University of Sheffield
Department of Civil and Structural Engineering

A THESIS SUBMITTED IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY IN CIVIL ENGINEERING

Characterising English Residential Housing Using Street View Capture and Deep Learning Techniques

Menglin Dai *MEng (Hons.)*

in the

May 22, 2023

Declaration

It is certified that all contents cited in this thesis from other people's work have been specifically acknowledged by clear cross-referencing to author and work. I understand that failure to do this amounts to plagiarism and will be considered grounds for failure.

Name: Menglin Dai

Signature: Menglin Dai

Date: 22/05/2023

Abstract

Building retrofit is an important facet in the drive to reduce global greenhouse gas emissions. However, delivering building retrofit at scale is a significant challenge, especially in how to automate the process of building surveying. On-site survey by expert surveyors is the main approach in the industry. This can lead to a high workload if planning retrofit at a large-scale. Remote sensing technology can efficiently capture urban environmental data which contains substantial information that is essential in identifying building retrofit needs. However, how to extract the information required for retrofit is still a challenge.

Automatically recognising critical building components such as windows is the first step towards scalable building retrofit. A substantial number of retrofit-related properties can be directly inferred or indirectly achieved by integrating other types of data sources such as thermal images with known building components' positions. The process of automatically recognising objects at pixel level is commonly referred to as facade segmentation, which aims to divide a facade into several groups, each with distinct semantic meanings. The state-of-the-art works on facade segmentation are predominantly based on rectified images which would lose useful information needed for the retrofit purpose such as the side of buildings. In addition, existing datasets do not focus on English houses. Moreover, the ontology of facades in the existing facade segmentation datasets is inconsistent and they are not targeted at tackling building retrofit. Therefore, there are significant research gaps in 1) determining facade ontology for retrofit, 2) building English house-oriented datasets and 3) developing approaches for the specific retrofit-assisting facade segmentation task.

Two datasets were constructed first based on the determined facade ontology which considers the need for building inspection, energy analysis and stock analysis. The raw images of the two datasets were captured in regions of Sheffield, UK. Then two deep learning technique-based models were developed on the first built dataset. The two semantic segmentation models, called FacMagNet-l and -s are designed specifically for this task. FacMagNet-l aims to tackle class inter- and intra- size discrepancy problems. Accuracy is the priority over computational cost in designing FacMagNet-l. FacMagNet-s, which aims to balance the accuracy and computational cost, is a reduced version of FacMagNet-l. The two models have achieved 81.01% and 77.87% in mean intersection-over-union (mIoU) metric. The mIoU metric measures the overlapping ratio of the prediction and ground-truth. The results set a new standard for state-of-the-art semantic segmentation models on the built dataset.

The state-of-the-art and adaptability of the proposed FacMagNet-s model are further validated using the recently announced Oxford RobotCar-Facade dataset. The results have shown that the model has achieved a competitive performance compared to state-of-the-art approaches on the Oxford dataset. The representativeness of the two datasets built in this PhD project is validated quantitatively by applying trained models on the Oxford dataset. The representativeness is further qualitatively validated using raw data captured by the same rig in building the English house datasets in different geographic locations. These representativeness experiments show that built datasets can properly represent English housing stock.

Acknowledgements

Finally, after another four years in Sheffield, I will graduate from The University of Sheffield with a PhD degree. I still remember the night in the spring of 2013, more precisely, 26-Jan-2013, I dragged two huge pieces of luggage and carried a backpack, landing at Manchester airport, and had a nice talk about British pop music with the driver who I had booked to drive me to my accommodation. I also still remember the cold rains that night and the feeling of living in a remote foreign country for the first time. In my first four years in Sheffield, I acquired my MEng Civil Engineering degree. Besides, I have gained much precious experience such as working in a British hydrology consultancy, RPS Group, doing research as an undergraduate student under the supervision of Dr Jurgen Becque, meeting many valuable friends, etc. More importantly, I learned to be independent, confident and integrate into local society. Without the MEng course, I would probably be in another place in the world and losing so many precious memories in Sheffield.

Firstly, I want to thank my supervisors: Prof. Martin Mayfield-Tulip and Dr Danielle Densley Tingley. They are the best supervisors I could ever have for my PhD study. Martin, thanks for the coincidental talk on the open day; without that talk, I would probably never have thought about doing a PhD outside the conventional civil engineering area. Also, thanks for keeping finding funding for me as well during my PhD. Dani, thanks for all your efforts in guiding, supervising and supporting me throughout my PhD. You really inspire me in many ways and I have learned a lot from you. Thanks go to Dr Hadi Arbabi and Dr Wil Ward for their support and collaboration over the past two years. I also want to thank other members of my research group, in alphabetical order of family name, with special mention to Dr Oktay Cetinkaya, Dr Maud Lanau, Dr Xinyi Li and Dr Gregory Meyers.

I want to acknowledge my parents Xinhua Dai and Qin Pan for their unlimited and voluntary support, mentally and financially, especially my mom. She called me frequently to check my health and provide assistance and support. Gratitude goes to my girlfriend Shanshan Gao as well, for her mental support and physical companionship for several years. She is always there and does not abandon me for the distant relationship. Thanks also go to my best friends in the department: Xitong Sun, Zhangjie Peng, Weijiang Yu and Shiyao Wang for our friendship and for always being supportive during my PhD journey.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

1.1.1 Research Background

Under the circumstances of increasingly severe global climate change, reducing greenhouse gas (GHG) emissions is a worldwide challenge. In an effort to keep the global temperature rise well below 2°C, the Paris Agreement was made during the 2015 United Nations Climate Change Conference (United Nations, 2016), in which involved parties must regularly report their carbon emissions to strengthen transparency and put forward their nationally determined contributions, as well as developing mitigation strategies and technologies (United Nations, 2016). To reach the 2°C target, many governments have imposed their ambitious national goals. In 2019, the UK government committed to reducing the UK's net GHG emissions by 100% of their 1990 levels by 2050 (Great Britain, 2019).

Globally, buildings contribute a large amount of GHG emissions during their construction and operation phase (Ma et al., 2012), as well as consuming a significant proportion of end-use total energy is (Lucon et al., 2014). In the European Union, the energy demand of buildings is estimated to reach 40% of EU energy consumption, and buildings are responsible for 36% of GHG emissions (European Commission, 2020). Despite the relatively low building renovation rate, around 1%, and that 75% of the building stock is energy inefficient in the European Union (European Commission, 2020), existing buildings will continue to contribute a large amount of GHG emissions if there is no intervention. The building renovation rate is the ratio of buildings which have been renovated for energy efficiency in each year over the total building stock. In the UK, across all sectors contributing towards GHG emissions in 2019, residential buildings are responsible for 15% of the total GHG emissions (Department for Business, Energy & Industrial Strategy, 2021) and consume 29% of the total energy (Department for Business, Energy & Industrial Strategy, 2020) in the UK. In the meantime, the GHG reductions from the residential sector and efforts to adapt the current housing stock for the climate change risks are stalled (Department for Business, Energy & Industrial Strategy, 2021; Commitee on Climate Change, 2019). Due to this situation, the UK Climate Change Committee has requested housing retrofit to be an infrastructure priority (Commitee on Climate Change, 2019).

Building retrofit has significant potential to decrease GHG emissions owing to the uptake of energy efficiency. Prior to deploying a retrofit plan for an individual building, a data collection process needs to be conducted in order to collect data required to assess the building's energy condition (Ma et al., 2012). This process collects the key building data especially for thermal characteristics e.g. building geometry, construction materials, glazing ratio, window/door type, etc., and fault information. However, the building survey data are not always available and commonly rely on on-site surveys (Densley Tingley, 2022). Such a highly labour-intensive and time-consuming process makes it extremely difficult to conduct building retrofit on a large scale. The state-of-the-art works contributing to large-scale retrofit are predominately based on using existing building data and available energy simulation software packages to automate the energy analytic process (Hong et al., 2020). If the decarbonisation of the building stock is to be accelerated and scaled, it is vital that the methods of automating the gathering of spatial information are developed. The generation of this information from spatial data is a key aspect of delivering upon this challenge. Within this, collecting and integrating building data in an efficient way is still a big data challenge in both industry and academia for individual house retrofit on a large scale.

Vehicle-mounted sensors provide a common way of collecting urban environmental data at scale. A famous example is the Google Street View (Anguelov et al., 2010) project which captures global image data in the urban environment. More examples have emerged in the past few years with additional point cloud and thermal image data available to support the autonomous driving research (Huang et al., 2018; Choi et al., 2018). Given the successes of these vehicle-mounted sensor platforms, a multi-spectral data collection platform has been built to support the urban building retrofit research (Meyers et al., 2019). The platform is vehicle-mounted and designed to collect visual images, thermal images, hyperspectral images and point clouds with a high automation level. The platform is named MARVEL (Multispectral Advanced Research VEhicLe). The collected data contains substantial building information which can be used to identify the building retrofit requirements, e.g. point cloud data contains building geometry information, thermal image and hyperspectral image data can be used to identify thermal faults and facade component material types.

Although MARVEL is designed to capture multi-spectral data and its motions are recorded by a GNSS (Global Navigation Satellite System)/IMU (Inertial measurement unit) localisation/acceleration unit. How to automatically identify building components from collected data is still a challenge. Manually annotating captured image data is an unacceptable option as it requires heavy labour and funding. Identifying targets in an image at pixel level refers to semantic segmentation, which is a significant research problem in the computer vision community. The semantic segmentation technique aims to segregate an image into individual groups, each with distinct semantic meanings. Using the deep learning technique on solving semantic segmentation problems on these datasets has achieved remarkable success. In the past five years, deep learning models have dominated all image semantic segmentation ranking lists such as Cityscapes (Cordts et al., 2016) and Apolloscape (Huang et al., 2018).

In summary, reducing building carbon emissions is an urgent issue. Scalable building retrofit plays an essential role in reducing building carbon emissions. With more efficient data capture systems becoming available, the development of these systems has provided a feasible way of collecting built-environmental data at scale. Automatically identifying key building elements is a critical step towards scalable retrofit using remote sensing data. For this purpose, existing publicly available datasets recognising building facades and state-of-the-art approaches should be reviewed first to identify potential research gaps. The following subsections will further consolidate the research target and identify potential data requirements.

1.1.2 Target House Elements and Data Inclusivity

This section provides an overview of the key building elements required and other features which the collected data should have in order to assess building retrofit. According to the low carbon domestic retrofit survey guidance (Smith, 2011), latest literature Densley Tingley (2022) and RICS professional guidance on surveys of residential properties (RICS, 2016), the key elements of building fabrics are listed below with descriptions of how they would contribute to building conditions.

1. Walls and Roofs: their construction types such as whether or not a wall has a cavity will affect the calculation of the thermal transmittance (U-values); it is also critical to know whether any areas of these two elements require maintenance, e.g. due to damp.

2. Windows and Doors: the air-tightness of windows, including their frames and glazing types, and doors will affect the U-value calculations as well; whether or not these elements require replacement is significant in a building inspection survey.

3. Chimneys: they might help to estimate the house ventilation type. More importantly, as chimneys are commonly built on top of roofs it is essential to differentiate them from the roof to provide a more integrated roof structure.

Apart from identifying the key elements listed above, the data collected must also contain other essential features and elements in order to further assist building retrofit; these necessities are listed below.

1. Including surrounding environmental information: building retrofit analysis at large scale, e.g. community level or higher, is not only about assessing individual buildings. According to the residential property inspection guidance (RICS, 2016), external features such as trees, plants, drives, boundary walls also need to be assessed for identifying safety issues and maintenance requirements. Furthermore, environmental information can be used to assess urban street-level quality (Xia et al., 2021).

2. Imaging residential buildings from multiple perspectives can provide more information than a plain facade image. In Zeppelzauer et al. (2018), the authors have demonstrated that multi-perspective building images contain features that can be used to estimate building age which is an important indicator in building energy modelling.

3. Capturing high-resolution and low-noise images: collected images should have sharp boundaries of building elements and contain distinct details to be used further for extracting building properties. For example, wall construction types could be inferred from wall textures. Stone walls of Victorian buildings show different textures compared to brick walls. Cavity walls typically have different outer brick layouts compared to solid brick walls. By manually measuring the dimensions of standard building objects such as bricks and mortar joints, the resolution has been determined to be above 150 pixels/meter. The noise level which can be measured by Peak Signal-to-Noise Ratio (PSNR) should be higher than 50dB.

1.1.3 Deep Learning Technique

This section provides an overview of the main technique which is used in this thesis. Deep learning technique is a branch of machine learning. Machine learning (ML) technique is the science and art of enabling a system to improve from experience without being explicitly programmed. The technique has provided a substantial number of options in automating the process of data analysis. Machine learning approaches have greatly reshaped many areas owing to their high efficiency in extracting desired information from large quantities of data by converting real-world data to some forms of abstract understanding. Machine learning is commonly defined as a set of algorithms which can learn from data as in Goodfellow et al. (2016).

Mitchell (1997) provide a concise definition of the phrase 'learning' in the machine learning area: for a certain type of task T, and performance evaluation metric P, a computer algorithm is able to 'learn' from experience E, meaning that once the algorithm is advanced through E, the algorithm performance which is measured by P is improved on T. The machine learning definition has indicated that the source data, i.e. experience E, is critical to the success of a machine learning algorithm. Moreover, choosing appropriate machine learning algorithms is also significant to the task T. Machine learning algorithms contain two major sectors based on different forms of experience E: supervised and unsupervised learning. The supervised learning algorithms will learn from annotated data. Annotated data has corresponding input-output pairs and a supervised learning algorithm will learn the mapping relationship between these pairs. The unsupervised learning will learn from unlabelled data by extracting the data source's valuable structure properties.

Deep learning technique, which is also a machine learning algorithm, has been briefly mentioned in the former section as it has achieved great success in autopilot oriented datasets. These datasets commonly contain a building sector but do not specifically include residential buildings and not at the component-level. Deep learning approaches have also achieved remarkable success in general image semantic segmentation tasks. Bearing this in mind, utilising deep learning technique to identify residential building facades seems highly likely. However, when this PhD project began in 2019, works using deep learning techniques on facade segmentation were very limited. Furthermore, deep learning-based semantic segmentation models vary distinctively in their structures and training strategies for different tasks. Thus, exploring a suitable model structure and determining its training strategy is significant.

Data scarcity is a universal obstacle among all deep learning tasks. Training deep learning models requires a significant amount of data. As aforementioned, deep learning models will learn from experience E to improve the performance measured by P on task T. This indicates that scale and diversity of a dataset would determine the upper-limit of the generalisation of an applied deep learning model to the real world. The generalisation of a deep learning model to the real world is the key aspect of whether the model can be deployed in the real world.

1.2 Research Scope

1.2.1 English Houses

This section helps to set the research context, by providing an overview of the English housing Stock. Four key documents are used to demonstrate the house information statistics: English housing survey report (Ministry of Housing, Communities & Local Government, 2018), BRE (Building Research Establishment) English housing stock report (Piddington et al., 2020) and TABULA building typology database (Loga et al., 2016).

In both TABULA building type database and the BRE report, house typologies are distinguished by their age bands and types. The BRE report has five age bands which are pre-1919, 1919-1944, 1945-1964, 1965-1980, 1981-1990 and post-1990, while the TABULA database further divides houses built after 1990 into three periods which are 1991-2003, 2004-2009 and post-2010 (Piddington et al., 2020; Loga et al., 2016). Buildings established pre-1945 are likely to be built with solid walls (Piddington et al., 2020) which commonly lack insulation layers (Li and Tingley, 2021). These buildings account for approximately 40% of the total stock (Piddington et al., 2020).

House types are divided into five categories in the BRE report which are terraced, semi-detached, detached, bungalow and flat. The presentations of the first four building types are clear, which are one- to two-storey buildings with pitched roofs (Piddington et al., 2020). Flats are more varied in their definitions, especially for high-rise buildings, their appearances are different to others. In the UK, the majority of the housing stock is one-to-two storey, accounting for 79.1% of the total UK housing stock (Ministry of Housing, Communities & Local Government, 2018). In Wales the number is significantly higher at 89% (Piddington et al., 2020). Although around 20% of people live in flats, 70% of them live in flats with fewer than three storeys and only 10% of them live in flats with more than five storeys, which is equivalent to 2% of the total households (Ministry of Housing, Communities & Local Government, 2018).

This PhD project focuses on building datasets and developing semantic segmentation approaches for English houses which are up-to three storeys. The focused building type covers 96% of the total housing stock according to English housing survey results (Ministry of Housing, Communities & Local Government, 2018).

1.2.2 The Data Capture Platform

1.2.2.1 Acquisition Platforms and Cameras Comparison Study

This section provides a comparative study of platforms which could be used for scalable built-environment data acquisition. The comparative study is based on the latest (6th edition) RICS guidance on earth observation and aerial surveys (RICS, 2021), which came into effect from 4 January 2022, other developed platforms for city modelling and current sensors are used for capturing facade images. The current research on facade segmentation is predominantly based on terrestrial image data, but there are emerging works using aerial oblique view data (Mao et al., 2022) as well.

According to the RICS guidance, for civil engineering and infrastructure use case, UAV (Unmanned Aerial Vehicle) and manned helicopter are recommended. UAV is recommended for small-size area and helicopter is recommended for medium-size area surveys. Vehicle, e.g. a van, is also a popular platform in capturing built-environment data as briefly reviewed in section 1.1.1. UAV has two types, multirotor such as the DJI Matrice series and fixed-wing such as the eBee-X drones. DJI is a multirotor drone manufacturer with the largest market share (Clark et al., 2017) and eBee-X is one of an industrialised solution of city modelling (Hu et al., 2022).

The target building type in this research is houses below three storeys, usually about 6-8 metres high (two-storey) but can be up to 10-12 meters (three-storey), if the average height of a storey is assumed to be c.3 meters. Narrow streets are common in the UK, this means, if using aerial photography to capture house facades, the minimum flight height regulations should be considered first in order to cover the whole facade area properly. A demonstration diagram using an oblique view camera for facade data capture is shown in figure 1.1. An oblique view camera has multiple sensors, one points towards the ground providing nadir view images. As an example, its field-of-view (FOV) is shown in light blue in figure 1.1. Other sensors are inclined at an angle and provide oblique views.

As shown in the diagram, if an aircraft is flying at its minimum height H and trying to capture the facade of building A, due to convex structures such as roof overhang or outer doorways, a portion of a facade will not be captured. The higher the altitude H, the angle θ will be larger which means a larger area of the facade will be occluded. This problem would be severe for houses with doorways or lean-to roofs. If the aircraft has to be a certain distance from targets to get an appropriate image angle, and the street whose width is denoted as L, is narrow, building B could occlude the lower part of building A.

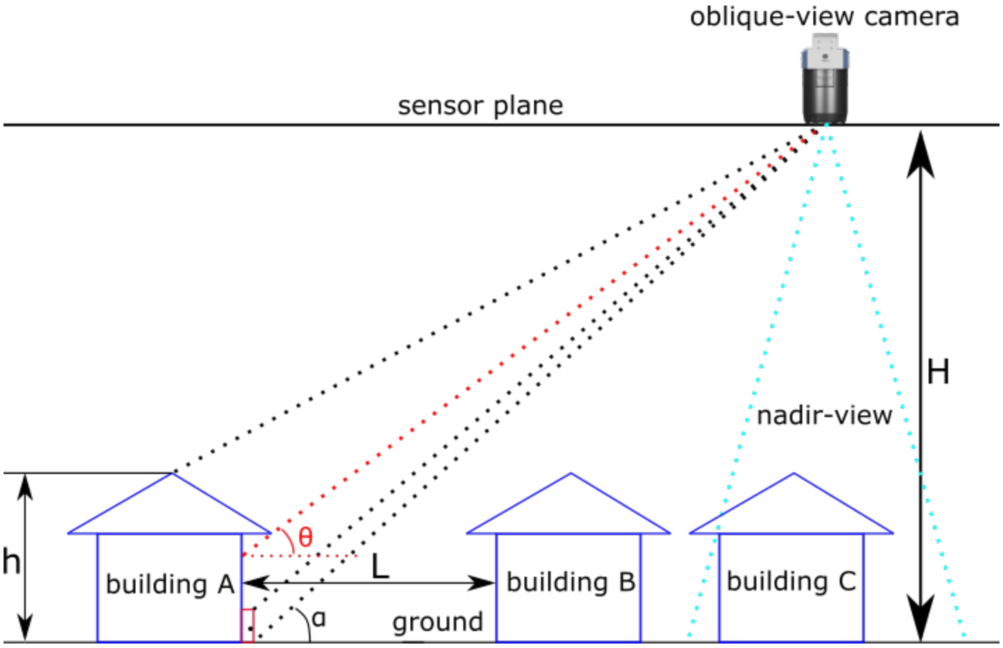


Figure 1.1: *Demonstration diagram using oblique view camera to capture house facades. A typical oblique-view camera has five lenses, four of which are inclined at an angle which is commonly 45° (oblique view) and the one remaining is perpendicular to the flying plane (nadir view). The letter h denoted the height of the building, L is the street width, H is the flying height of the platform.*

In this case, helicopter might not be an ideal choice. In the UK, helicopters are not allowed to fly "closer than 500 feet (c.150 metres) to any person, vessel, vehicle or structure" (Department of Transport, 1996). In order to achieve a favourable camera angle to avoid occlusions caused by roof overhang structure or outside doorways, a helicopter might have to be far from the target. For example, if a light-ray angle, α of 45° is needed for a door, the helicopter needs to be approximately 150m horizontal to the target. In addition, light rays reflected from the surface of the door will be occluded by building B if the height of a building, h, is larger than the street width L. As an example, a street that is 6 metres wide could lead to occlusion if building B has a height of 7 metres. In similar cases, targets, especially for ground floor features, could be easily occluded by their neighbouring structures.

Vehicle and UAV are two potential platforms to capture facade data. In comparison with UAVs, vehicles are more advantageous in capturing features of ground level. As an example in figure 1.2, if using a spherical view camera for facade capture, convex structures such as the roof overhang will not result in loss of features. However, terrestrial data capture is less favourable in capturing roof structures than using aerial platform due to limited FOV, especially if there are skylights or dormers.

