric machine learning for earth observation. *Arxic*, 2024. URL `https://arxiv.org/abs/2312.05327`.

[30] Ava Vali, Sara Comai, and Matteo Matteucci. Deep learning for land use and land cover classification based on hyperspectral and multispectral

earth observation data: A review. *MDPI*, 2020. URL https://www.mdpi.com/2072-4292/12/15/2495.

[31] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pages 5998–6008, 2017. URL https://arxiv.org/abs/1706.03762.

[32] Qiusheng Wang. Leafmap: A python package for interactive mapping and geospatial analysis. *Journal of Open Source Software*, 8(88):1–5, 2023. doi: 10.21105/joss.04850.

[33] Yichen Wang, Yuting Huang, Induja R. Nimma, Songhan Pang, Maoyin Pang, Tao Cui, and Vivek Kumbhari. Validation of gpt-4 for clinical event classification: A comparative analysis with icd codes and human reviewers. *Wiley*, 2024. URL https://onlinelibrary.wiley.com/doi/10.1111/jgh.16561.

[34] Rikiya Yamashita, Mizuho Nishio, Richard K. G. Do, and Kaori Togashi. Convolutional neural networks: An overview and application in radiology. *Insights into Imaging*, 9(4):611–629, 2018. doi: 10.1007/s13244-018-0639-9. URL https://doi.org/10.1007/s13244-018-0639-9.

[35] Z. Zhu, C. E. Woodcock, J. Rogan, and C. Holden. The influence of spatial resolution on land cover classification accuracy: A case study with google earth imagery. *Remote Sensing of Environment*, 239:111629, 2020. doi: 10.1016/j.rse.2020.111629.

Ringraziamenti

Arrivati a questo punto, voglio ringraziare chi mi ha accompagnato in questo viaggio e mi ha sopportato per tutta la sua durata.

Mamma, Papà, mi avete sostenuto quando ero felice e soprattutto quando non lo ero. Anche se per la maggior parte del tempo lontani, mi siete stati vicino più di quanto immaginiate, anche con poche semplici parole. Non c'è nulla di più che un figlio possa volere dai propri genitori.

Fra, anche se sono il tuo fratello maggiore, in questi due anni mi hai insegnato più cose di quanto immagini. Spero di rimanerti sempre accanto per esserti da supporto come tu lo sei stato per me. E per mangiare quello che prepari.

Ai miei amici di giù, con cui sono cresciuto. La vostra capacità di farmi sentire a casa anche dopo lunghi periodi, come se il tempo non fosse mai passato, hanno reso questo percorso meno faticoso. Grazie per esserci sempre, anche quando il tempo e le distanze sembrano metterci alla prova.

Gio, il sostegno che mi hai dato è quello di una sorella acquisita. La tua amicizia ha un valore inestimabile e sono fiero di averti come punto di riferimento. So che te lo prometto sempre, ma il momento della mia pasta e patate arriverà presto.

Ai miei compagni di viaggio della Bicocca. Non avrei potuto desiderare di più da un gruppo come il nostro. Mi avete regalato momenti indimenticabili e tutto ciò che abbiamo condiviso ha reso questa esperienza unica. A mille altri spritz.

Alla Tessera, il numero di cavolate che abbiamo tirato fuori dal cilindro è incalcolabile. Ogni giornata è stata una nuova occasione per superare il limite della demenzialità, e ammettiamolo: ci siamo riusciti alla grande. Grazie per la tua capacità di trasformare anche la più banale delle situazioni in qualcosa di epico e per il tuo spirito indomabile. Non cambiare mai.

Alle mie coinquiline:
Angie, condividere casa con te è stata una fortuna (questa volta non per scherzo). Grazie per le chiacchierate infinite, i consigli e le cene improvvisate. Ma soprattutto grazie per tutti i post-it. Indipendentemente da cosa ci riserva il futuro, rimarranno sempre il simbolo di ciò che abbiamo condiviso.
Ale, hai completato un quartetto memorabile. La tua genuinità e la tua risata (molto facile) hanno contribuito a creare un ambiente speciale. Ovviamente sto scrivendo tutto questo nella speranza che non mi rubi la scala del letto. Grazie a entrambe per aver creato l'atmosfera di casa che c'è.

Per ultimo, ringrazio me stesso. Per avercela messa tutta anche nei momenti più difficili. Perché anche quando hai perso la bussola, hai continuato a lottare, in attesa di suonare i tamburi della liberazione.

Grazie,
Antonio (o come preferisci chiamarmi)