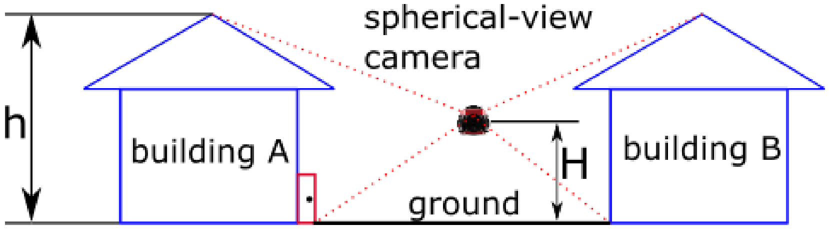


Figure 1.2: *Demonstration diagram using spherical view camera to capture house facades. A typical spherical view camera will have one lens point upwards and the others point parallel with the ground. The letter h denoted the height of the building, H is the position height of the camera.*

This section reviews the state-of-the-art vehicle-based and UAV-based platforms that could be or have been used for facade image capture, or have been built for city modelling purposes in table 1.1. Four platforms are included which are Google Street View van which was built for urban environment mapping (Anguelov et al., 2010), Oxford RobotCar (Maddern et al., 2017) which was built in 2014 but its data were used to build the latest facade segmentation dataset (Wang et al., 2022), DJI Matrice 300 which is an industry-level multirotor drone and eBee-X which is an industry-level fixed-wing drone. This review includes their cruising distances, equipped or potentially equipped sensors with their effective pixels.

It is noted that only enterprise-level UAVs are included in this review to provide a fair comparison to professional terrestrial platforms. As stated, UAV can be categorised into two types, multirotor and fixed-wing. In general, fixed-wing UAVs will have longer cruising range than multirotor UAVs (Boon et al., 2017). Popular manufacturers include DJI and Parrot for their multirotor aircraft and eBee and Skywalker for their fixed-wing models. Therefore, both fixed-wing and multirotor models are included in this review. This review does not include hand-held sensors which have been predominantly used for building facade segmentation datasets, due to their limitations on data capture efficiency.

Table 1.1 has shown that compared to UAVs, vehicle-based platforms have a longer cruising range. Although eBee-X has a maximum cruising range of 55 km, its signal transmission distance limit can only fly within 8 km. The latest sensors for vehicle or UAV are equally powerful in taking images. In order to capture facade images, using UAVs or vehicles are both feasible solutions. Multirotor drones are more flexible in operation than fixed-wing drones. Therefore, for the specific facade image capture task, multirotor drones are more advantageous. Compared to vehicles, multirotor drones have more restrictions such as keeping drones 'within visual line of sight'. Such restrictions could make using UAVs in high-density residential areas become complicated. Different camera lens are then reviewed below. The CCD and CMOS sensors and various camera types are also reviewed.

Table 1.1: *A review of state-of-the-art data capture platforms including vehicle-based and fixed-wing and multirotor UAV-based models.*

Platform	**Google Street View**	**Oxford RobotCar**	**eBee-X**	**DJI Matrice 300**
Year announced	2010	2014	2018	2020
Carrier type (model)	various oil-driven vehicles (Chevrolet)	electrical vehicle (Nissan LEAF)	fixed-wing drone	multirotor drone
Camera type	custom controlled-distortion	trinocular stereo + wide-angle	oblique (two oblique and one nadir views)	wide angle + zoom
Sensor	custom CMOS	Sony ICX445 (1/3-inch CCD)	1-inch RGB	1/2.3-inch and 1/1.7-inch CMOS
Effective Pixels	5MPx	1280 × 960	5472 × 3648	20 and 12MPx
Max. cruising or transmission range (battery life)	1000km	270km	8km (90min)	15km (55min)

Imaging sensors for cameras include CCD (Charge-Coupled Device) and CMOS (Complementary Metal Oxide Semiconductor). Both of the sensors work by capturing light photons and converting them into electrons. Over decades, Litwiller (2001) stated that CCD was the choice of high image quality preference but CMOS was a cost-effective solution. However, nowadays, CMOS is a more popular choice owing to technology development (TELEDYNE FLIR, 2021). Table 1.1 also shows that latest data capture platforms prefer to use CMOS sensors.

Camera lenses are critical in photography. The basic classification of lenses include prime and zoom lenses. Prime lenses have fixed focal length and zoom lenses can adjust their focal length within a fixed range as required. Lenses can also be categorised based on their angle of view which can usually be specified by their focal length but it also depends on their film format. A normal lens will provide a view which approximates the view of human eyes. Wide-angle and telephoto lenses will provide wider and narrower views compared to normal lenses, respectively. The four platforms that have been reviewed all employ wide-angle cameras.

Another consideration of cameras for environmental data capture is camera types, although it also based on which platforms have been employed. Oblique view camera can only be installed on aircraft. In addition, UAV manufacturers provide sensors that could be installed on their platforms. Those sensors are typically designed for their platforms and some of them are designed for specific tasks such as mapping or anti-terrorism. For example, DJI recommends its Zenmuse-h20 camera which integrates prime wide angle and zoom visual cameras, thermal camera and laser distance measure (DJI, 2022). eBee-X recommends its

S.O.D.A. 3D camera which is specifically designed for mapping (AgEagle, 2022). Cameras which can be mounted on a vehicle are various, from common monocular cameras to integrated models such as spherical and stereo cameras. Among them, the spherical camera is a choice for mobile mapping such as in the Google Street View project (Anguelov et al., 2010). This type of camera integrates several individual sensors to provide a spherical view of the environment.

1.2.2.2 Multi-spectral Advanced Research VEhicLe (MARVEL)

In this PhD project, the built multi-spectral advanced research vehicle (MARVEL) platform is adopted for facade image capture. The camera used to capture images in the platform is a Ladybug5+ spherical camera. The camera rig comprises six separate 2/3-inch Sony IMX264 CMOS sensors with 2048 × 2448 effective pixels and wide-angle lenses. The cameras are oriented with one on the top pointing upwards and the other five positioned horizontally along the sides forming a regular pentagon. The combined capture has a field-of-view (FOV) of 90% of full sphere. Figure 1.3 demonstrates the MARVEL rig.

MARVEL is designed to capture buildings up-to three storeys which are beside roads while driving. Standardised widths for new UK roads were set in 1993. However, to tackle narrower roads, the design single-side width is reduced to 2.25 m from the standardised 3.65 m. A single building storey is estimated to be 3 meters. Under these assumptions, the approximate required FOV for a three-storey house is 84.3° which is within the sensor's FOV. More detailed design specifications are available in the MARVEL paper (Meyers et al., 2019).

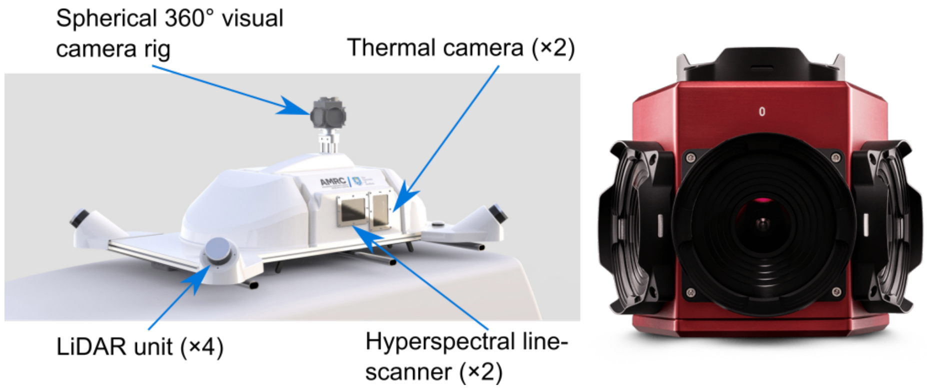

Spherical 360° visual camera rig
Thermal camera (×2)
LiDAR unit (×4)
Hyperspectral line-scanner (×2)

Figure 1.3: *The developed multi-spectral data collection platform, MARVEL (Meyers et al., 2019). The visual camera rig is installed on the top of the platform; one LiDAR unit is installed on each of the four corners of the platform; thermal cameras and the hyperspectral line-scanners are installed on both sides of the platform. The spherical Ladybug 5+ camera is shown on the right hand side of the figure.*

1.3 Aims and Objectives

The first aim of this thesis is to build datasets focusing on English housing stock with up-to three storeys. The second aim is to develop approaches for facade semantic segmentation using street view images. The research questions, and the objectives for answering them, that form the skeleton of this research are outlined below.

1. What is the state-of-the-art (SOTA) in facade semantic segmentation including datasets and approaches?
   - (a) The SOTA approaches developed for facade semantic segmentation and existing datasets are comprehensively reviewed;
   - (b) Research gaps between SOTA and the research aims of this thesis are identified.
2. How can key English house components (determined in section 1.1.1) be automatically identified from street view images?
   - (a) A dataset construction pipeline is proposed and developed;
   - (b) Deep learning-based approaches for built facade segmentation datasets are designed or explored;
   - (c) A publicly available benchmark dataset based on Oxford is used to validate the generalisation and adaptability of built datasets and designed approaches.

1.4 Road-map and Contributions

The thesis contains five chapters. Chapter 1 introduces the research background and proposes research questions. In Table 1.2, the road-map and contributions of this thesis are summarised by showing research challenges, associated contributions to knowledge which have been made in this thesis with corresponding chapters. The research problems are based on objectives described previously. All contributions, outcomes and findings are summarised and discussed in chapter 5.

Table 1.2: *Proposed research questions and challenges addressed in this thesis by relevant chapters.*

Research Challenge	**Thesis Contributions**	**Chapters**
Construct street view-style building image datasets fitting retrofit needs	1. Review existing facade segmentation datasets and identify research gap on datasets availability 2. Develop an annotation pipeline for the specific task 3. Propose annotation rules fitting the task needs 4. Validate the feasibility of the annotation pipeline	2, 3
Develop building facade segmentation approaches for street-view facade images	1. Review SOTA approaches on facade segmentation and deep learning architectures 2. Develop deep learning-based facade segmentation approaches for built datasets 3. Design comparative studies with other SOTA approaches and architectures to validate the performance of developed approaches and representativity of built datasets	2, 4

1.5 Publications to Date

The following papers have resulted from works conducted in this thesis:

1. Dai, M., Ward, W.O., Meyers, G., Tingley, D.D. and Mayfield, M. (2021) 'Residential building facade segmentation in the urban environment', *Building and Environment*, 199, p.107921.

2. Dai, M., Ward, W.O., Arbabi, H., Densley Tingley, D., Mayfield, M. (2022, September) 'Scalable Residential Building Geometry Characterisation Using Vehicle-Mounted Camera System', *Energies*, 15(16), p. 6090. https://doi.org/10.3390/en15166090

3. Dai, M., Densley Tingley, D., Jurczyk, J., Ward, W.O., Arbabi, H., Mayfield, M. (2023) 'Residential Building Material Stock Characterisation and Quantification Using Computer Vision Techniques with Drive-by Image Capture', Ongoing.

4. Arbabi, H., Lanau, M., Li, X., Meyers, G., Dai, M., Mayfield, M. and Densley Tingley, D. (2022) 'A scalable data collection, characterization, and accounting framework for urban material stocks', *Journal of Industrial Ecology*, 26(1), pp.58-71.

5. Ward, W.O., Li, X., Sun Y., Dai, M., Arbabi, H., Densley Tingley, D., Mayfield, M. (2023) 'Estimating Energy Consumption of Residential Buildings at Scale with Drive-by Image Capture', *Building and Environment*, 234, p.110188.

Work from this thesis has also been presented in the following conferences:

1. Dai, M., Meyers, G., Tingley, D.D. and Mayfield, M. (2019, October) 'Initial investigations into using an ensemble of deep neural networks for building façade image semantic segmentation' in Remote Sensing Technologies and Applications in Urban Environments IV (Vol. 11157, p. 1115708). International Society for Optics and Photonics.

2. Ward, W., Dai, M., Arbabi, H., Sun, Y., Tingley, D. and Mayfield, M. (2022, September) 'Measuring the Cityscape: A Pipeline from Street-Level Capture to Urban Quantification', in IOP Conference Series: Earth and Environmental Science (Vol. 1078, No. 1, p. 012036). IOP Publishing.

Chapter 2

Literature Review

2.1 Introduction

In (Koziński et al., 2014), the authors define facade segmentation as ‘The goal of facade parsing is to segment rectified building images into regions corresponding to architectural elements’. Facade segmentation, as a branch of image semantic segmentation, has attracted much attention in recent years (Dai et al., 2021; Zhang et al., 2022; Ma et al., 2022). Automatically understanding facade decomposition would contribute towards many applications such as digital entertainment, building information modelling (BIM), etc.

Image semantic segmentation is a branch of image segmentation, a high-level computer vision task which aims to partition a digital image into various subgroups of pixels. The image segmentation family contains three main branches: the semantic segmentation, the instance segmentation and the panoptic segmentation. The differences between the three branches is the fineness level of the partitioned pixel subgroups. Semantic segmentation is the foundation of the other two branches. Semantic segmentation tasks aim to classify each pixel into different categories with distinct semantic meanings (Liu et al., 2019). Instance segmentation detects different objects and further identifies each individual instance of interest belonging to the same category (He et al., 2017). Panoptic segmentation unifies the semantic segmentation and the instance segmentation by categorising objects of interest into things and stuff (Kirillov et al., 2019). ’Things’ refer to countable objects such as animals and trees which are partitioned at instance level. ’Stuff’ is an amorphous region of similar texture or material, such as road and sky, which is partitioned at semantic level.

The classification of image semantic segmentation methods varies. A common viewpoint is that the classification depends on whether or not the methods are deep learning technology-based (Liu et al., 2019). The methods which do not employ deep learning technologies are usually referred to as 'traditional methods'. On the contrary, methods that do employ deep learning technologies are named 'deep learning methods'. Deep learning-based methods have a considerably shorter history than traditional methods but generally outperform them (O'Mahony et al., 2019). Traditional methods are usually about designing handcrafted feature extractors. Deep learning-based methods are about designing learnable feature extractors. A feature, especially of an image, is a piece of information which relates to whether a certain region in an image has specific properties. Features include various forms such as edges, corners, colours, etc. (Nixon and Aguado, 2019). Basic features such as points and edges which do not contain any shape information are commonly defined as 'low-level', and features concerning finding shapes and objects are termed 'high-level'. Compared to traditional methods, the feature extraction process commonly requires more human-intervention than the deep learning-based methods.

In this chapter, the deep learning-based general-purpose image semantic segmentation is first reviewed. The review will include state-of-the-art model architectures and frequently-used publicly available datasets for autonomous driving. Some deep learning basics are concisely included as an introduction to this research domain. Then, the facade semantic segmentation research is comprehensively reviewed. The review will contain techniques developed over the past decade and publicly available datasets.

2.2 Convolutional Neural Network-based Image Segmentation

2.2.1 Introduction

Convolutional neural network (CNN) is an assembly of a series of mathematical operations. The overall target of CNN is to minimise a designed loss function. Hornik et al. (1989) establish that multi-layer feed-forward networks which include networks that only have one hidden layer can theoretically approximate any Borel measurable functions to any given accuracy. The convolution layer is the core building unit of a CNN. It contains a set of learnable filters which can be computed as below in the forward propagation phase:

$$y_{u,v} = \sum_{i=0}^{m-1} \sum_{j=0}^{n-1} k_{i,j} x_{i+u,j+v} + b_{i,j}. \tag{2.1}$$

In which y is output, x is input, k is a filter with size (m, n) which should be smaller than the input size and b is bias. It is noted that the filter k is also widely referred to as *kernel* and the output y is sometimes named *feature map*. In comparison with node-based neural networks, convolution operation is more efficient in processing multi-dimensional input, both in reducing memory requirements and improving statistical efficiency (Goodfellow et al., 2016).

Similar to the node-based neural network, training a CNN is based on the backpropagation (BP) algorithm which adjusts kernels' weights at the direction of gradient descent at a given point. As an example, in a neural network with a single hidden-layer, the weight adjusted from the output layer is calculated by:

$$w'_{hj} = w_{hj} - \Delta w_{hj} \tag{2.2}$$

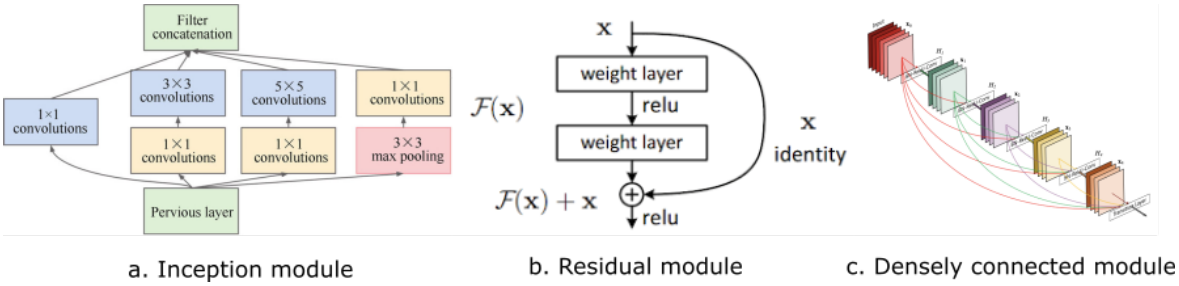
The Δw_{hj} is the gradient of the loss function to the weights connected to the output j, multiplied by a given *learning rate* η:

$$\Delta w_{hj} = -\eta \frac{\partial E_k}{\partial w_{hj}} \tag{2.3}$$

As the principle of CNN indicates, the performance of a CNN model can be influenced by its model architecture, loss function, optimiser, weight initialisation as well as hyper-parameters such as learning rate and batch size. Among these topics, developing more powerful CNN architecture has attracted much attention.

The state-of-the-art (SOTA) architectures developed for semantic segmentation have the same encoder-decoder structure (Hao et al., 2020). This paradigm was first introduced by (Long et al., 2015) in 'Fully convolutional networks (FCN) for semantic segmentation' in 2015. The work is also widely accepted as the foundation of the CNN-based semantic segmentation models. FCN pioneers semantic segmentation using convolution layers to replace dense-connection layers to enhance pixel-wise predictions. The author also develops the skip-connection which aims to recover information lost due to pooling operations by combining the finer low-level feature maps and coarser high-level feature maps.

The encoder-decoder structure consists of two connected modules: the encoder network and the decoder network. The encoder network aims to extract image features and the decoder network recovers the spatial resolution. The encoder network is also called the backbone network. The backbone network is the main structure of a CNN model for image classification tasks without dense connected layers. Commonly-used backbone networks include VGG (Simonyan and Zisserman, 2015), Inception (Szegedy et al., 2015, 2016, 2017), ResNet (He et al., 2016*a*), DenseNet (Huang, Liu, Van Der Maaten and Weinberger, 2017) and architectures based on the more recent neural architecture search (NAS) technology (Tan and Le, 2019). Figure2.1 visually compares the three SOTA backbone network architectures.



a. Inception module b. Residual module c. Densely connected module

Figure 2.1: *The summary of three convolution neural network modules; a. is the inception module (Szegedy et al., 2015), b. is the residual module (He et al., 2016*a*) and c. is the dense connection module (Huang, Liu, Van Der Maaten and Weinberger, 2017).*

Among these frequently-used backbone networks, VGG is a stack of 3 × 3 convolution kernels which aims to enlarge the receptive field of the designed model. As shown in figure 2.1a, the inception module is a structure which contains various sizes of convolution kernels in a parallel way. The effective design can increase the depth and width of the network while keeping the computational budget constant. Figure 2.1b shows the residual module which contains a shortcut structure across convolutional layers in the network. The module can more effectively transmit weight through the model in the back-propagation process. ResNet realises the potential of training networks with thousands of layers: in 2016, a 1001-layer ResNet was produced and it outperformed its shallower-version counterparts (He et al., 2016*b*). Figure 2.1c shows the densely connected module which also contains shortcut structures. Unlike the residual module, the densely connected module collectively aggregates former feature maps with the following ones. The dense connection can efficiently use the former information and thus requires fewer feature maps.

2.2.2 A Review of Model Architectures

Following FCN, a series of encoder-decoder networks is proposed such as SegNet (Badrinarayanan et al., 2017), U-Net (Ronneberger et al., 2015) and DeconvNet (Noh et al., 2015). These models have a similar U-shape architecture. SegNet and DeconvNet use the unpooling operation which is the reverse operation of the max-pooling operation to recover the spatial information, but U-Net uses the concatenation operation instead. The unpooling operation comes from Zeiler and Fergus (2014). This operation records the locations of the maxima within each max-pooling region to obtain an approximate inverse of the pooling operation. The concatenation operation directly concatenates feature maps from shallow layers with corresponding ones in deep layers in the channel direction.

Table 2.1: *An incomplete list of the U-Net family models. It is noted that the table is only for the demonstration of the adaptability of the U-Net architecture. Therefore, some models such as the ResU-Net which have different implementations with the same name are not included here.*

U-Net variants	R2U-Net (Alom et al., 2018)	Attention U-Net (Oktay et al., 2018)	Deformable U-Net (Jin et al., 2019)	ResU-Net (Chu et al., 2019)	DENSE-INception U-Net (Zhang et al., 2020)	RAU-Net (Chen et al., 2020)
Description	Combining the recurrent convolutional neural network (RCNN) with the U-Net model and identifying the efficacy of the recurrent structure in the medical segmentation task.	Embedding the attention mechanism into the U-Net; the proposed attention gate module contributes to small object predictions.	Using the deformable convolution which is designed for objects with irregular shapes to replace the normal convolution in the original U-Net.	Using ResNet to replace the encoder network of the original U-Net, this work also utilises a conditional random field (CRF) model and the morphological operation as the post-processing.	A U-shape model combining multiple techniques including the dense connection, the inception module and the residual module.	Embedding the ResNext block and locality sensitive hashing attention module (LSH). The LSH is a spatial attention module which enables the model to learn positions of important areas.

The models with the encoder-decoder architecture can easily be adjusted to improve the model performance by embedding different backbone networks or modules such as the feature pyramid (Chen et al., 2018; Diakogiannis et al., 2020) and the attention (Oktay et al., 2018; Ni et al., 2019) modules. For example, the U-shape architecture of the original U-Net is expanded to a series of models. Table 2.1 shows an incomplete collection of models with the U-Net architecture and their brief introductions.

The U-shape architecture is widely adopted in semantic segmentation models. However, the encoder-decoder architecture also has other structures including the hierarchical shape structure such as the RefineNet (Lin, Milan, Shen and Reid, 2017), and the global convolutional network (GCN) (Peng et al., 2017). The RefineNet exploits details from various stages of convolutions and fuses them together to produce a high-resolution prediction. The GCN exploits feature maps after every pooling layer. The feature maps are fed into the global convolutional module which contains two branches: each branch has the integration of larger kernels with a size of $1 \times k$ and $k \times 1$ (k is the kernel size), instead of using small-size square ones, like 3×3. The branch is designed to exploit the benefit of large-size kernels which have a larger receptive field and without introducing heavier computation cost. The model also designs a boundary refinement unit to sharpen the output boundary other than using conventional conditional random field (CRF) method. The unsymmetrical convolution design is also explored in ACNet (Ding et al., 2019)

A feature pyramid is another architecture commonly used in semantic segmentation models. The feature pyramid itself is not a new thing in the image processing area (Adelson et al., 1984). The feature pyramid module has a hierarchical structure for the purpose of multi-scale feature representation (Lin, Dollár, Girshick, He, Hariharan and Belongie, 2017). The semantic segmentation models utilising the feature pyramid module include the PSPNet (Zhao et al., 2017) and the DeepLab-v2 (Chen et al., 2017). The PSPNet introduces the spatial pyramid pooling module to tackle the issue of similar objects being confused by exploiting the contextual information. The pyramid module performs the pooling operation with different sizes of pooling kernels which distribute as a pyramid structure. Following that, the upsampling operation is applied to the feature maps from the pyramid modules before concatenating with the feature map from the last convolution layer of the backbone network.

The attention mechanism itself is from the human visual systems (HVS). The HVS contains two types of attention: the top-down and the bottom-up (Treue, 2003). The top-down attention is controlled by the high cognition system and the bottom-up attention is how the visual system naturally responded to the environment, i.e. perception system. For example, when people look at an image without any purposes, the most conspicuous object is normally observed first. This kind of attention is the bottom-up attention mechanism. However, if people look at an image with a purpose beforehand such as guessing the age of buildings, the attention will be attracted by the building area of the image. The process is operated by human cognition system and is called the top-down attention mechanism. The two attentions are combined to create an integrated saliency map which flags regions of interest in the retina image.

The attention mechanisms are proposed for the CNN by simulating the human attention mechanisms described above. The CNN attention mechanism contains two different types: the hard attention and the soft attention. The hard attention mechanism imitates the top-down attention mechanism. The common hard attention mechanisms use the reinforcement learning to give the desired region credits (Mnih et al., 2014) or use the recurrent neural network Li et al. (2016). The soft attention is to simulate the bottom-up attention mechanism. The soft attention is to embed a carefully crafted attention module into the deep learning model and to learn the attention map through the training process (Woo et al., 2018; Fu et al., 2019).

The soft attention can impact on feature maps at the spatial-wise or the channel-wise or in both facets. Chen et al. (2020) applies a spatial-wise attention which is firstly introduced in the natural language prediction (NLP) task in (Vaswani et al., 2017). The mechanism is initially calculated with three variables: Q (Query), K (Key) and V (Values). The compatibility between the the Q and the K is calculated first and then multiplied by the V to yield the attention map. Vaswani et al. (2017) designed a spatial-wise attention module in which the compatibility is the dot product of the Q and the transpose of the K with a following softmax. In Chen et al. (2020), the Q, K and V are the same feature maps from a layer. The channel-wise attention learns weights impacted on channels of feature maps of a layer. An example is the SENet (Hu et al., 2018). The squeeze-excitation module achieves the receptive field of the whole feature map of each channel by using the global average pooling layer and then passing the outputs through a sigmoid layer to compute weights impacting on the initial feature maps.

Integrating the spatial and the channel attention mechanisms has become popular recently. Fu et al. (2019) designs the dual-attention network which combines spatial and channel attentions in a parallel way, then fuses the outcomes from the two branches. Woo et al. (2018) cascades the channel and the spatial attentions instead of connecting them in the parallel way. Choi et al. (2020) designed a height-driven attention module which emphasises the informative features or classes selectively according to their vertical positions. Yu et al. (2020) designed a context prior attention module that aims to solve the classification confusion problem in classes with context dependencies.

2.2.3 Urban Datasets

Building an urban image-based dataset is very common and crucial to support the research of autonomous driving. Publicly-available examples include CamVid (Brostow et al., 2009), KITTI (Geiger et al., 2013), Cityscapes (Cordts et al., 2016), Mapillary Vistas (Neuhold et al., 2017), the ApolloScape (Huang et al., 2018) and the latest A2D2: Audi Autonomous Driving Dataset (Geyer et al., 2020). Although these datasets do not consider buildings at the component level and not particularly focus on residential buildings, they still provide valuable experience in developing a new dataset. CamVid and KITTI datasets are the two earliest urban datasets developed for the purpose of autonomous driving. Both of their data are collected using carefully designed vehicle-mounted rigs which ensure the efficiency of the data collection process. CamVid contains over 700 annotated images. The annotation quality is ensured by designated quality inspectors. The KITTI dataset contains 400 annotated frames in which 200 are for training and 200 for validation.

The Cityscapes dataset contains 25,000 images collected from over 50 different cities by a moving vehicle. Of those images, 5000 are fine labelled and the remaining 20,000 are coarsely labelled. A vehicle-mounted camera rig is used to collect these data. The dataset is split into train, validation and test sets based on pre-processed labels to maintain uniformity in each set. The Mapillary Vistas dataset was planned to reflect the diverse scenes of the urban environment at their greatest extent; the dataset does not rely on an ad-hoc data collection system to collect images as in the Cityscapes dataset. Instead, it collects images with a wide range type of rigs such as mobile phones and head- or car-mounted equipment such as a Garmin or GoPro action camera. The dataset also has a wider range of collection area than the Cityscapes dataset. The Mapillary Vistas dataset considers the scene changes in detail with regards to weather, season and even time of day. The size of the dataset is also considerably larger than that of Cityscapes reaching 25,000 fine-grained annotated images.

The ApolloScape dataset provides a comprehensive data source for autonomous driving. The dataset contains multiple types of data such as visual images, point clouds and depth maps. A vehicle-mounted data acquisition system is constructed for the data collection task and an efficient data labelling pipeline is developed. The data acquisition system contains a collection of sensors such as cameras, LiDAR units and a GNSS/IMU unit. The labelling pipeline is designed to be highly efficient by fusing the 3D annotation process with the 2D annotation process. The dataset evolves over time and the latest release contains 146,997 frames with annotations. Similar to the ApolloScape, the A2D2 dataset also employs a vehicle-mounted data acquisition system which is installed on a mid-size SUV. The dataset contains 41,280 annotated frames with corresponding depth images and camera poses as in the ApolloScape dataset. Compared to the rig designed for the ApolloScape dataset, the A2D2 acquisition rig contains more cameras and LiDAR units.

From CamVid to A2D2, the urban scene datasets for computer vision tasks have progressed rapidly in the past decade, owing to the development of different sensors and the needs of autonomous driving. Generally, the early-built datasets such as the KITTI and the Cityscapes contain less data and sensors in their acquisition system than the later ones. At the side of the data acquisition method, most of the datasets employ a vehicle-mounted collection platform but the Mapillary Vistas utilises a wide range of cheaper mobile sensors. The ApolloScape dataset developed a novel data annotation pipeline which fuses the 3D labelling and the 2D labelling which greatly reduces the cost of the dataset annotation.

2.2.4 Section Conclusion

The convolutional neural network (CNN) technique has contributed significantly towards the development of the semantic segmentation community. The state-of-the-art CNN-based semantic segmentation approaches are built on the encoder-decoder architecture. Various techniques have been developed to improve the performance of semantic segmentation models by designing new structures, e.g. enlarging receptive field, utilising class dependencies, etc. Efforts have been made on both macro and micro aspects of model structures: the U-shape structure, feature pyramid and attention mechanism are three macro developments of model structures, and dilated convolution and residual module are located at the micro side. Datasets have progressed significantly since the deep learning technique became popular. Of urban datasets, the dataset scale has progressed from a few hundred images to tens of thousands images with significantly more diverse scenes covered.

2.3 Building Facade Segmentation

2.3.1 Introduction

The history of building facade semantic segmentation can be traced back to the 1970s (Ohta et al., 1978). The paper develops a region growing-based method of segmenting buildings and windows from a scene image. In the past decade, the topic of facade segmentation has attracted a lot of attention and is still an active research area. This review focuses on facade segmentation works from the past decade, i.e. 2012–2022.

The literature review is divided into two sections: the first section reviews all publicly available facade segmentation datasets; the second reviews developed approaches and their performance. The collection of facade segmentation-related papers is achieved using the Google Scholar database. Key words ‘facade segmentation’ and ‘facade parsing’ are used and irrelevant papers are manually filtered out of the search results. Then, the citation history of each facade segmentation dataset is tracked to find potential missing documents. In particular, Kelly et al. (2017) do not have key words in their title but had still developed a method for facade segmentation. Apart from (Femiani et al., 2018; Wang et al., 2022), non-peer reviewed documents are not included. The reason these two documents are included is specifically stated as such: Femiani et al. (2018) is a frequently referenced literature in the facade segmentation community, Wang et al. (2022) has not been published but have uploaded their dataset online which is then used to benchmark in this thesis. There are works aiming to recognise windows such as (Neuhausen and König, 2018). These works are not included as these works are not for recognising the whole facade.

2.3.2 Publicly Available Datasets

From the literature and using Google search service, nine publicly available facade segmentation datasets are found. Descriptions of these datasets are listed below.

1. **Oxford RobotCar-Facade (Wang et al., 2022):** The Oxford RobotCar-Facade dataset is the latest publicly available dataset with labels of building components. The dataset contains 500 annotated images, 400 of which are for training and the remaining 100 are for validation. The raw data is collected by a vehicle-mounted camera which is forward facing. Therefore, most of the images in this dataset show a street-canyon view. Buildings are separated into five classes which are wall, window, door, balcony and shop. The dataset is based on Oxford, in the UK.

2. **TMBuD (Orhei et al., 2021):** There are 1120 images in total in this dataset and 300 of them are annotated. Smartphone is considered as the input sensor device and the image size is determined to fit the generic smartphone's filming size. The dataset is designed to contribute to the augmented reality domain. Each building in this dataset is captured from several perspectives. Architectural styles vary distinctively in this dataset. For example, Gothic, Byzantine and modern styles are all identified in this dataset. The labelling rule is set to partition an image into eight classes. Five of them are irrelevant to buildings: background, sky, vegetation, ground and noise. It is noted that the noise class is considered for temporary objects such as people and cars. Buildings are divided into three classes: window and door, the remaining area of a building is labelled as building class.

3. **ECP (Teboul et al., 2010):** This dataset contains 104 images of Haussmannian style buildings. Raw images are manually rectified and cropped. This dataset partitions a building into six categories which are wall, balcony, roof, shop, door and window. The dataset was not divided for training and validation beforehand.

4. **Art Deco (Gadde et al., 2016):** This dataset contains 79 annotated images of Art Deco architectural style buildings. Most of the backgrounds are cropped and images are rectified. The dataset is designed as a supplement of the ECP dataset to demonstrate that encoding specific architectural rules into the segmentation algorithm will lead to a better result. The annotation rule is the same as the ECP dataset. Occlusions are presented in this dataset with hand-annotated ground-truth for the labels behind the vegetation.

5. **RueMonge 2014 (Riemenschneider et al., 2014):** This dataset contains 428 images and 219 of those are annotated. The 428 images of 60 Haussmannian buildings were taken continuously along a 700-metre road in Rue Monge, Paris. Since the dataset

is captured continuously, it contains several images of the same building from slightly different perspectives. In this case, overlapping can be frequently observed. The training and validation data are pre-specified in this dataset with a rough ratio 1:1.

6. **CMP (Tyleček and Šára, 2013):** This dataset contains 606 annotated images with various architectural styles. The raw data is collected in two different ways. 312 images are collected by authors of the dataset and the rest are collected using unlabelled data of ECP and ZuBuD (Shao et al., 2003) databases.

7. **Graz50 (Riemenschneider et al., 2012):** According to the literature, the Graz50 dataset contains 50 annotated images, the majority of which are Gruenderzeit architecture style which is common in Germany and Austria. However, the link to Graz50 dataset was invalid at the time of writing. The dataset only contains four classes: window, door, wall and sky.

8. **LabelMeFacade (Fröhlich et al., 2010):** The raw images of LabelMeFacade are a subset of LabelMe database (Russell et al., 2008). LabelMeFacade contains 945 annotated images of substantially different types of architectural styles. The 945 images are partitioned into 100 images for training and 845 images for validation. The division ratio is rare. The author of this thesis has identified different types of buildings, not limited to: skyscrapers, modern non-commercial buildings, neoclassical buildings, Baroque buildings, Gothic churches, etc. Unlike other datasets, excluding the latest TMBuD dataset, the dataset also contains images taken at night. The views of this dataset also vary distinctively: a substantial number of images show street-canyon views, a portion of images show the top part of tall buildings, etc.

9. **eTRIMS (Korč and Förstner, 2009):** eTRIMS dataset is the earliest dataset for the facade segmentation task. The dataset has 60 annotated images. The dataset has two versions, the four-class version does not contain building subset items, the eight-class version contains two extra classes: window and door.

Table2.2 shows a brief summary of these datasets. The number of annotated images and image sizes are determined by downloading corresponding datasets to find the latest version. Therefore, the values may vary compared to their introduction documents. For example, the TMBuD paper introduces 160 images but the authors have added a further 140 images to their database since the dataset was announced. Furthermore, the authors of CMP dataset claim that a portion of their raw data has 6MP effective pixels. However, in the published version, all images are down-sampled to below 1MP. The RueMonge 2014 dataset claims it has 428 images in the dataset but only 219 of them are annotated. In the 'Effective pixels' column, the term 'fixed' is used when all images in the dataset have the same size, the term 'varied' is used to indicate that the sizes of images vary in both pixel quantities and height and weight, the term 'regular' indicates that the height and weight sizes vary in this dataset but pixel quantities remain approximately the same.

Table 2.2: *This table demonstrates an overview of publicly available building facade segmentation datasets. This study includes names of these datasets, numbers of annotated images, effective pixels and locations of data capture. The last column of the table records the image style. The table is arranged by order of year published. It is noted that the URL of the Graz50 dataset is now invalid and thus the size information cannot be provided.*

Name (known as)	Citation	Number of Annotated Image	Effective Pixels	Location of the Data Capture	Image Style
Oxford RobotCar Facade	Wang et al. (2022)	500	Fixed, 1280 × 960 = 1.2MPx	Oxford, UK	driving-view, multi-storey and two-storey residential houses
Timisoara Building (TMBuD)	Orhei et al. (2021)	300	Fixed, 768 × 1024 = 0.8MPx	Timisoara, Romania	street view, mixed architectural styles and uses
Paris ArtDeco	Gadde et al. (2016)	79	Varied, min≈ 0.2MPx max≈ 0.5MPx	Paris, France	cropped, multi-storey buildings
ETHZ CVL RueMonge2014 (VarCity)	Riemenschneider et al. (2014)	219	Fixed, 1067 × 800 = 0.9MPx	Paris, France	cropped, multi-storey buildings
CMP Facade	Tyleček and Šára (2013)	606	Varied, min≈ 0.1MPx max≈ 1.0MPx	Various cities across Europe and USA	cropped, multi-storey buildings
ICG Graz50	Riemenschneider et al. (2012)	50	N/A	Various European cities	cropped, mixed architectural styles
Ecole Centrale Paris Facades (ECP)	Teboul et al. (2010)	104	Varied, min≈ 0.1MPx max≈ 0.3MPx	Paris, France	cropped, multi-storey buildings
LabelMeFacade	Fröhlich et al. (2010)	945	Regular, ≈ 0.3MPx	Various cities across the world	various image styles, mixed architectural styles and uses
eTRIMS	Korč and Förstner (2009)	60	Regular, ≈ 0.4MPx	Various European cities	cropped, multi-storey and two-storey residential houses

The earliest dataset, eTRIMS was published over ten years ago and the latest ones, Oxford RobotCar and TMBuD were published this year (2022) and last year (2021), respectively. A six-year time gap is identified before TMBuD was constructed for which no new facade segmentation datasets are publicly available. Three of these datasets have fewer than one-hundred annotated images and the ECP dataset has just about one-hundred image-mask pairs. Paris is the most frequent location adopted for raw data captures. Four of the nine datasets are rectified and cropped in pre-processing. The pre-processing removes the majority of the background and means images only have frontal-parallel views of buildings. The image size of current publicly available facade segmentation datasets limits to c.one mega pixel and four of the eight (excluding Graz50) datasets have effective pixels lower than half a mega pixel.

The annotation strategy varies significantly across the nine datasets. Table 2.3 summarises how these datasets are labelled. These datasets are labelled in two different ways: the first is the pixel-wise which aims to use polygons to follow the boundary of each object as accurately as possible, the other is the region-wise which uses bounding boxes to cover objects to be annotated. The pixel-wise annotation is more accurate in theory than the region-wise approach, especially in non-rectangle objects such as chimneys. However, the author has noticed many labelling errors in LabelMeFacade dataset, thus its annotation accuracy might still be lower than a region-wise annotated dataset.

Table 2.3 shows that five of the eight datasets are labelled at the pixel-wise style. Window and door are two universal categories in these datasets. Depending on whether other objects belonging to the building are labelled separately, the wall category is determined. The CMP dataset has the finest labelling rule which divides a facade into twelve categories. The CMP dataset also separately annotates windows with glass appearance and with coverage into two different categories: window and blind. In addition, there are various strategies for treating occluded objects. eTRIMS, LabelMeFacade and TMBuD individually annotate occluded objects into separate categories.

Figure 2.2 shows visualised examples of the eight available datasets in the published year order. ECP, CMP and ArtDeco are the three region-wise annotated datasets. Their examples show that all elements in these datasets are annotated as bounding boxes and, therefore, sacrifice annotation accuracy to some extent. Errors are observed in LabelMeFacade such as unlabelled elements and occlusions. Oxford dataset has a unique street-canyon view among these datasets.

Table 2.3: *The table demonstrates labelling correlations of each dataset. Labelling accuracy indicates which methods a dataset has chosen. Labelling categories are divided into three categories: the first is building which contains all objects belonging to a building, the occlusion is objects which partially cover buildings, background is other objects appearing in an image.*

Dataset / Division	eTRIMS	ECP	ArtDeco	CMP	RueMonge2014	LabelMeFacade	TMBuD	Oxford RobotCar
Labelling Accuracy	pixel-wise	region-wise	region-wise	region-wise	pixel-wise	pixel-wise	pixel-wise	pixel-wise
Building	building window door	wall window door roof balcony shop chimney	wall window door roof balcony shop	facade window door blind pillar molding cornice pillar sill balcony shop deco	wall window door roof balcony shop	building window door	building window door	facade window door balcony shop
Occlusion	car vegetation	-	-	-	-	car vegetation	noise vegetation	-
Background	pavement road sky	sky		background	background	pavement road sky various	ground sky	background

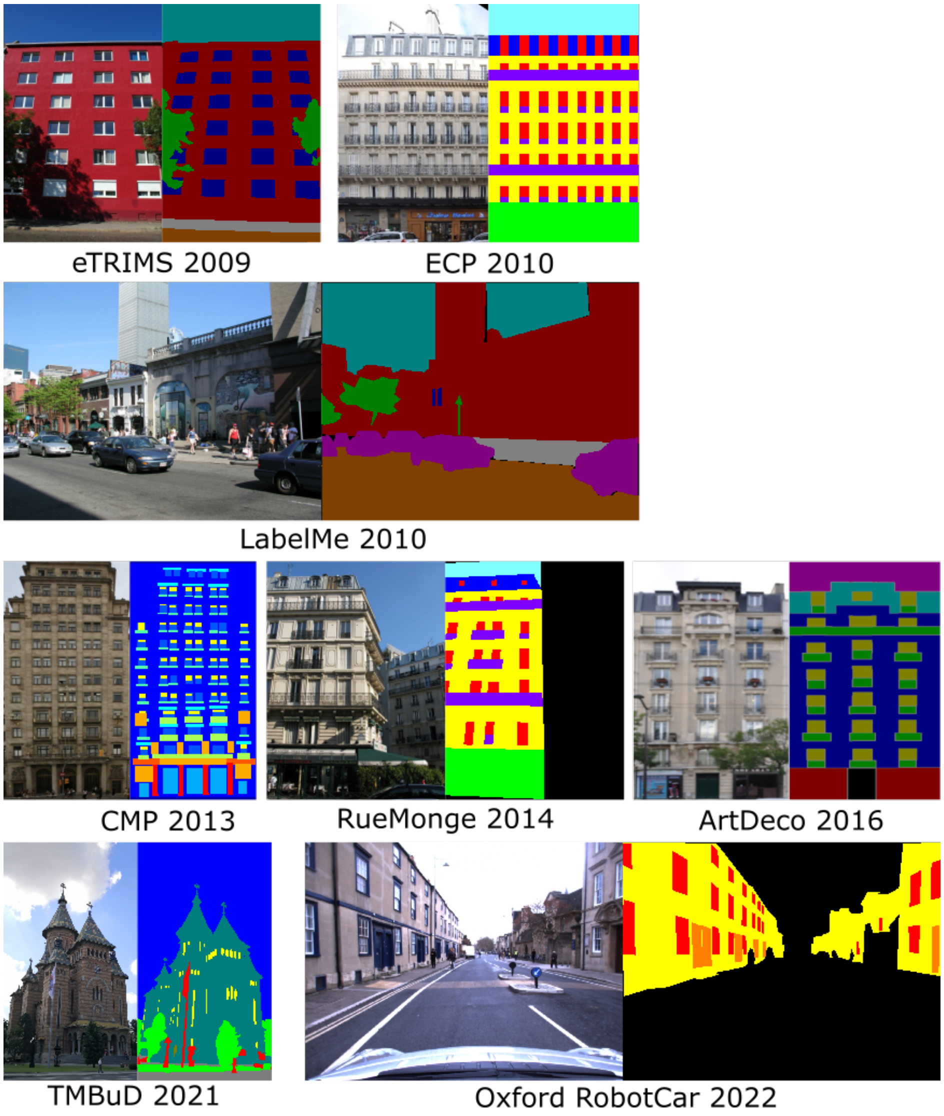


Figure 2.2: *Visualised examples of publicly available datasets in age order. Grammar parsing-oriented annotation is the mainstream dataset style (four of the seven including Graz50, three of the six without Graz50) in the facade segmentation community. ECP and ArtDeco datasets show a more concise segmentation than CMP. RueMonge 2014 is developed for 3D reconstruction as well, therefore, it only labels buildings on the same street. eTRIMS is a dataset sitting in between street view and rectified styles. The recently published facade datasets are all street view styles.*

The use frequency of these datasets is counted manually based on the collected papers. Among the nine datasets, ECP is the most frequently-used dataset: it is used twenty-six times. The second is eTRIMS which is used eleven times. The least-used is LabelMeFacade which is only used five times and four of them are before 2017. Figure 2.3 shows the statistics.

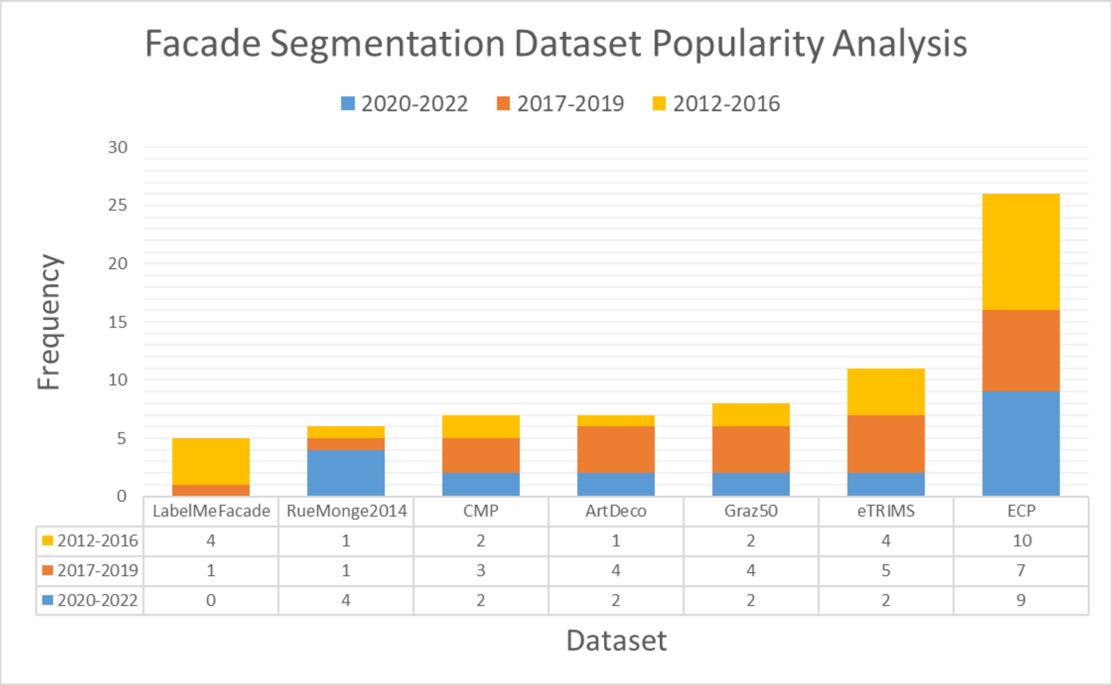


Figure 2.3: *Popularity analysis of dataset usages based on collected facade segmentation papers. The analysis is divided into three time intervals: 2012-2016, 2017-2019 and 2020-2022. 2016 is the the first year the deep learning technique was introduced to the facade segmentation community. In 2020 the deep learning technique became the dominant technique in this community.*

2.3.3 Approaches

2.3.3.1 Overview

In previous facade segmentation literature reviews, the approaches developed for facade segmentation are commonly categorised into two different groups. The two groups are sometimes named 'top-down' and 'bottom-up' as in (Rahmani and Mayer, 2018), alternatively, they are considered to be 'conventional' and 'deep learning-based' as in (Liu et al., 2022). The definitions of these two groups vary accordingly. The 'top-down'–'bottom-up' classification categorises approaches using shape grammar as 'top-down' and other classification-based approaches as 'bottom-up'. The categorisation is visualised based on the characteristics of these two approaches. Grammar rule-based approaches firstly divide a facade into bigger parts and recursively split them into facade components. Bottom-up approaches classify each pixel or super-pixel into predetermined groups. The 'top-down'–'bottom-up' classification has not been used since 2019. More and more literature categorises facade segmentation approaches into deep learning and non-deep learning categories as in the general image segmentation community. However, the author of this thesis argues that if the machine learning-based classification approaches are categorised into a separate group, the categorisation rule will be more clear to demonstrate the characteristics of each type of approach. Therefore, in this literature review, facade segmentation approaches are classified into three groups: grammar parsing-based, machine learning-based classification and deep learning-based classification approaches.

The three groups of facade segmentation approaches have shown an 'evolution' characteristic. Grammar-based approaches were popular before 2015, and the latest grammar-based approach was published in 2016 (Gadde et al., 2016). From 2015 to 2018, machine learning-based classification approaches became the state-of-the-art. After 2018, deep learning-based approaches dominated the facade segmentation research area. As a research branch of the general image semantic segmentation research, the development of facade segmentation has shown a clear coupling relationship with the development of image semantic segmentation, while with a small time lapse. For example, 2015 is the year that the breakthrough work, fully convolutional neural network, FCN, was announced (Long et al., 2015). The fully-convolutional deep learning technique was first introduced to the facade segmentation area in 2016 (Schmitz and Mayer, 2016). Furthermore, the use of transformer technology on semantic segmentation was studied in 2021 (Zheng et al., 2021) while the technology was first applied on facade segmentation the following year (Zhang et al., 2022). The next three parts review the facade segmentation papers of the past decade. Findings and conclusions are summarised in the 'Section Conclusion' as in Section 2.2.

2.3.3.2 Grammar Parsing-based Approaches

The concept of grammar is initially from the Natural Language Processing (NLP) community (D'Ulizia et al., 2011). The grammar types used widely in facade segmentation are the shape and split grammar which pre-define a set of rules as object shape and component relationship constraints. Then the defined rules are implemented to split facades recursively until no more rules can be applied. A tree-structure is common in representing defined grammars. The grammar rules can be general, e.g. using a rectangle shape to define all facade components (Gadde et al., 2016). However, more rules are very specific and highly reliant on architectural styles. A common architectural style which has been widely studied is the Haussmannian architecture in the ECP (Teboul et al., 2010) dataset. As examples of specific grammar rules, Teboul et al. (2011) defines rules such as 'balcony running across the whole width of a wall', 'no wall area between shop and door', 'roof window should be as high as the whole roof', etc. It is clear that these rules depend on specific architectural styles and rely upon rectified facade images.

In the past decade, most works using split grammar rules on facade semantic segmentation focus on how to optimise the facade-split procedure. The latest grammar-based facade segmentation approach is (Gadde et al., 2016). This work designs a method of using more generic rules to replace specific rules to avoid hand-crafted, expertise-involved grammar designs. The authors constructed the Paris Art Deco dataset to validate the capability of the developed approach on different styles of architecture. The generic grammar defines the building in axial direction, i.e. vertically and horizontally to separate floors and differentiate facade component relationships. Reinforced learning technique is adopted to optimise the grammar implementation procedure.

Koziński et al. (2014); Kozinski et al. (2015) designs a hierarchical representation of facade segmentation utilising the facade components' alignment and adjacency. The optimisation approach is developed based on linear programming and further progressed to use Markov random field (MRF). Martinovic and Van Gool (2013) utilise a Bayesian model to learn grammar rules from labels and then utilising reversible-jump Markov chain Monte Carlo (rjMCMC) method to generate predictions from raw images. Teboul et al. (2012) is the first to use reinforcement learning technique on parsing shape grammars. Riemenschneider et al. (2012) strengthen a facade as a group of irregular rectangular tiles which is defined by a split line across facades in both horizontal and vertical direction instead of patches. The design can reinforce the component position grammar unlike methods using patches. A dynamic programming (DP) model is adopted to optimise the label transition.

In summary, grammar parsing-based approaches utilise specific rules of the building facade to construct parsers. This is feasible because buildings are constructed based on explainable rules as man-made structures, e.g. component usually shows a rectangular appearance. However, every architectural style has its own distinct aesthetic features. The situation limits the generalisation capability of grammar-based approaches. Although the most recent work (Gadde et al., 2016) tried to use more generic grammar to replace architecture-sensitive grammar. However, in the cross-dataset analysis of this approach, trained parsers still generalise poorly across different datasets. Furthermore, grammar-based approaches are established on cropped-rectified fronto-parallel facade images. This prerequisite ensures all facade aesthetic rules can be maintained. However, if building facade images are captured from various perspectives, e.g. street view capture, grammar-based approaches are not applicable.

2.3.3.3 Pixel-level Classification-based Machine Learning Approaches

Grammar-based models can generate rectangular-shaped predictions which is beneficial for applications in the computer graphics community, e.g. constructing virtual building models for digital entertainment, especially under the Manhattan-world assumption. However, the objective of these approaches is not to generate pixel-level accuracy predictions rather, the priority is to maintain regular shapes of facades to validate the feasibility of designed grammar rules. Therefore, the datasets frequently used in validating grammar-based models, e.g. ECP and Art Deco, are not labelled at pixel-level accuracy as summarised in Table 2.3. Owing to machine learning-based classifiers developed in the general image semantic segmentation community such as random forest (Schroff et al., 2008), pixel-wise facade segmentation approaches have been developed over the past decade.

Using the variants of random forest (RF) for facade segmentation is popular in classification-based facade segmentations. Earlier approaches using RF-based classifiers in the past decade were established by Fröhlich et al. (2012, 2013). Fröhlich et al. (2012) designed a classifier called iterative context forests (ICF). This approach combines RF with the auto-context concept that is an incremental method which uses previous-level's classification results as features for its subsequent layer (Tu and Bai, 2009). ICF is designed to contain three stages, initial feature maps are colour features only, the following classification results from RF are iteratively stacked together. This method is the first piece of work using auto-context in facade segmentation. In the following year, the same authors developed a general image semantic segmentation approach which combines RF with Gaussian process (GP) (Fröhlich et al., 2013). This approach is then tested for facade segmentation in their paper.

Jampani *et al.* use auto-context and combine boosted decision trees to tackle the facade segmentation problem (Jampani et al., 2015; Gadde et al., 2017). In their works, the authors also design a similar three-stage auto-context architecture as in (Fröhlich et al., 2012). In comparison with Fröhlich et al. (2012), Jampani et al. (2015) and Gadde et al. (2017) use different features for both initialisation and auto-context stages and various learning techniques. These approaches achieve higher benchmark results on publicly available datasets than (Fröhlich et al., 2012). This shows that auto-context based approaches rely on carefully designed features and classifiers.

More recently, Rahmani *et al.* have utilised structured random forest (SRF) on the facade segmentation task (Rahmani et al., 2017; Rahmani and Mayer, 2018). Structured random forest was developed in 2011 (Kontschieder et al., 2011). In comparison with conventional RF, SRF considers the correlations of neighbouring pixels in classification. Therefore, this approach is more beneficial in capturing local context or structured information than a plain RF approach. Rahmani et al. (2017) introduced the SRF to the facade segmentation community in 2017; in their subsequent work, the authors designed a feature extraction pre-processing and fed the outcome feature maps to a SRF classifier (Rahmani and Mayer, 2018) instead of using raw images as the SRF input. A pre-trained regional proposal network (RPN) is adopted in the feature extraction process to extract window and door features. RPN is advantageous in maintaining the rectangular shape of objects.

Cohen *et al.* introduced the dynamic programming (DP) technique to the facade segmentation community Cohen et al. (2014, 2017). Dynamic programming is a greedy optimisation strategy which solves a problem as a number of simpler sub-problems. In Cohen et al. (2014), the problem 'facade segmentation' is divided into three sub-problems: 1. find a row of elements; 2. constrain co-occurrence elements, and; 3. shape optimisation. The architectural rules are hard-coded as in grammar parsing-based approaches. In the subsequent work, the authors extend their approach by further exploiting the regularity and symmetry features of architectures using SIFT features to better handle occlusions (Cohen et al., 2017).

Martinovic *et al.* propose a three-layer approach named ATLAS (Martinović et al., 2012; Mathias et al., 2016). In general, the bottom layer is coarse semantic segmentation. The middle layer is a detection layer for window and door followed by Markov Random Fields. The top layer applies weak architectural principles to refine the outcome from the middle layer. In the primary version of this approach (Martinović et al., 2012), recursive neural network (RNN) is adopted in the bottom layer as a pixel classifier and very specific architectural constraints are applied in the top layer, e.g. balcony running across the whole building on the second and fifth floors. In the final version of this approach (Mathias et al., 2016), pixel segmentation in the bottom layer is replaced by superpixel segmentation using multiclass support vector machine (SVM) and the second layer is strengthened.

The state-of-the-art machine learning-based pixel-wise classification approach (Rahmani and Mayer, 2018) has outperformed the grammar parsing-based approach in benchmark datasets. However, machine learning-based approaches still have some drawbacks such as the system could be very complex, e.g. the ATLAS model has multiple layers and each layer has multiple individual modules. A clear contribution of the state-of-the-art machine learning-based approaches compared to grammar-parsing based approaches is that SOTA approaches do not need to encode architectural hard constraints into models, which improves the model adaptability to various architectural styles. However,it must be noted that SOTA (Rahmani and Mayer, 2018) is not, strictly-speaking, a machine learning approach as it uses RPN technique which belongs to deep learning technology.

2.3.3.4 Deep Learning-based Approaches

As reviewed in Section 2.2, with the announcement of the FCN model (Long et al., 2015), 2015 was an important year for the image segmentation community. The first work to use the deep learning technique on facade segmentation was Brust et al. (2015). This is an early attempt at using convolution neural network (CNN) on facade semantic segmentation while it still uses fully connected layers for pixel-wise classification. The first work to introduce the fully convolutional technique to the facade segmentation community was Schmitz and Mayer (2016). This work uses weights from AlexNet (Krizhevsky et al., 2017) which is trained on ImageNet dataset (Deng et al., 2009) and uses a concatenation operation to combine low-level features with high level features. The results show that this work has achieved competitive results in comparison with the ATLAS model on the eTRIMS dataset.

In the following year, Kelly et al. (2017) explored facade segmentation based on SegNet (Badrinarayanan et al., 2017) and pre-trained Bayesian SegNet (Kendall et al., 2017). For a multi-class pixel-wise classification problem, a common output of using CNN is to predict probabilities of every class to every pixel. Then the label with the highest probability is assigned to the pixel. In this work, the authors propose a multi-output Bayesian neural network model based on the Bayesian SegNet, which is named at SegNet-Facade. The model creates output modules for each semantic class and every output layer contains four channels which are edge, negative, positive and unspecified. The authors of this article claim that the multi-output design is to achieve sharper features.

In their work, Femiani et al. (2018), further advanced the developed SegNet-Facade. It is noted that this work has not been published in a peer-review journal or conference yet although it is referenced multiple times, therefore, this review covers this article. Instead of using Bayesian neural network, the base model is changed to a normal SegNet with convolution layers. The authors have proposed two additional operations or modules on the basic multi-output SegNet. The first refinement is to use separable convolutions which were developed in MobileNet (Howard et al., 2017). The separable convolution separate a common 3×3 convolution kernel into two kernels which are 3×1 and 1×3. The design is used to reduce computational costs while Femiani *et al.* adopt the separable convolution to strengthen the straight-line features' extraction of a building. The second refinement is to add a recursive neural network (RNN) module after the output layers. This refinement is claimed to improve the performance of neighbouring elements' predictions. However, according to their experiments, the three different model architectures do not show distinct performance growth. In the lateral comparisons, F1 score rises when the separable module is implemented but a high decrease on precision metric is also observed.

The application of CNN techniques on facade segmentation began to gain popularity in 2020. Liu *et al.* utilise the symmetry characteristic of buildings (Liu et al., 2020, 2022); in their first paper, a novel reflective symmetry loss is proposed which aims to minimise the sum of coordinates variance of each object (Liu et al., 2020). A RPN module is also adopted as in Rahmani and Mayer (2018) to refine the output shape of objects. In their following work, the authors extend their symmetry loss to further exploit the transnational symmetry characteristic of facades (Liu et al., 2022). Their model architecture contains two branches. One is semantic segmentation branch to generate pixel-wise predictions and the other is instance detection branch which utilises an object detection CNN to produce bounding boxes around windows and doors. The outcomes from the two branches are refined by the translational symmetry loss which assumes objects of the same category would have the same dimensions.

Ma *et al.* have published a series of works on facade segmentation in the past three years (Ma, Ma, Xu and Zha, 2020; Ma, Xu, Ma and Zha, 2020; Ma et al., 2022). Ma, Ma, Xu and Zha (2020) developed a pyramid module to capture multi-scale context information. The separable convolution technique is also adopted in this work to enhance the extraction capability of straight-line features of facades. Ma, Xu, Ma and Zha (2020) aimed to use multi-view images to better handle occlusion problems. The designed model uses a series of facade images which are taken from multiple perspectives as the input. All these images pass through the same backbone network to extract features. The achieved attention maps are then fused together to enhance the target facade's representation. Ma et al. (2022) further explore how to handle the occlusion problem. This work designs a stage-wise feature learning strategy. The authors utilise a Bayesian CNN to produce an uncertainty map and progressively use a separable convolution-based module to enhance the prediction confidence on occlusions.

Zhang et al. (2022) introduce the latest transformer technique to the facade segmentation. Transformer, as the grammar parsing, is a technique initially from the natural language processing (NLP) community which was first introduced in the prominent paper 'Attention is all you need' (Vaswani et al., 2017) and then promoted to semantic segmentation in Zheng et al. (2021). This technique can better handle long-range relationships than the CNN architecture. Zhang et al. (2022) designed a complex model which combines PSPNet (Zhao et al., 2017) as a semantic segmentation branch and a transformer-architecture object detection model, and DETR (Carion et al., 2020) as an object detection branch. The model also uses the symmetry loss developed by Liu et al. (2020) as part of its training loss. Overall, the structure of the model developed in this work still follows the 'semantic segmentation for facade + object detection for elements' paradigm which was first developed by Rahmani and Mayer (2018).

Kong *et al.* and Tao *et al.* propose that pixel-wise facade parsing may not be the ideal form of representing facade decomposition (Kong and Fan, 2020; Tao et al., 2022). In their views, detection-based representations are better at representing a building decomposition. Kong and Fan (2020) use a YOLOv3 (Redmon and Farhadi, 2018) object detection model to recognise facade elements and use a PSPNet to identify facades. The final representation of a facade decomposition is a hybrid form of pixel-wise mask with bounding box. Tao et al. (2022) use bounding boxes only to represent a facade. The authors design a self-attention module encoding the facade layout regularity features to enhance the model performance.

In summary, in the area of facade segmentation, deep learning-based approaches have outperformed conventional machine learning-based approaches in terms of accuracy. In the past three years, all works in facade segmentation are based on deep learning techniques. How to exploit facade prior knowledge, such as symmetry and straight-line features, to refine segmentation results is still the focus in the facade segmentation community. A common architecture is to combine a semantic segmentation model with an object detection model. More recently, works aiming to explicitly encode facade priors into deep learning models have started to attract attention.

2.3.4 Section Conclusion

The research topic, facade segmentation, has experienced three stages in the past decade: grammar parsing, machine learning-based classification and deep learning-based classification. Deep learning-based classification approaches are the current SOTA. Most of the approaches developed for recognising facade components are based on cropped-rectified facade images. These images show a frontal-parallel view of a facade and most of the environmental information is cropped. In total nine publicly available datasets for facade segmentation were found; seven of them were announced over seven years ago and two of them were announced in the last two years.

2.4 Chapter Discussion and Conclusions

2.4.1 Chapter Discussion

The number of publicly available datasets for facade segmentation is still a barrier in this area. Before the time of writing, only seven publicly available facade segmentation datasets were announced. Four of them contain approxaimately or fewer than one hundred images. The dataset with the largest volume, LabelMeFacade, is rarely used in the facade segmentation community. Some literature complains about the quality of this dataset (Wang et al., 2022) and the candidate has also found many obvious annotation imperfections. In comparison with other publicly available built-environment datasets, facade segmentation datasets are significantly behind in their volumes, diverse scenes, quality and coverage.

Moreover, the drivers of facade segmentation research are considered limited. Most of the approaches developed for facade segmentation are based on rectified images which show a frontal-parallel view of facades. This image pre-processing paradigm is to ensure man-made patterns can be preserved, which is beneficial to computer graphics applications, especially for procedural modelling. However, as discussed in Section 1.2.1, other applications such as understanding built-environment require images to be built from multiple perspectives and preserve environmental information. The limited driver makes the definition of facade segmentation ambiguous. A definition of facade segmentation is given by Koziński et al. (2014) which claims facade segmentation is to use rectified images to segment building elements. The definition is very narrow which eliminates works using street view images out of the facade segmentation domain. This could further affect the later researchers' perspectives.

CNN-based technique is a powerful enabler to facade segmentation research. Since 2019, CNN has become the dominant technique used in this area. The SOTA approaches of facade segmentation are designed to encode man-made structure priors into their models. However, if street view images are used for facade segmentation, many architectural priors are no longer valid e.g. straight-line appearance and component alignment. Prior to this literature review, only Wang et al. (2022) had adopted built-environment features in driving-environment images to tackle facade segmentation in their datasets, however, their paper has not been peer-reviewed yet. Furthermore, SOTA approaches using CNN techniques commonly use FCN and PSPNet as their base segmentation models, therefore, whether or not other powerful architectures of CNN such as U-Net and DeepLab series can show better performance in facade segmentation still lacks exploration.

Urban datasets reviewed in section 2.2.3 could inspire the development of the urban residential building facade segmentation dataset in many ways. For example, in the data capture process, the impact of environmental variables such as weather and season need considering. The vehicle-mounted rig is the main-stream method of capturing the urban environmental data which provides an efficient solution of data capture and is naturally the method of collecting data in autonomous driving. The annotators need to be carefully selected as professional annotators and amateurs would make different decisions. The state-of-the-art urban datasets commonly contain over 10,000 images with annotations to cover as many diverse scenes as possible.

2.4.2 Chapter Conclusions

In this chapter, the research area facade segmentation is comprehensively reviewed including the publicly available datasets and developed approaches. General-purpose CNN-based semantic segmentation models are also reviewed. In summary, the research gap between the state-of-the-art of facade segmentation and the target of this thesis can be identified as:

1. There are currently no facade segmentation datasets focusing on English houses except for the Oxford RobotCar. However, this dataset was constructed after this thesis was written. In order to assist environmental understanding, a high-resolution, multi-perspective, English house-based dataset is essential.

2. Previously used facade decomposition schemes for facade understanding vary. Different application scenarios might be the major reason for this diversity. Among the eight publicly available facade segmentation datasets, eTRIMS, ECP, ArtDeco and CMP are for procedural modelling in the computer graphics area. TMBuD dataset is for Artificial Reality development and Oxford RobotCar Facade focuses on autonomous driving-style data. Therefore, to fit the building retrofit need, a new definition of facade decomposition is needed.

3. Facade segmentation using street view images can be tackled using a general-purpose semantic segmentation model. However, as built-environment is a man-made scene, whether or not other prior knowledge or dataset features can be adopted to improve model performance also needs to be explored.

Chapter 3

Data Collection and Annotation

3.1 Introduction

Datasets are the foundation of any data-driven predictive task. The dataset quality and quantity have a crucial impact on the performance of an applied data-driven predictive model. The diversity and richness of the dataset define how closely the dataset can represent the real-world situation. Building a dataset can be a time-consuming and high labour/cost piece of work, especially for high-level computer vision tasks such as supervised semantic and instance segmentation. Apart from the universal data collection and cleaning process, this type of task requires a labelling process which provides human understanding and knowledge to the dataset. A selected model can be trained on the dataset to learn the knowledge which is the mapping relation function $Y = f_{(X)}$ of the input data X and output Y from the labelled dataset.

Although different public datasets of the urban environment have become more widely available since the ImageNet project (Deng et al., 2009), as reviewed in section 2.3.2 and summarised in section 2.4.2, there currently lacks an English house-based facade segmentation dataset. Therefore, collecting adequate required data becomes critical. As introduced in section 1.2.2, a multi-spectral urban data collection platform has been previously built to substantially improve the automation level of the process of collecting the urban data. While using the urban data collection platform, a substantial quantity of data can be collected in a highly efficient manner, however, the collected image data still needs to be labelled for training a deep learning model to characterise building components.

In summary, the overall contribution of this chapter is the development of an urban residential building facade dataset annotation protocol, and a dataset based on this protocol isconstructed. The developed annotation protocol is further used to construct another facade segmentation dataset. The more detailed contributions are listed below:

1. produce a labelling framework of annotating residential building facades and their components from urban-style images;
2. construct a dataset based on the protocol with professional annotators involved aiming for developing solutions to pixel-wise building component detection;
3. design and perform an annotation feasibility experiment to validate the designed protocol;
4. introduce a larger dataset with fewer trained annotators aiming for exploring a more efficient and cost-effective solution of a scalable facade annotation task.

3.2 Annotation Protocol

3.2.1 Category Selection

Recognising a building as an integrated object is not a subjective and complex task, however, recognising buildings as groups of individual components each with distinct semantic meanings is not straightforward. The first ambiguity is the decomposition of building facades. A building facade would contain multiple components, commonly including windows, doors, walls, etc. However, these components can also be decomposed further. For example, windows can be decomposed into frames and glass panels. For example, Mao *et al.* aim to only extract the glass area of windows from oblique aerial images (Mao et al., 2022) which needs to decompose a window further. Moreover, the wall areas sometimes contain pipes, decorated pillars, balconies, etc. Decomposing building facades into very fine levels will substantially increase the annotation cost and may not be necessary. Therefore, the definition of the decomposition needs to balance the needs of a task and the cost of the annotation strategy.

Referring to Table 2.3, the facade segmentation annotation rules defined in publicly available datasets vary in many ways. A conclusion was that datasets which were constructed to validate grammar-parsing models have finer decomposition than other datasets. However, the conclusion is broken by the latest Oxford RobotCar dataset (Wang et al., 2022). The coarsest decomposition only contains three classes which are building, window and door (Korč and Förstner, 2009; Fröhlich et al., 2010). The finest decomposition contains twelve classes in CMP dataset (Tyleček and Šára, 2013).

The other ambiguity is how to define components of a building such as windows and doors. The main obstacle is that such components have various distinctive appearances and sizes, for example, windows can be conveniently defined as a rectangular glass panel enclosed by structural frames. Then how to treat many other forms of windows, such as opened, highly distorted or occluded windows, becomes problematic. In CMP dataset, occluded areas of windows are annotated as a blind or cornice. The strategy defines window as a glass opening while it is still unknown why subsequent datasets did not follow this strategy. Particularly as the definition could reduce the difficulty of the window recognition task by helping a designed model focus on the simplest type of a window.

Deep learning models extract features to recognise targets, therefore, it will be stronger in processing objects with more consistent appearances. Apparently, compared to the proposed window definition, i.e. glass within a frame, ambiguous windows will lose many features such as shape or surface material. However, from a human's perspective, those ambiguous windows can still be recognised as windows with intuitive conjectures. As the mechanism behind why people can recognise objects in different states, and what features the deep learning models use for recognising these objects are still unknown, definitions of building components with ambiguous conditions may largely affect the model performance.

A simple test was conducted by asking five colleagues in the RISE group to recognise highly-distorted and occluded windows using the Google Street View (Anguelov et al., 2010) platform. These window objects could be highly-distorted, occluded, blinded or on other occasions where the objects are not in a normal window appearance. Opinions were collected from them as to whether an object can still be treated as a window and the reason why this alternative object can be recognised. The test can also help the author to answer two questions: 1. how a window might be recognised, whether it is recognised by its own features or by context information, and 2. to what extent an object can be regarded as it is still in its normal form. An agreement was made it was unclear if some unusual windows should be treated as normal windows. This is because if they are isolated from the environment, they cannot be recognised by the majority of the test participants. For example, with windows with a highly distorted view or that were occluded, test participants agreed that these objects could be recognised by the human cognition system's inference capability but cannot be treated as its normal form as it is quite subjective.

As examples presented to colleagues are still limited, the alternative class decision is further explored by manually building a 300-image dataset using Google Street View images. Only windows and doors are annotated in this dataset. More details about this dataset are available in Appendix A.

Table 3.1: *Category descriptors, with properties that can be inferred through the visible-light image semantic segmentation, as well as information that could be obtained by incorporating other multi-spectral data, such as LiDAR, thermography and hyperspectral data.*

Category	Wall	Roof	Window	Door	Chimney
Description	The continuous vertical structure encloses the building's interior area. Other walls used to divide an area of land are not included	The covering of a building in the horizontal plane support by a wall with all attached components such as soffit and rain gutter	The opening in a wall and roof with glazing coverings and frame, other similar objects such as doors and vehicle windows are not included	The movable barrier made of a panel which provides access to the inside of the building. Similarities such as vehicle doors and gates are not included	The architectural ventilation structure which conducts smoke and combustion gases up from a fire or furnace vertically, terminating at or above roof level
Directly Inferrable Properties	Total area of external building element; total building height; number of storeys; orientation	Roof pitch; total building height; roof surface area	Number of windows; number of storeys; partial room layout; window type; total glazing area	Occupancy; partial room layout	Quantity; chimney type
Inferrable with Multi-spectral Data	Thermal bridge; material; cavity type	Thermal bridge; material	Glazing type; thermal transmittance (u-value)	Material	Usage

Considering the house inspection needs introduced in section 1.1.1 and the outcome of the human-level recognition test stated above, a house is categorised into seven classes: window, alternative window, door, alternative door, chimney, roof and wall. A pseudo-class representing the 'background' categorises all features that do not belong to any of the other classes. Relevant objects in an image are all labelled regardless of whether they occur in the foreground or background. Choices on labelling rules were considered given desirable properties of a given feature; for example windows were considered with their frames. A full taxonomy of the categories, with descriptions and information inferable from their localisation is given in Table 3.1.

Occlusion is an inevitable feature in the urban data captured. Two strategies are employed, designed to annotate objects partially covered by two types of occlusions: solid and sparse. Solid occlusion occurs when objects such as signs and vehicles appear in front of objects. Solid occluding objects are considered as 'background' and are effectively ignored. Sparse occluding objects are those such as trees and railings. Unlike solid occlusion, objects occluded by sparse obstacles are still partially visible, but may not show any explicit structure. Labelling sparsely occluded features is a dilemma, as if we label these as background, a substantial quantity of information will be lost, and may detrimentally affect training. The trade off we make is that if any part of an object is not occluded, the area is still labelled with its corresponding category, otherwise it is ignored.

3.2.2 Annotation Pipeline

Constructing a large-scale dataset requires multiple decisions including the data scope and the annotation protocol. As the ultimate goal of this dataset is to provide a substantial understanding of the residential building facade, the annotation protocol must be efficient enough to facilitate the influx of new data to be manually labelled. In the previous section, the categorical divisions of a facade have been determined. However, a complete annotation protocol still requires an annotation software package and labelling principles.

The semantic segmentation annotation is based on drawing polygons around the target objects and then filling in pre-defined label codes. Numerous tools can be applied to the task. The tools which can be used range from the free Windows Paint and the commercial Adobe Photoshop software packages, to other off-the-shelf and online software such as Amazon Sage Maker, Hive and LabelMe. The core function of drawing polygons is the same amongst all of these tools. In the thesis, an open-source software package, LabelMe[1], is adopted and modified to fit the task's needs. The software package is free and easy to use. In Figure 3.1, the LabelMe operation window is demonstrated. Also, since it is open-source, changes are easy to be made for a specific task. The LabelMe can only process images in a singular manner and store label images in separate folders. Furthermore, the software cannot apply the same labelling colour codes in different images. However, to create a dataset, labels need to be in universal colour codes and stored in required folders. Therefore, the software package is edited first to fit the needs of creating the dataset. A post-processing function is developed to automate the image renaming and relocation process.

In developing the annotation protocol, images collected by MARVEL, the developed vehicle-mounted data capture platform introduced before are used. Google Street View images were also used for the early-stage pipeline design and test.

The annotation rule is demonstrated in Table 3.2. The rule is made based on the building facade component definitions in Table 3.1. In an annotation process, precisely following the

[1] available at: `https://github.com/wkentaro/labelme`

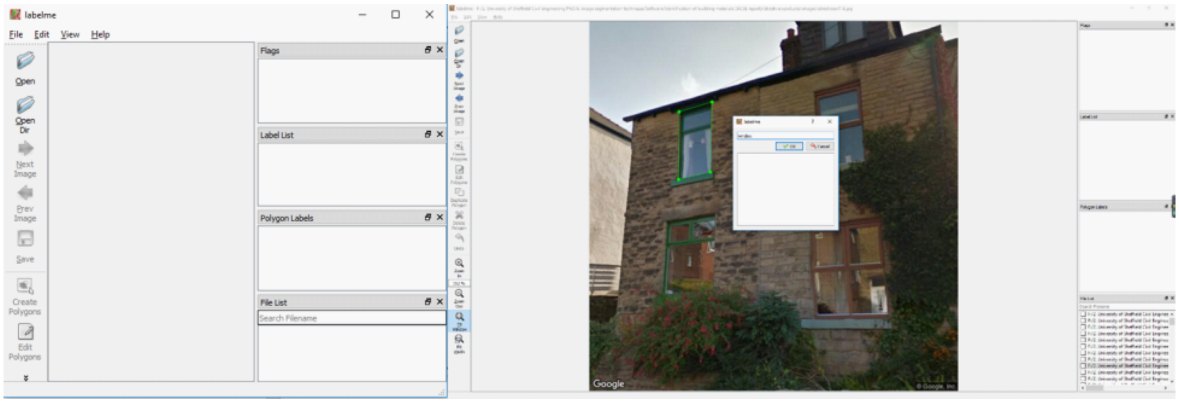


Figure 3.1: *The left-hand image is the initialised dialogue box of LabelMe, the right-hand side is the dialogue box after an image is loaded. Click the 'Open Dir' button to load the raw images folder. When images are loaded, click 'Create Polygons' to start to annotate by creating a polygon by following the boundary of the destination object.*

boundary of instance frames is very challenging because it is very time-consuming. However, a high-quality annotation will be very beneficial for a deep learning model to learn features as it reduces noise. Therefore, a trade-off between generating a high-quality annotation and minimising time expenses needs to be taken. Motivated by the requirement, a trade-off is taken by sacrificing some of the annotation accuracy by slightly over-covering the frame boundaries but avoiding missing parts of the frames. The reason for the accuracy sacrifice is for future thermal performance analysis and identifying the types of window glazing purposes.

The recommended labelling sequence is decided based on the normal building construction sequence:

1. Wall
2. Roof
3. Chimney
4. Window & alternative window
5. Door & alternative door

The reason for the proposed labelling sequence is to avoid potential boundary gaps between connected instances due to the degree of labelling precision defect. Among the first three classes, the previously annotated classes need to slightly exceed the boundary of the subsequently annotated class. For example, when the labelling activity is performed on a fresh image, a wall needs to be annotated first; it is performed by covering the whole wall area, the

Table 3.2: *Annotation rule descriptor sheet.*

Category	Annotation rule
Window&Alt-window	'Window' class is defined as an opening area on a wall comprised by glass and frame structure in an image. The 'window' class should only contain windows which have no external cover and where the whole frame is visible. Windows with a clear internal cover (e.g. curtains, blinds) are also classified in the 'window' class. All instances which are imperfect (e.g. partially visible, large shape distortion) will go to the 'alt-window' class.
Door&Alt-door	Similar to the window annotation rule, a door annotation should include the door frame and leaf. The 'door' class should only contain doors which are completely visible.
Chimney	Covering the whole visible area of a chimney. The annotation should start from the linkage between the chimney bottom and roof.
Roof	Pitched roof is the majority of roof types in the project. A pitched roof is usually comprised of multiple parts, e.g. rafter, joist, coverings, etc. Therefore, a comprehensive roof annotation should include all visible roof parts of a building. In addition, it is observed that building extensions are common in UK houses, the extension roofs should also be classified into the 'roof' class.
Wall	The wall class should only contain structural walls which are part of the building facade and should not contain other types of wall (e.g. boundary wall, fence, etc.). Also, stairs should not be included in a wall annotation. Some attached objects on the walls can be covered in a wall annotation (e.g. pipe, antenna, CCTV, etc.), but large objects should be avoided during the wall annotation (e.g. billboard, waste bin, etc.)

potential windows and doors (with both of the usual and alternative categories) objects on it. On the border of the wall and other instances except the five classes mentioned above, the wall annotation should slightly exceed the boundary itself. In addition, the other instance, e.g. a roof, should follow the border precisely to cover the part to which the wall annotation exceeds. Similarly, a roof annotation should slightly exceed the boundary of the chimney on it. For the last four classes, annotations are drawn by directly covering the corresponding area. Thus, the finished facade annotation should result in an integrated area. In addition, the wall or roof instances should be annotated individually as all visible wall areas of each building or a single roof regardless of potential overlapping with other instances in the same class.

Figure 3.2 demonstrates an example of the annotation sequence. Figure 3.2a shows that the annotation starts from a single wall and slightly exceeds the boundary of the top of the roof; Figure 3.2b shows the top roof annotation and follows the roof's boundary as precisely as possible and should exceed the chimney boundary slightly; Figure 3.2c-d shows the annotation of chimney and window classes; Figure 3.2e shows that if there are different buildings in an image, these buildings should be annotated individually; Figure 3.2f shows the finished annotation.

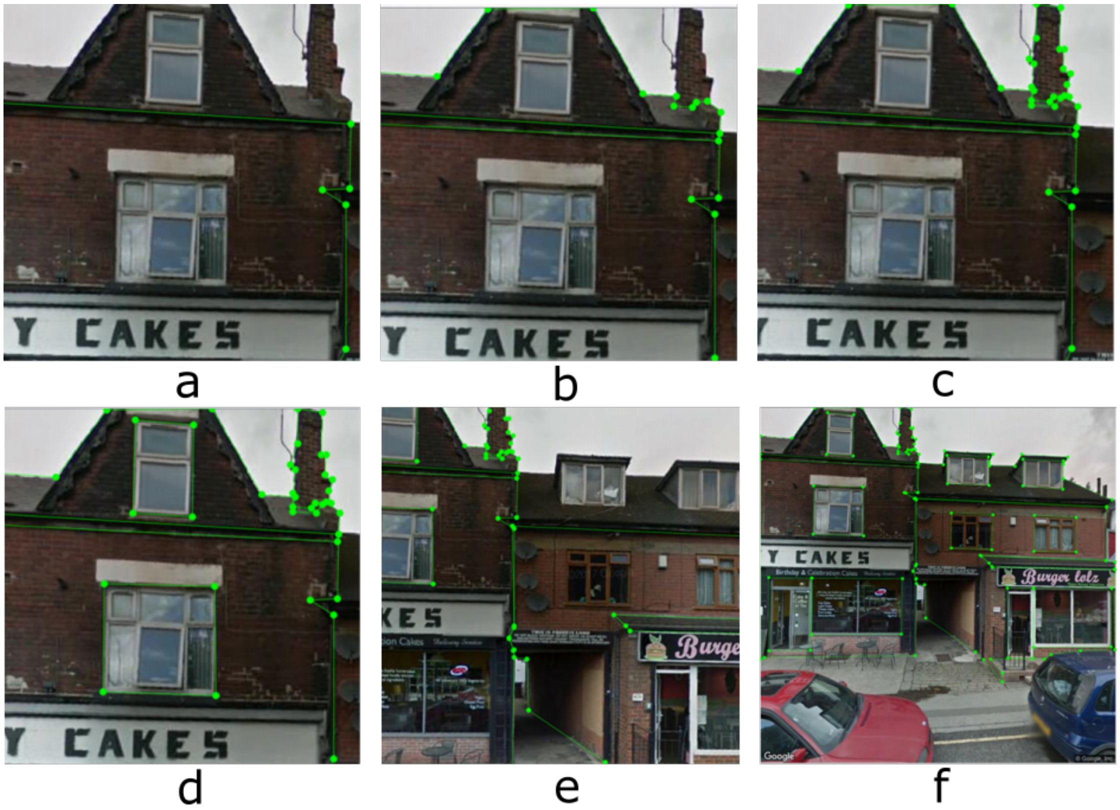


Figure 3.2: *The proposed annotation sequence. Annotation should start from bases, i.e. wall and roof, to attachments, i.e. window, door and chimney. Different buildings should be annotated individually and boundary should be annotated slightly overlapped. a. an annotation should start from an individual wall by following its boundary using polygons; b. then the roof should be annotated; c-d. chimney and window which are attachments are annotated further; e. annotation for the neighbouring house starts; f. a finished annotation figure of two houses.*

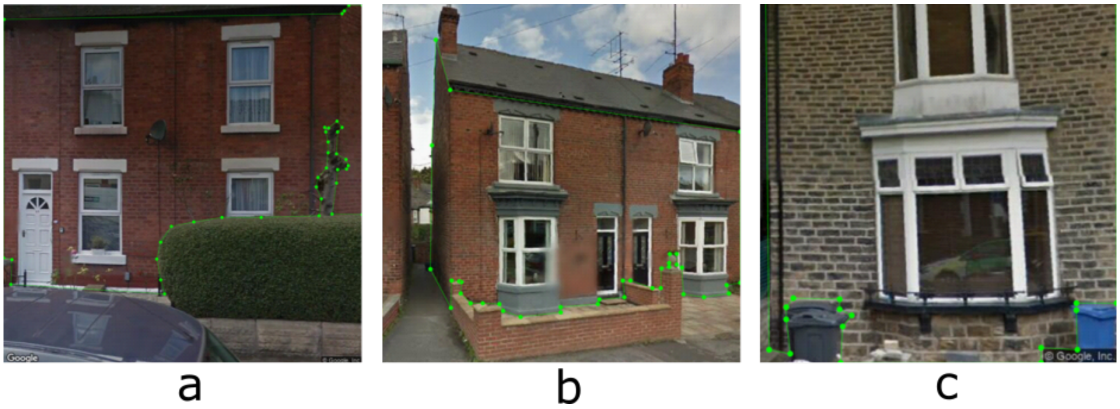


Figure 3.3: *Examples of wall annotation. a.& c. dense occlusions should be avoided in annotations, in these two cases, they are fence, tree trunk and waste bins; b. boundary walls should not be labelled.*

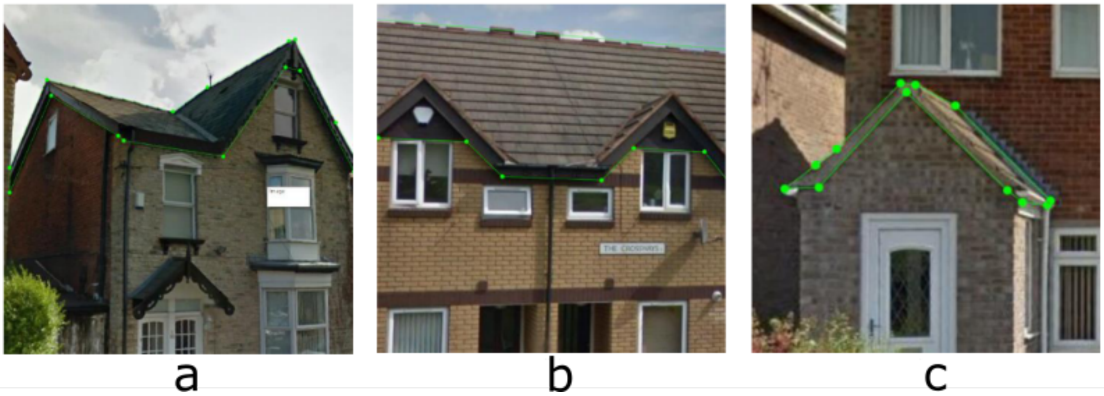


Figure 3.4: *Examples of roof annotation. a. all parts belonging to a roof including eaves and rakes should be included; b. parts with materials that are different from the wall should be annotated as roof; c. ground floor roof structure should be labelled.*

Figure 3.3 shows examples of wall annotation. Figure 3.3a shows that the antenna and pipes can be included in a wall annotation; large and dense objects, which are, in this image, a tree trunk and a fence, should not be included. Figure 3.3b shows boundary walls should not be labelled as a ‘wall’ annotation. Figure 3.3c shows waste bins should not be included in a wall annotation.

Figure 3.4 shows examples of roof annotation. Figure 3.4a shows a roof annotation should include all elements of a roof such as eaves, rakes, etc. Figure 3.4b shows if a gable is made from wood or other different materials to walls, it also should be included in the roof annotation. Figure 3.4c shows except the primary roof, the roof of a building extension should be labelled as the ‘roof’ class as well.

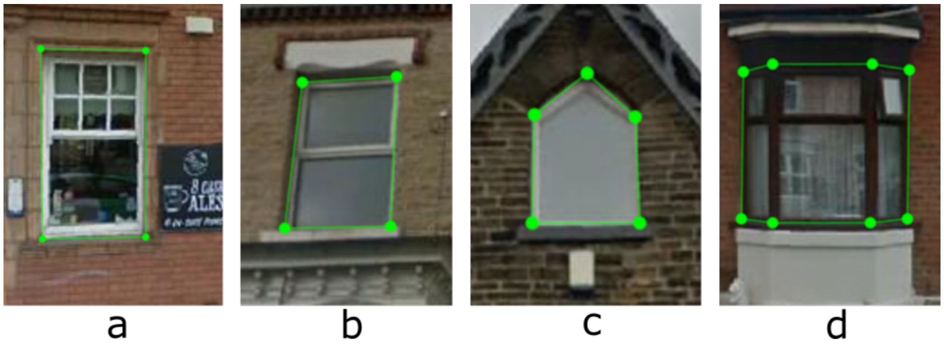


Figure 3.5: *Examples of typical window annotations: including window frame but not including window sill and support beam as well as making a trade-off by slightly over-covering the window frame. a. a typical window with internal blind; b. a slightly inclined window; c. a window belonging to a loft; d. a typical bay window.*

Figure 3.5 shows annotations of typical windows. In this project, building window is the only interest. Window annotation in this project needs to include glass and window frame but does not include windowsill or support beams and pillars if applicable.

Figure 3.6 shows annotations of alternative windows. Figure 3.6a-b are open casement and awning windows. In this situation, the philosophy is to try to generate an annotation to cover the opened window and the area which is not covered by the window but inside the window frame. Figure 3.6 c-e show occluded window examples. In these occasions, windows are partially covered by other objects. Annotations need to be made based on the area density of the covering object: if the object is fairly sparse (e.g. trees, bushes, handrails, etc.), which makes the window still partially visible behind the object, the mask should cover the area where the window is covered; if the object is very dense, the mask should avoid the area. Assumptions are essential to deciding the boundary of the window covered by sparse objects. Figure 3.6 f-h show highly distorted window annotation examples. If the window frame of a window is still visible, the instance should be annotated. Figure 3.6 i-j show examples of skylights.

Figure 3.7 shows annotation examples of bay and bow windows. A bay window has three faces which form a half hexagon. A bow window has multiple separate windows to form an approximately curved front face. The two types of windows are difficult to be classified into alternative or standard classes. The reason is the two window types have a three-dimensional structure in comparison with planar windows. Thus even if images of these types of windows are captured at a large viewing angle, there are still visible intact faces. Rules are made on

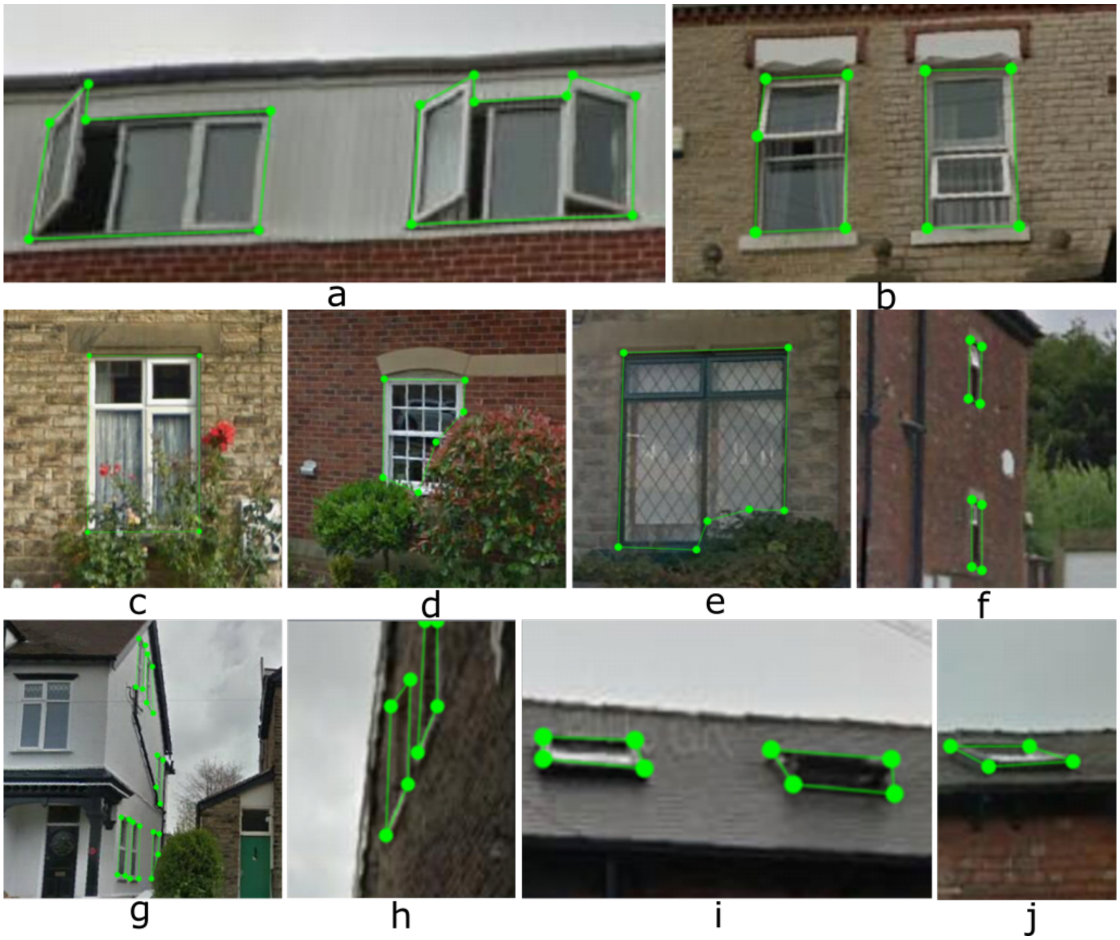


Figure 3.6: *Examples of alternative window annotations include opened, partially visible, highly-distorted and dormers. a-b opened window annotation should follow the boundary of window frame; c-e vegetation-occluded window annotation should depend on the density of occlusions; f-h highly distorted window should also be annotated; i-j dormer should be annotated as the alternative whether it is opened or closed.*

bay windows that only if the picture is taken roughly perpendicular to the front window, which means the majority of the area of the three sides of the bay window is visible, the bay window will be classified as ‘window’. Otherwise, it will be classified as ‘alternative window’. Looking at a bow window from different angles, in most of the cases, the majority of all separate windows are completely visible. In this situation, the bow window should be classified as a ‘window’ and if the viewing angle is so large that part of the window is hidden; this window should be put into the ‘alternative window’ class.

Another difficult decision is annotating windows around a door. For example, the window right above the door and within the door frame is usually called a ‘transom window’. How-

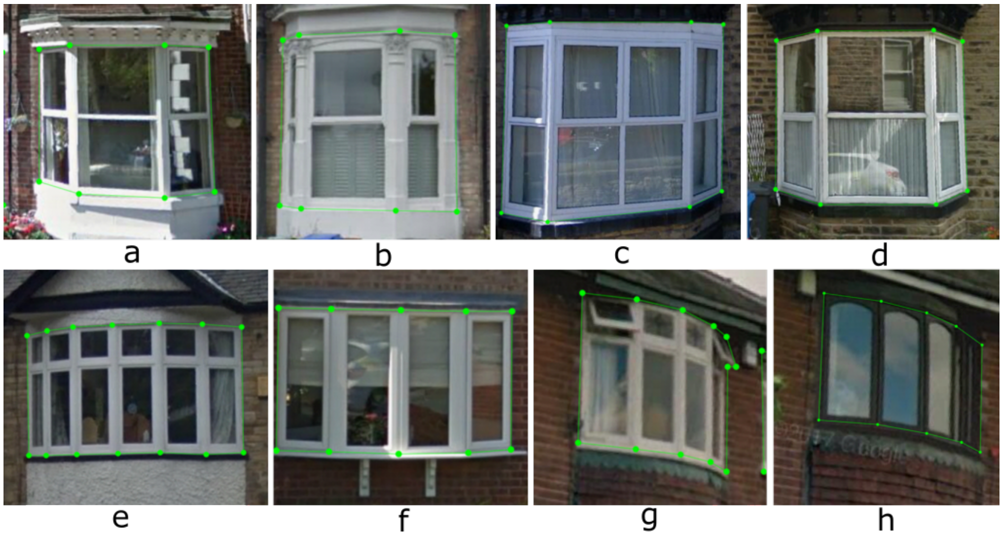


Figure 3.7: *Examples of bay and bow window annotations. The top four examples show bay window annotations and the bottom four show bow window annotations. Bay window examples a, c, d should be put into standard window class and b should be in the alternative window class. Bow window examples e and f should be in the standard class and g and h should be in the alternative.*

ever, this type of window can still be recognised as part of a door as they are surrounded by a same door frame. This decision will not add additional confusion as the door labelling rule defines a door including the door frame and area within it. If a window is not included in a door frame, the window will be annotated following the window annotation rule. Annotation of very tiny windows is another problem. These windows commonly have a blurry boundary, extremely low resolution and are highly-distorted. Regardless of what angles these windows are captured at, they are annotated as alternative windows. Figure 3.8 shows some examples of annotating windows around a door and very tiny windows.

In comparison with windows, doors do not have many variants and are more simple to annotate. Only doors on a facade are labelled. The main consideration includes if the door is partially covered by other objects or are significantly distorted. These doors need to be put into 'alternative door' class. Other considerations include whether the door is open or not; if the door is open, cover the whole frame and classify it as 'alternative door'. It is noted that in this project, garage doors and commercial building doors are not considered. The consideration is out of scope as this project is only about residential buildings. Also, doors made from steel bars, i.e. gates (normally found in terraced houses), are not considered. Figure 3.9 shows examples of alternative door annotations.

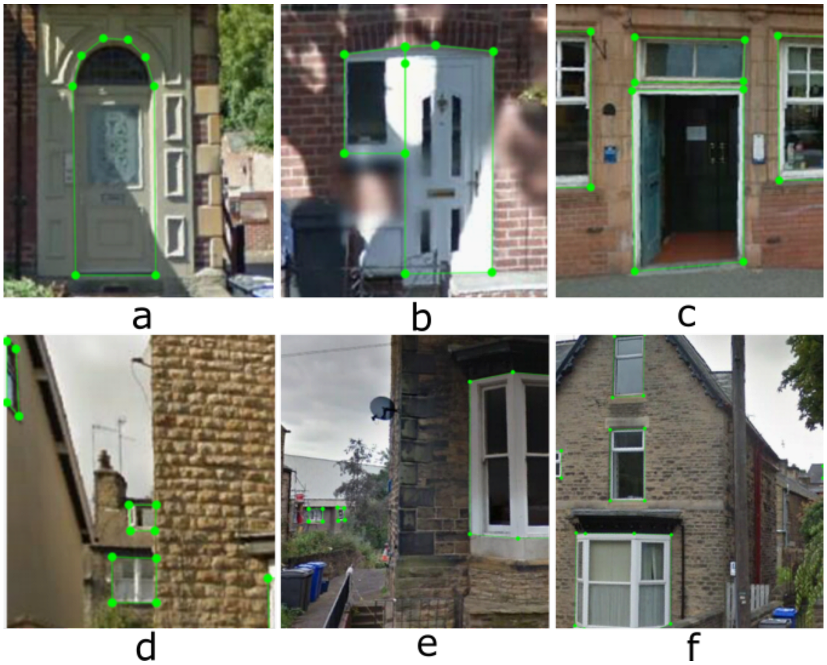


Figure 3.8: *Annotation examples of windows around doors and tiny windows. In example a, the window above the door is annotated as part of the door as it is within the door frame. Examples b and c are annotated as separate windows as they are not surrounded by the door frame. Example d, e, f show examples of tiny window annotations.*

Figure 3.10 shows some other abnormal annotation conditions. Figure 3.10a-b show uncertain area examples. Figure 3.10a shows two doors in shade. Under this condition, it is hard to estimate whether the area above the two doors is window or wall. Thus, the area above is not annotated. In Figure 3.10c, the area behind the plant is also uncertain and hence not labelled. Figure 3.10b and d show instances made of unusual materials. These examples should be annotated as usual but be classified as the alternative. Figure 3.10e-f show unusual bay windows. The annotations are created by following the boundaries of these instances. This is because windows here are installed separately unlike that in common bay windows: glass is installed within the same frame.

Figure 3.11 shows four high-quality fully annotated examples. In the four examples of the figure, all instances occurring in the figure are correctly, tightly annotated, i.e. no missing instances, no incorrect labels and no gaps. Furthermore, all annotations precisely follow the object boundaries which maintain the shape information of the targets.

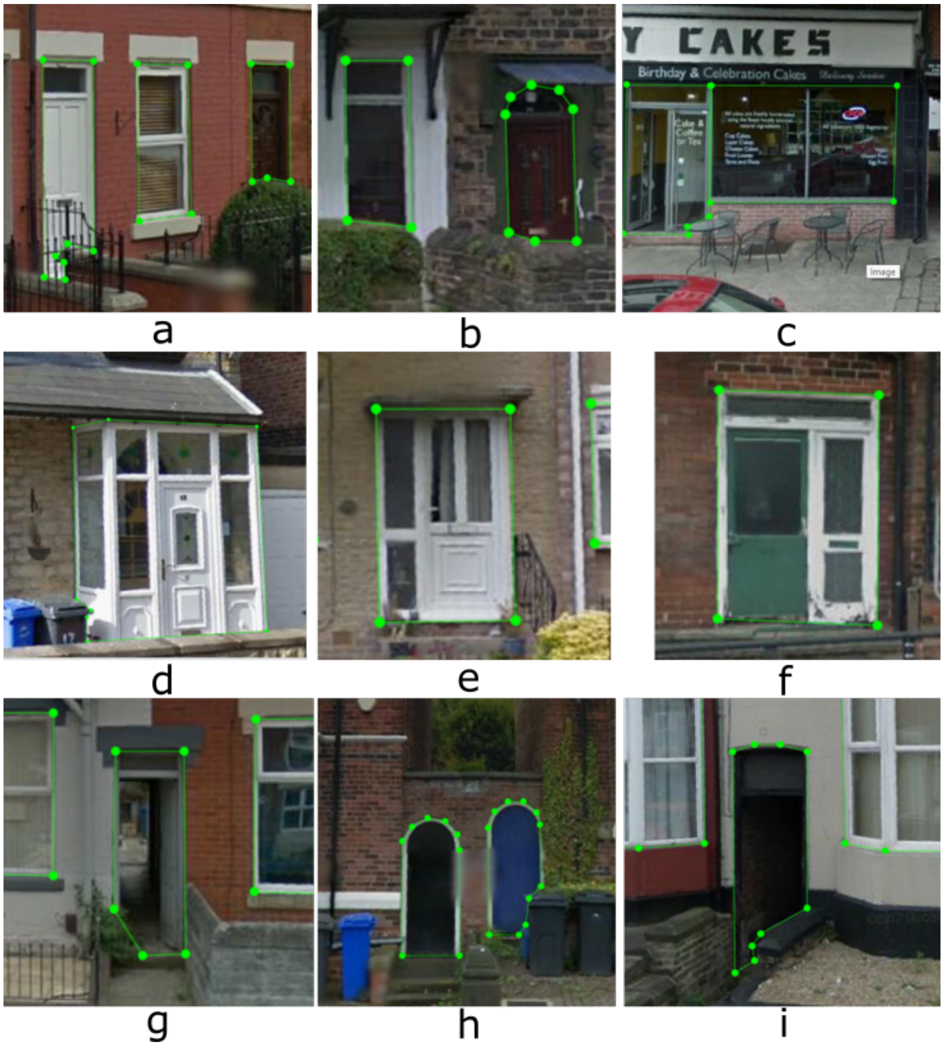


Figure 3.9: *Alternative door annotation examples. Examples a,b and c show partially occluded, and opened door annotations. Examples d, e and f show annotation of extended-frame doors: all areas within the door frame should be labelled as the same door. Examples g, h and i show annotation of outer corridor doors.*

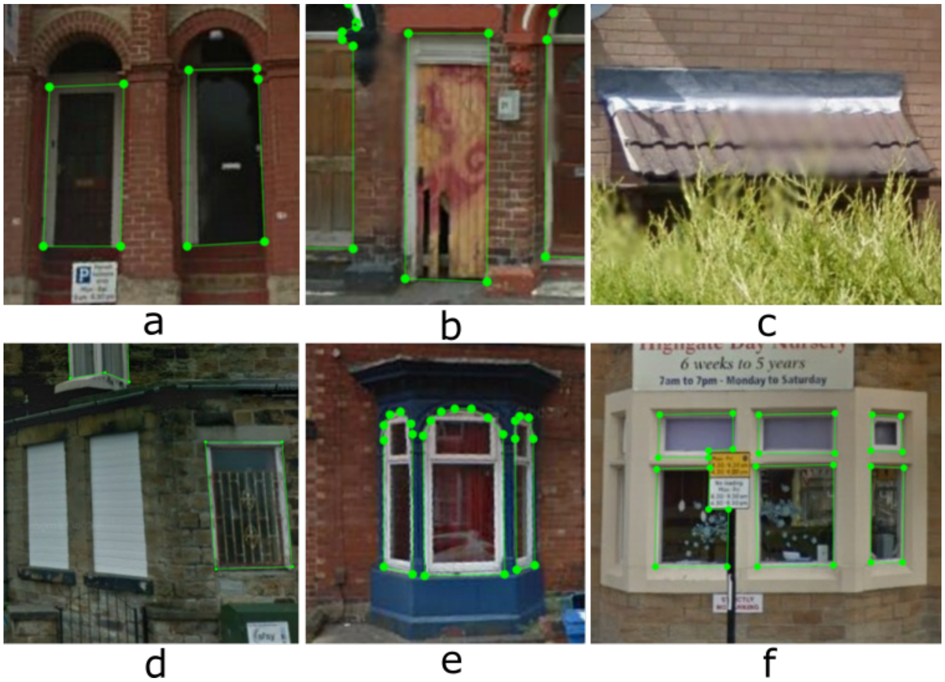


Figure 3.10: *Examples of abnormal instances' annotations. a. due to illumination conditions, if what is above the door is ambiguous, this should not be labelled; b. unusual appearance of objects should be labelled as the alternative; c. what is behind the vegetation is unclear, this should be labelled as the alternative; d. sealed window should not be labelled; e.-f. heavily decorated bay window annotation should avoid decorations.*

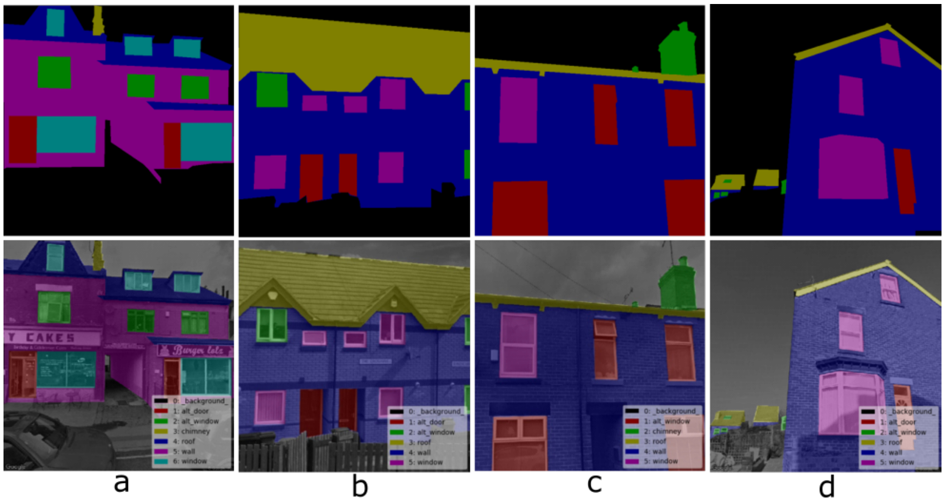


Figure 3.11: *High quality annotation examples. It is noted that these examples are only for illustration purposes and colour scheme inconsistency can be ignored, when the actual dataset was being built, the colour scheme was followed consistently which means across the whole dataset, the objects belonging to the same class should be labelled in the same colour. a. houses with ground-floor shops; b. fence-occluded terraced houses; c.-d. houses with limited views of roofs.*

3.3 Crookesmoor Dataset and Annotation Feasibility Experiment

3.3.1 Crookesmmor Dataset

In the previous section, a building facade annotation pipeline was constructed. In this section, a mid-size dataset is first constructed using the developed annotation protocol. The images used for building the dataset were captured in the city of Sheffield, UK. The buildings in this dataset are visually matched with the British residential building typology database (Loga et al., 2016). The database classifies the building typologies based on building types, e.g. detached, terraced, and sub-classified based on building age bands. The building age in this area range from 19th to the 21st century which covers the majority of the age bands determined by the database. The three main building types defined in the database include single-family, multi-family and terraced houses. Examples of each one are observed in the captured images. A map of the data capturing route is shown in Figure 3.12 with examples of each building type highlighted, along with its corresponding location.

The raw data collected by MARVEL platform is spherical-view videos. The video data can be easily transformed to image data by isolating the data into individual frames. This type of image is still in spherical view and thus highly distorted. A cube mapping technique is applied on the captured images. The technique is to map the spherical-view data to six environmental mapping images, each with an image size of 2048 × 2048 pixels (Lambers, 2020). The six images form a cube covering the entire FOV with a front, right, back, left, top and bottom view. The top and bottom views, which predominantly show the sky, road and sensors, respectively, were not used in training or prediction. The benefits of using the cube mapping technique include 1. a primary view can be set to focus on a priority direction, 2. compared to spherical view images, cube mapping images are less distorted although they are not rectified.

The dataset is built with 997 urban scene building images [2]. The dataset is randomly split into training, validation and test sets with ratio 80%, 10% and 10%, respectively. Thus, the training set has 797 images, the validation and the test set have 100 images each. The ratio selected is a commonly used means of creating an evaluation dataset, as seen, for example, in (Robinson et al., 2018; Syrris et al., 2020). The dataset was annotated by professionally trained annotators which will potentially provide a higher-quality annotation result than inexperienced annotators. These annotators are occupational dataset builders of machine-

[2] The 997 images were manually picked from image frames of the captured video images by the candidate. Images with overlapping buildings as well as images without facades are avoided.

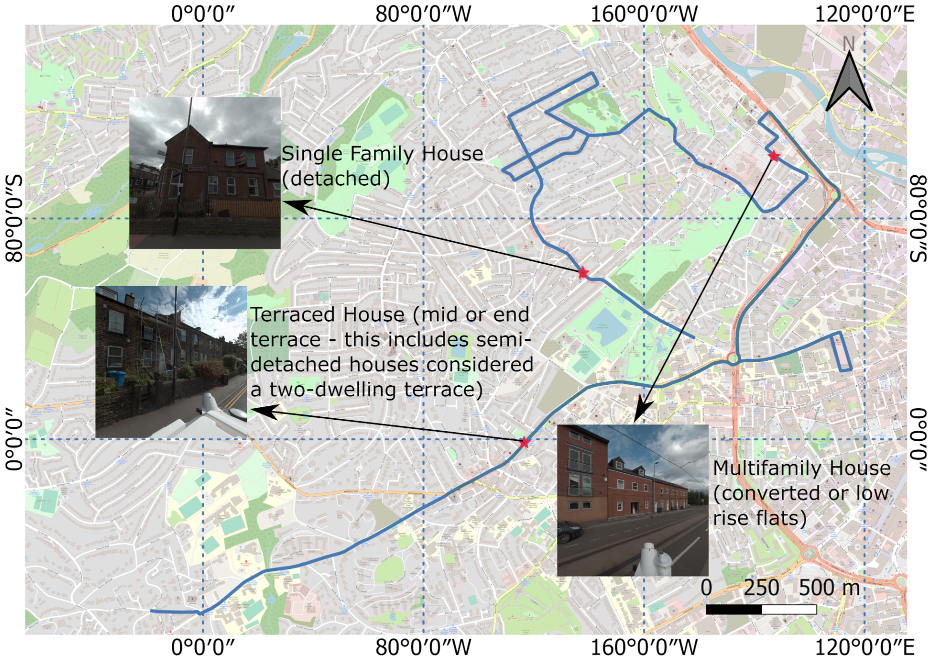

0°0′0″
80°0′0″W
160°0′0″W
120°0′0″E
80°0′0″S
0°0′0″
N
Single Family House (detached)
Terraced House (mid or end terrace - this includes semi-detached houses considered a two-dwelling terrace)
Multifamily House (converted or low rise flats)
0 250 500 m

Figure 3.12: *The data collection route is marked in blue. The route is selected in a typical suburb of the North of England residing on the outskirts of Sheffield city centre. The route contains a wide range of residential building typologies defined in the TABULA database Loga et al. (2016). Examples of the three main residential building typologies with corresponding descriptions are marked by the red stars.*

learning tasks. An iterative training procedure was taken to train these annotators to be familiar with the facade annotation task by the candidate[3].

The annotation quality is assessed in detail in Table 3.3. Figure 3.13 shows examples of those images with problematic annotations. The dataset evaluation work provides guidance for future annotation work, especially if hiring lower-cost amateur annotators. These frequent labelling errors can be directly listed in the annotation requirement list to avoid them in the future. The annotation error descriptions are also listed below.

[3] The candidate acknowledges that the Crookesmoor dataset is outsourced to URBAN DATA DYNAMICS LTD for annotation and it is contracted by Dr Gregory Meyers. Dr Gregory Meyers also takes the role of communication with the outsourcer including providing annotation instructions and feedback information produced by the candidate.

Table 3.3: *Problematic annotation summary.*

Problem numbering	Annotation problem description	Problematic image quantity
0	Unlabelled lean-to roof	63
1	Window on a roof annotation error	27
2	False label assignment	22
3	Other annotations missing	19
4	Billboard inclusion	6
5	Others	15

- The most frequent problem is the missing lean-to roof annotations. Lean-to roofs which have the same appearance as top roofs are commonly built on ground-floor walls. Sixty-three images have this problem. This is a significant problem since the shed roof is very common and if these roofs are labelled as walls, it might confuse the model in predicting wall and roof. Examples are demonstrated in Figure 3.13P0. For instance, in P0-a, the second building from the left side has a lean-to roof above the window but it is labelled as part of the front wall. As the high frequency of the problem occurs, it is necessary to pay special attention to it for future annotation projects.

- The second problem is annotating windows on a roof. For a roof window structure, the front side is usually the window with a roof on the top side. The two side parts could be made of brick or artificial material board or glass. If the side parts are made from glass, they should be annotated as window. However, if they are made from other materials, they should be accounted as wall, not roof or background. Examples are shown in Figure 3.13P1. The first example shows the condition that the roof window wall is not labelled and the remaining three show walls incorrectly categorised as roofs.

- The third and fourth problems are wrong classification and missing annotation, respectively. Examples are shown in Figure 3.13P2. P2-a annotated a window as a door, P2-b annotated a roof as a door.

- The fifth problem is billboard avoidance. It was decided that large billboards attached on walls would not be labelled. However, annotators occasionally ignore this setting. P3 shows examples in which billboard is included as part of a wall.

- Apart from the five main problems, other labelling errors include unusual circumstances such as how to define a church tower, and whether solar panel should be included as part of a roof. These problems are very unique and thus cannot be grouped properly in the guidance. As these problems are rare, these problems are recognised as data noise.

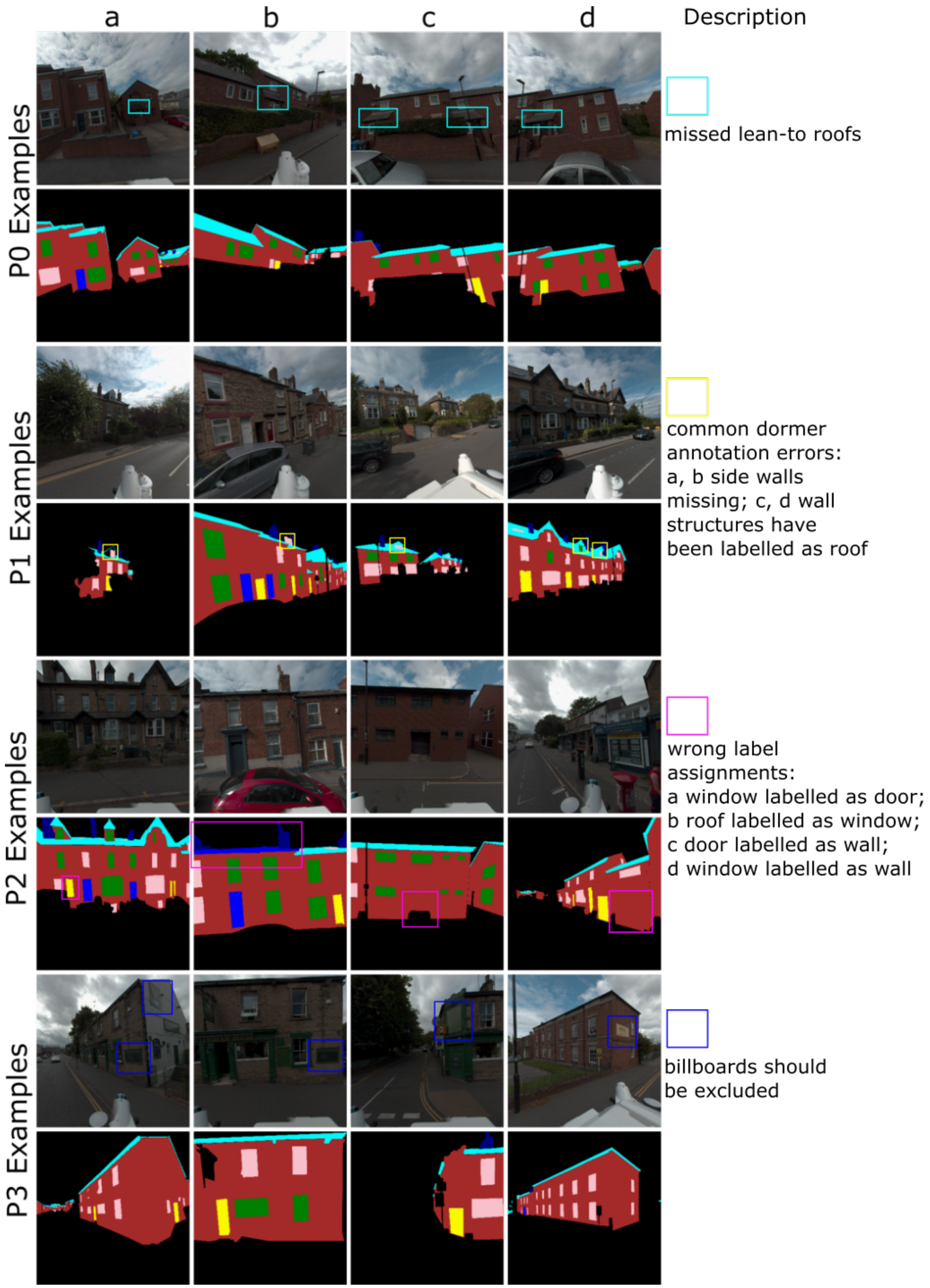

a
b
c
d
Description
P0 Examples
□ missed lean-to roofs
P1 Examples
□ common dormer annotation errors: a, b side walls missing; c, d wall structures have been labelled as roof
P2 Examples
□ wrong label assignments: a window labelled as door; b roof labelled as window; c door labelled as wall; d window labelled as wall
P3 Examples
□ billboards should be excluded

Figure 3.13: *Visualised problematic annotations. P0 demonstrates the missing lean-to roof annotation problem; P1 shows the dormer annotation error; P2 shows wrong label assignment problem and the missing annotation problem; P3 shows the billboard inclusion problem.*

3.3.2 Labelling Strategy Feasibility Test

3.3.2.1 Test Setup

Although trained deep learning-based models can largely reduce the labour cost of a task, preparing the data required for them is still extremely costly, both in labour and financially. Therefore, testing the proposed annotation strategy before promoting it to the whole dataset is essential to avoid unnecessary costs. The first 240 annotated images in the Crookesmoor dataset are used to test the validity of the proposed annotation strategy.

The labelling strategy feasibility test is designed to validate: 1. the feasibility of the class definition, 2. the feasibility of the occlusion labelling, 3. the feasibility of the alternative class strategy. The annotated multi-class labels are isolated first to create binary labels for each class. An original U-Net model with an extra encoder and a decoder block is adopted for each class in the testing work to examine the labelling strategy feasibility. A binary classification task is easier than a multi-class classification task. Therefore, using individual models on each individual class can potentially reduce the liability of the task to the model applied. As introduced in Chapter 2, the model has a symmetry U-shape structure and wide expandability. The mentioned model components in this section such as convolution and pooling were introduced in section 2.2.

The model which has been built is demonstrated in Figure 3.14. The U-shape model contains a contracting path, an expansive path and skip structures. The contracting path of the architecture used here has six convolutional blocks. Every block has two convolution layers with a 3×3 size filter with a stride of 1×1, dropout layer, batch normalisation and rectifier activation. In addition, zero padding is applied in the convolution process to maintain the feature map dimension. These blocks will increase the number of feature maps from 3 to 1024. Max pooling with a stride of 2×2 is applied to each of these blocks except the last one. These max-pooling layers will decrease the feature map size from 256×256 to 8×8. The expansive path will increase the feature map dimension from 8×8 to 256×256 with 3×3 filter and stride of 2×2 deconvolution layer. The deconvolution layer will double the dimension of a feature map by two and decrease its number by two also. In every block of the expansive path, feature maps from the contracting path will be concatenated with the feature maps from the expansive path; and two convolution layers the same as those in the expansive path will be applied to reduce the number of feature maps. In the end, a convolution layer with a stride of 1×1 and sigmoid activation will be applied to reduce the number of feature maps to 1 that reflects the probability of the foreground segmentation.

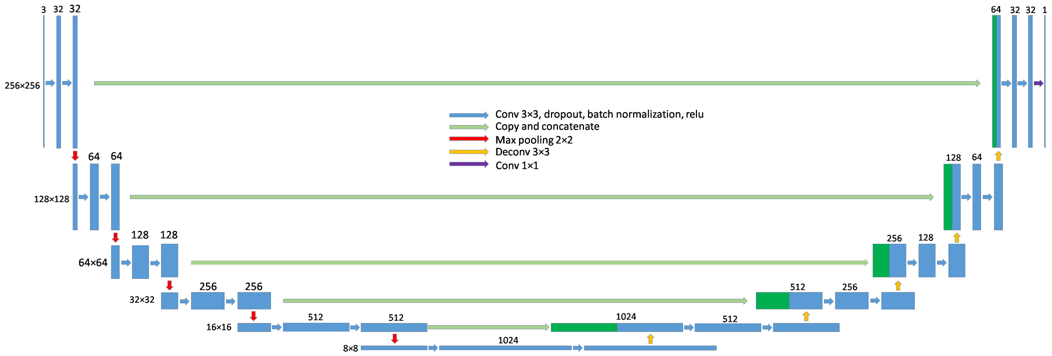


Figure 3.14: *The U-Net model built for the annotation feasibility test.*

The 240 images are randomly split into three parts for training, validating and testing the U-Net model. The choice of the split ratio is very important but empirical. The split ratio varies in different datasets, MS-COCO (Lin et al., 2014) uses the ratio (50%, 25%, 25%) for their training, validation and testing dataset, respectively; Cityscape (Cordts et al., 2016) uses the ratio (60%, 10%, 30%) for their dataset split. Considering the datasets used in this thesis have a much smaller size and the tradition of (80%, 20%) training and validation datasets split ratio in the machine learning community, a unique ratio (80%, 5%, 15%) is adopted. The reason a more consistent ratio (80%, 10%, 10%) is not adopted is that it was decided that the test set would be larger than the validation set to better reflect the generalisation of models. Thus, 192 images for the training dataset, 12 images for the validation dataset and 36 images for the testing dataset are achieved.

Applying data augmentation method into deep learning was originally introduced in the AlexNet paper (Krizhevsky et al., 2017). This method has been applied to medical imagery to intentionally produce more training images from the original ones before feeding the data into the U-Net model demonstrated in Figure 3.14. This is realised by performing multiple augmentation methods on original data, e.g. flip, rotate, shift, shear, brightness adjustment, etc. The horizontal flip is implemented with a 50% chance of occurring. Also, the width and height shift is applied with 50% chance of occurring with 10% moving distance. In addition, the hue is adjusted by 0.1. Building facade image itself limits the application of many other data augmentation methods compared to medical image datasets. For example, vertical flip and right-angle rotation cannot be used here since buildings will not be either up-side-down or falling-over.

A binary-entropy loss is adopted in this test for each individual model except for the chimney recognition. Binary-entropy loss is a commonly-used loss function in the semantic segmentation community which discretely measures the per-pixel similarity between the model predictions and the desired true values:

$$L_{bce} = -\frac{1}{N}\sum_{i}^{N}[y_i \log \hat{y}_i + (1 - y_i)\log(1 - \hat{y}_i)], \tag{3.1}$$

where y is the label, $\hat{y}$ is the predicted probability distribution, N is the number of pixels and i is the pixel index.

Chimney is visually significantly smaller than other objects. Therefore, a dice loss (Milletari et al., 2016) term is added to the binary-entropy loss to add a constraint on overlapping measurement:

$$L_{dice} = 1 - \frac{2\sum_i^N y_i\hat{y}_i}{\sum_i^N y_i^2 + \sum_i^N \hat{y}_i^2} \tag{3.2}$$

In tests of each class, built models are trained through the adaptive moment estimator (Adam) optimiser (Kingma and Ba, 2014). Unlike the traditional stochastic gradient descent (SGD) optimiser in which the learning rate is a constant, the Adam optimiser can update the learning rate by utilising the first and second moments of gradients. Other hyper-parameters are set as: dropout rate = 0.2, batch size = 3 and max epochs = 50. Dropout rate 0.2 is a commonly-used choice and batch-size is the maxima that the owned hardware equipment can carry.

The tests are assessed by the evaluation metrics: accuracy, precision, TPR (true positive rate), TNR (true negative rate) and the F1 score. These metrics are calculated through the binary classification confusion matrix. A confusion matrix is the summary of predictions. In a confusion matrix, the foreground prediction is known as positive and the background is designated as negative. The binary classification confusion matrix contains four values: the true positive and false positive (TP and FP, respectively) represent the correctly and falsely predicted true values, the true negative and false negative are the correctly and falsely predicted negative values. The five equations of evaluation metrics are tabulated in table 3.4.

The accuracy measures the percentile of correctly predicted pixels; the precision measures the percentage of correctly predicted foreground pixels over the whole foreground; the recall is the rate of the true positive over all ground-truth positives; the TNR measures the ability of predicting negative values; the F1 score measures the prediction accuracy considering both the precision and the recall values.

Table 3.4: *Metrics have been used in the annotation feasibility test.*

Metric	Calculation Function
accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$
precision	$\frac{TP}{TP+FP}$
recall	$\frac{TP}{TP+FN}$
TNR	$\frac{TN}{TN+FP}$
F1 score	$\frac{2\times precision\times recall}{precision+recall}$

All individual networks have been implemented with TensorFlow library (Abadi, 2016) and trained on an NVIDIA Quadro M1200 GPU with 4GB memory. It took an average of 50 minutes to train an individual model.

3.3.2.2 Test Results and Discussion

Table 3.5 shows the results of each individual test. While it can be seen that all models achieve very high accuracy and TNR values, the TPR values for positively detecting doors, alt-doors and alt-windows are low. High class imbalance is visually inspected in window, door and chimney classes. This highlights that, due to the highly imbalanced positive and negative classes, using only accuracy and TNR metrics would have been unreliable for measuring true model performance here. When the data is highly imbalanced, the TN value has a high influence which could lead to the accuracy and the TNR values being extremely high, which cannot reflect the real performance of the algorithm. Suppose there is a binary classification problem with n=100 samples (corresponding to pixels in this problem). There are five samples belonging to the target class. Suppose a model only successfully classifies one of those. However, the same model correctly identifies 93 of the negative samples: thus TP=1; TN=93. The sum in the denominator (TP+TN+FP+FN) is always 100 regardless of the breakdown. Thus the accuracy is 94%, despite only correctly classifying 20% of the true positive samples. The class imbalance of the built dataset is analysed statistically at the end of this chapter. Figure 3.15 demonstrates the class imbalance problem of the built Sheffield Crookesmoor dataset.

Instead, the F1 score appears to be a more reliable indicator of model performance. From the F1-score, it can be seen that the 'wall', 'chimney' and 'roof' class models achieved good performance, the 'window' class model achieved satisfactory performance and the 'door', 'alt-door' and 'alt-window' class models do not perform well. In this situation, dividing a single object type into common and alternative divisions cannot feasibly be evaluated as the low

Table 3.5: *Performance metrics of all of the tests. The test names with prefix 'Per' indicate the classes of objects without occlusions and distortions, the test names with prefix 'Alt' indicate the alternative classes and the names with prefix 'Com' indicate the classes that combine the corresponding normal and alternative classes. The trained model has failed to predict any alternative windows in the test set and thus leads to the precision and recall values being zero.*

Category	Accuracy	Precision	TPR	TNR	F1 score
Wall	0.929	0.893	0.846	0.961	0.869
Roof	0.987	0.670	0.883	0.990	0.762
Chimney	0.998	0.813	0.839	0.999	0.826
Per-door	0.953	0.322	0.143	0.987	0.198
Alt-door	0.985	0.156	0.318	0.989	0.209
Com-door	0.987	0.307	0.739	0.989	**0.434**
Per-window	0.979	0.705	0.637	0.991	0.669
Alt-window	0.983	0	0	0.983	0
Com-window	0.981	0.735	0.870	0.986	**0.796**

performance could result from limited data. Control tests are designed to validate the feasibility of differentiating the window and the door class into their correlated alternative and normal classes. The two tests train single base models on combined normal and alternative classes. The results are identified as those with prefix 'Com' in Table 3.5. By comparing the 'Com' class tests with their corresponding alternative strategy applied tests, an obvious increase is observed.

Figure 3.16 shows some examples of the visualised prediction results against their annotations. The top line is the captured raw images, the second line is the manually annotated masks and the third line is the visualised predictions. The wall example shows that the dense occlusions can be avoided. Although some boundaries are not smooth enough, considering it is such a small dataset, the results can still validate the strategy of removing dense occlusions to be feasible. The roof surface is correctly predicted and the attachments are covered in the prediction. The chimney example also achieves a feasible result. The models applied fails on the two alternative classes. For the two classes which only have perfect conditions of objects, the models easily confuse the objects with a perfect or an alternative condition.

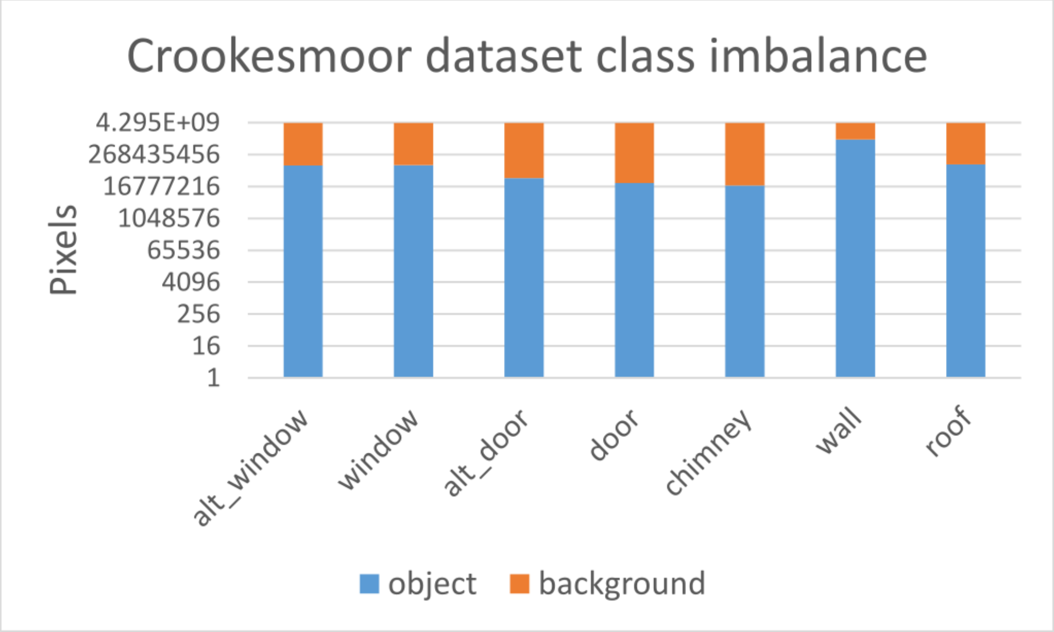


Figure 3.15: *A statistical analysis of class imbalance problem in Crookesmoor dataset. The figure has shown that pixel accuracy and TNR are not suitable for assessing madel performance in this task.*

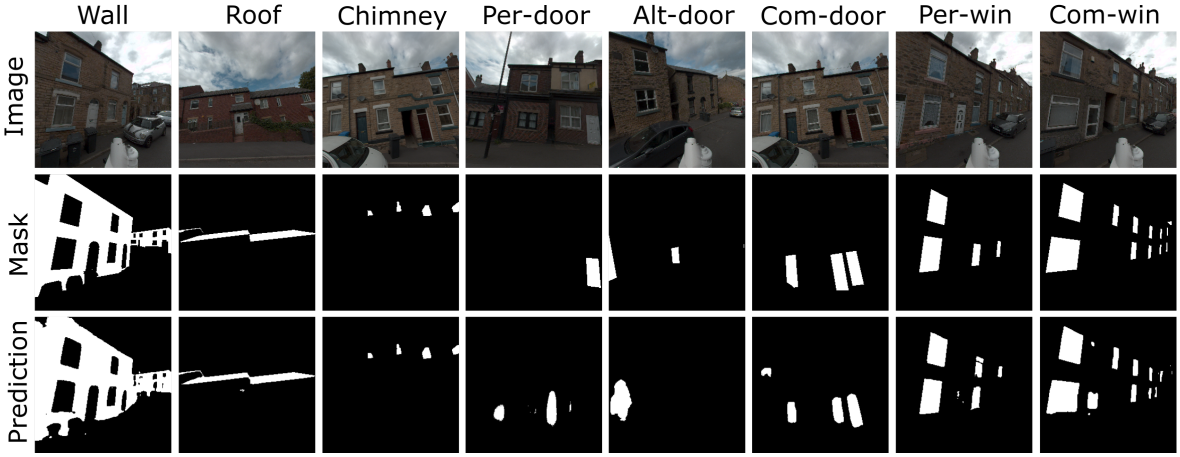


Figure 3.16: *The visualised results of the labelling strategy feasibility tests. The wall, roof and chimney have achieved satisfactory results; the door and the window tests show that the alternative strategy is not feasible in the building component recognition task.*

3.3.3 Generalisation Experiment

The challenge of whether a trained deep learning model can be applied further on unlabelled data depends on the generalisation of trained models and, more importantly, whether the dataset used for training contains adequate variations of targets. A generalisation experiment was collaboratively designed and conducted to validate the generalisation of the constructed dataset [4]. In this experiment, the U-Net models which were trained on combined window and door data in the previous subsection were adopted directly as a black-box ready-made package. An automatic element counter was developed by the candidate which computes the number of convex hulls in a prediction map. Noises are filtered out using morphology operations.

An independent batch of 42,451 images were captured in the southwest of Sheffield, UK with approx. 2500 inhabitants spanning 2.79 km^2 using the same data capture platform. This study area was deliberately chosen to be away from the area where the raw data in the Crookesmoor dataset was collected. The number of elements counted by the developed convex hull-counter were compared against manually counted values. The outcome shows the trained model can provide equivalent results as humans. In the key metric number of windows per-image and number of doors per-image, the trained model achieved 4.65 and 1.06 with human counting having 4.38 and 1.09, respectively.

3.3.4 Section Conclusions

Recalling the three purposes of the testing work which were validating the class definition, the occlusion labelling strategy and the alternative class separation strategy. Based on the results from all the tests, firstly, the alternative class separation is inappropriate as an obvious performance drop was observed. Secondly, the class definition and the occlusion labelling strategy is feasible. This is because the evaluation metrics show promising results in this small dataset and the visualised predictions mostly have smooth boundaries. The generalisation test shows that the built Crookesmoor dataset can be generalised to a larger area of buildings, at least in the Sheffield region.

[4] The candidate acknowledges that three colleagues in RISE research group took the role of manually counting elements of the collected data and comparing against the results from the developed automatic element counter. The three colleagues are Dr Hadi Arbabi, Dr Maud Lanau and Dr Xinyi Li. Comprehensive results are presented in (Arbabi et al., 2022) which have already been peer-reviewed and published.

3.4 Handsworth Dataset

In the previous sections of this chapter, the data semantic annotation procedure and proposed definitions of building facade decomposition are explored and validated. The data annotation procedure and facade decomposition definition constitute a building facade annotation protocol. A mid-size dataset, Crookesmoor dataset, was constructed using the annotation protocol. Inspired by the annotation pipeline, a larger dataset was constructed as a group-level project[5]. This dataset is introduced in this section, as it will be used later in this thesis.

In previous dataset construction works such as ImageNet (Deng et al., 2009), the strategy of using mixed professional and amateur annotators is widely adopted. The development of the Crookesmoor dataset validates a designed facade annotation protocol. However, the annotation service provided by the professional outsourcer which was used in developing the Crookesmoor dataset is expensive and time-inefficient, therefore, it is worth investigating a more flexible and cost-effective annotation outsource. Amazon launched their crowd-sourcing marketplace service, Mechanical Turk (MTurk). This service enables individuals and companies to outsource their tasks and distributes them to an available workforce.

Crookesmoor dataset validates the proposed labelling pipeline, however, the dataset scale is very limited, therefore, the second data capture was planned and conducted in Handsworth which is a typical suburb in the North of England. Figure 3.17 shows a snapshot of an area in Handsworth. After the raw data were captured, instead of using the cube mapping technique as in the Crookesmoor dataset, the frame images of each camera sensor were adopted for annotation without pre-processing. In comparison with images processed using the cube mapping technique, the frame images are highly distorted due to fish-eye lenses.

The collected images were picked first to avoid similar scenes and to maintain the dataset diversity. After the selection process, 6587 images were outsourced to Amazon MTurk for annotation. The annotation rule is simplified in comparison with the one provided to the Crookesmoor dataset outsource. The annotation guidance in the Handsworth dataset is listed in Table 3.6. Although the annotation guidance is simplified, common mistakes observed when constructing the Crookesmoor dataset, such as missing bay window roofs are specifically highlighted. A brief automatic annotation quality inspection was made after receiving the annotated data. The automatic inspection is to clean unlabelled and insufficiently labelled images. The automatic data cleaning detects 120 unlabelled and 561 insufficiently labelled images. The remaining 5906 images were kept and named as Handsworth dataset.

[5] The candidate acknowledges that Dr Wil Ward is responsible to outsource the data annotation project to Amazon MTurk and Dr Hadi Arbabi is responsible for taking the raw data capture.

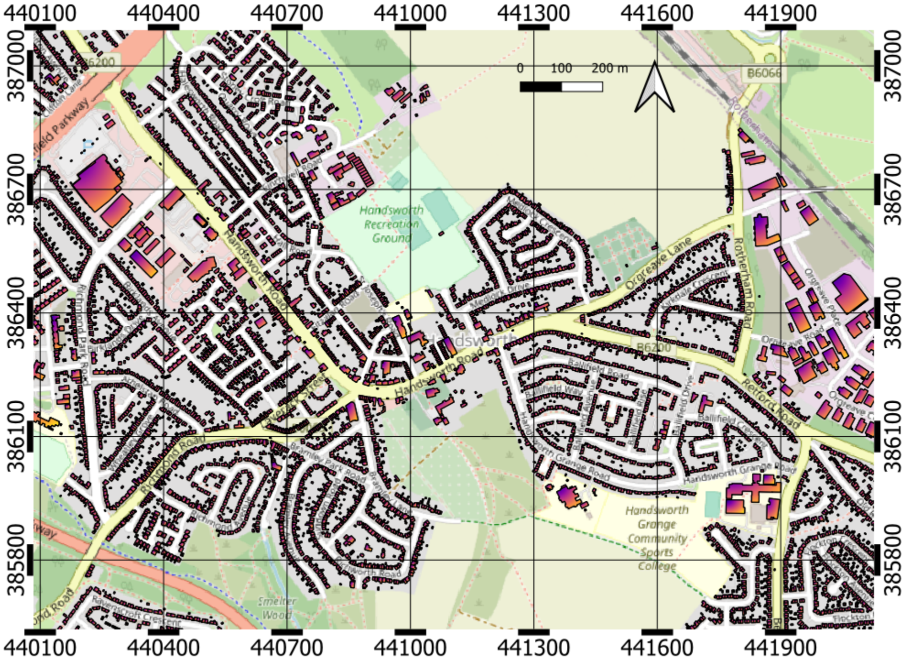


Figure 3.17: *A snapshot of the Handsworth area. The area is also a typical suburb in the North of England.*

3.5 Chapter Discussion and Conclusions

3.5.1 Chapter Discussion

3.5.1.1 Overview of Built Datasets

Any scalable machine learning-based solutions for the problem of semantic segmentation must rely on a high-quality training set. Such training data must be reliably and accurately annotated, and any frameworks for procurement must facilitate the influx of new data to be manually labelled. A residential building facade annotation framework is proposed in this chapter. The proposed framework has shown satisfactory feasibility in scalable annotations. The determined labelling rule successfully balances the level of detail of annotation required and the annotation efficiency. The determined labelling strategy partitions a residential building into five categories: wall, roof, window, door and chimney. The strategy is enacted in line with the needs of collecting information for building energy modelling and material stock analysis.

Index	Label	What does it include?
0	background	Anything that cannot be categorised as the other items, including occluding items such as trees. Some experts disagree on garden walls so they may appear as label 0 or 1.
1	wall	Main facade component, typically drawn first then overlaid with other labels. May include garden walls, background garages, etc.
2	roof	Visible roof components including eaves, may include bay window roof and dormers. May include skylights.
3	chimney	Visible chimney aspect, limited data due to small pixel sizes.
4	window	Glazing panels, possibly including skylights. Opaque windows, e.g. covered with signage, may be labelled as window.
5	door	Ground level doors, including inset doors. Should not include alleys.

Table 3.6: *Annotation guidance of Handsworth dataset*

A semantically labelled building facade dataset, Crookesmoor dataset, was constructed in this chapter. A portion of the Crookesmoor dataset was used to validate the feasibility of the determined partition strategy. The chapter also introduces another semantically labelled building facade dataset, the Handsworth dataset. The two datasets constitute all image data which are used in this thesis. All data used in this chapter was collected by driving a vehicle around suburban areas of Sheffield, the North of England. The two datasets vary in many facets. In the dataset scale, Crookesmoor dataset contains 997 images and the cleaned Handsworth dataset has 5906 images. The raw data in the Crookesmoor dataset is pre-processed with cube mapping algorithm first and then labelled by trained professional annotators. In the Handsworth dataset, raw frame images were directly outsourced to annotation.

The two datasets also use different outsourcers. Owing to the inefficiency and high-cost of the Crookesmoor dataset annotation outsourcer, Amazon MTurk service was used for the Handsworth dataset annotation task. The Crookesmoor outsourcer has a fixed team and the annotators were trained in this task during the annotation. MTurk has insecure annotators for which an annotator only relied on a literal guidance to learn the labelling rule. The MTurk service has shown advantageous labelling efficiency and flexibility. The designed data cleaning process has largely reduced problematic and unnecessary image-mask pairs. In comparison with using fixed-member outsourcer. MTurk has shown advantages in large-scale annotation tasks.

There are differences between Crookesmoor and Handsworth datasets in their presentations and annotations. One of the differences is the image processing technique. The type of raw data collected by the vehicle-mounted data capturing system is video. The data capturing system contains six cameras and, therefore, in each data capture journey, six videos are recorded. Frames at the same timestamp of the videos can provide a spherical view of the captured environment. Frames containing different residential buildings in the video file are manually selected. In the Crookesmoor dataset, the cube mapping technique was applied to generate image data. Cube mapping technique can generate six images from frames at the same timestamp which can form a cube covering the entire FoV. In the Handsworth dataset, instead of applying the cube mapping technique, the frame images from each camera are used directly for annotation without pre-processing. The frame images are highly distorted due to the wide-angle lens. Moreover, the buildings in the Handsworth dataset can only occupy c.1/3 of the total area. Sky and road commonly occupy the rest of the image. On the contrary, images generated through cube mapping can occupy the majority of the image.

In the Crookesmoor dataset, the building components are labelled to maintain their shapes at most extents except for meeting highly dense occlusions. This is because the shape feature may be significant in recognising the target objects and predicting boundaries for deep learning models. In the Handsworth dataset, it is argued that the occlusions should be completely avoided to maintain the surface material consistency of building components, therefore, the Handsworth dataset shows a more fragmented annotation than the Crookesmoor dataset.

3.5.1.2 Statistical Analysis of Datasets

The two developed datasets were analysed statistically and compared with publicly available datasets for facade segmentation. The comparative analysis has eight different datasets including the built Crookesmoor and Handsworth datasets. The other six datasets are Oxford RobotCar-Facade (Wang et al., 2022), TMBuD (Orhei et al., 2021), Varcity (RueMonge2014) (Riemenschneider et al., 2014), and three early datasets: eTRIMS (Korč and Förstner, 2009), ECP (Teboul et al., 2010) and LabelMeFacade (Fröhlich et al., 2010). The details of these datasets have been reviewed in section 2.3.2.

Among the eight datasets, eTRIMS and ECP have provided statistical details or been previously analysed such as in (Kong and Fan, 2020). Crookesmoor dataset annotations come in JSON files from the outsourcer, thus its statistics can be achieved by summarising information from them. The other five datasets have not been analysed statistically before. Therefore, a program was developed to automatically and statistically analyse datasets under this situation from their annotation images.

Table 3.7: *This table shows a statistical analysis of the developed building facade datasets and publicly available datasets. SCBuD and SHBuD represent the Sheffield Crookesmoor and Handsworth datasets. LabelMe is the abbreviation of LabelMeFacade. Oxford is the abbreviation of the Oxford RobotCar Facade dataset. The number of alternative- and normal-class objects of the Crookesmoor dataset have been combined in this table.*

Dataset Category	SCBuD	SHBuD	eTRIMS	ECP	LabelMe	Varcity	TMBuD	Oxford
Announcement Year	2019	2021	2009	2010	2010	2014	2021	2022
Number of images	997	**5906**	60	104	945	219	300	500
Image size (MPx)	4.2	4.2	0.4	≤ 0.3	0.3	0.9	0.8	1.2
	Number of objects							
Wall	4251	**183,946**	142	104	3593	456	2811	1105
Window	14,770	**195,868**	1016	2976	7664	5834	8770	8820
Door	2406	**26,250**	85	94	863	196	860	331
Roof	4052	**115,754**	N/A	104	N/A	219	N/A	N/A
Chimney	2536	**34,688**	N/A	N/A	N/A	N/A	N/A	N/A
	Average per-object resolution (Px)							
Wall	**234,021**	11,671	74,166	82,005	38,891	137,750	32,567	127,686
Window	**14,175**	2669	2634	855	1519	2503	1656	2118
Door	**23,613**	4110	3809	2726	2520	14,561	3338	4566
Roof	**27,795**	65,822	N/A	11,217	N/A	24,012	N/A	N/A
Chimney	**4101**	1567	N/A	1042	N/A	N/A	N/A	N/A
	Number of pixels (MPx)							
Wall	994.8	**2146.9**	10.5	8.5	139.7	62.8	91.5	141.1
Window	209.4	**522.8**	2.7	2.5	11.6	14.6	14.5	18.7
Door	56.8	**107.9**	0.3	0.3	2.2	2.9	2.9	1.5
Roof	112.6	**710.5**	N/A	1.2	N/A	5.3	N/A	N/A
Chimney	18.4	**54.4**	N/A	0.3	N/A	N/A	N/A	N/A

The program is based on the element counter software developed in section 3.3.3. Modifications were made to fit the statistic analysis task. The element counter was written using object-oriented programming (OOP). Therefore, functions can be conveniently added to the software without sabotaging its initial architecture. A function which can summarise foreground and background pixels was added and the function of removing noise is made void. The updated software was validated using Crookesmoor and ECP datasets before populating it to other datasets. The statistical analysis is tabulated in Table 3.7.

Among the eight datasets, the Handsworth dataset has the largest volume and the Crookesmoor dataset has the second-largest volume and the highest average per-object resolutions. The Crookesmoor dataset is annotated to preserve shape information, therefore, the number of walls can then be used to infer that the Crookesmoor dataset contains c.4000 different scenes of buildings. However, as the Handsworth dataset is labelled in a more 'fragmented' way, the number of walls does not indicate how many scenes are in the dataset.

However, the number of scenes may be estimated by the number of doors and chimneys.

Table 3.7 also shows that before 2014, facade segmentation datasets commonly had lower image sizes. Nowadays, facade segmentation datasets with image size above full-definition (HD) i.e. 1280×720 have become more readily available. The latest three datasets, Varcity, TMBuD and Oxford RobotCar-Facade are all above or approximate to the HD-level image size. However, an increase in image size (number of pixels) does not necessarily yield an increased resolution in object space (pixels/meter). The eTRIMS dataset has an average window resolution of 2634 pixels which is equivalent to datasets with significantly higher image sizes, especially Handsworth and Oxford RobotCar-Facade datasets. This situation poses challenges in developing facade segmentation models.

3.5.2 Conclusions

In this chapter, two street view facade segmentation datasets were built focusing on the UK residential housing stock[6]. One has 997 labelled images and the other contains 5906 images. A data labelling framework was developed, considering potential needs in assisting scalable building retrofit. The two datasets were constructed using different outsourcers. The comparison has shown that the fixed-team annotation is more advantageous than the crowd-source MTurk annotation when annotation accuracy is the foremost requirement of the task. MTurk is a more balanced choice than the fixed-team outsource in the residential building facade labelling task.

[6] The two built datasets will be made available online after embargo.

Chapter 4

Residential Building Facade Segmentation

4.1 Introduction

Chapter 2 systematically reviewed the state-of-the-art datasets and developed approaches for facade segmentation. Recalling the literature review, facade segmentation is a dynamic and fast-developing scientific field. Prior to 2020, most works focus on developing approaches for rectified single-facade images for procedural modelling. As facade segmentation is such a fast-growing field, the timeline of this chapter is important in comparison with the state-of-the-art works. The Crookesmoor dataset was built throughout 2019 given the facade segmentation project started at the end of December, 2018. The first developed model FacMagNet was built in 2020 and is presented in section 4.2. The model was re-developed the following year and this is presented in section 4.3. The Handsworth dataset was built in 2021. The contents in section 4.4 were finished in the same year. The timeline is structured in table 4.1 with the state-of-the-art (SOTA) works at the same time in facade segmentation.

The overall contribution of this chapter includes 1: development of novel scalable approaches to the automation of residential building facade component recognition; 2: exploration of an advantageous end-to-end deep learning model and training strategy. The Crookesmoor dataset was adopted for contribution one to develop the facade segmentation models. The Handsworth dataset was adopted for contribution two. The solutions outlined in this paper are summarised thus:

Table 4.1: *Thesis timeline with corresponding state-of-the-art works. The year 2021 was rather a quiet time for facade segmentation but many papers were published in 2022 according to the literature review. This might be because of the impact from the global Covid-19 pandemic.*

Time	Thesis work	State-of-the-art
November, 2019	Crookesmoor dataset was built	SRF+RPN(Rahmani and Mayer, 2018) and MULTIFACSEG-NET(Femiani et al., 2018)
August, 2020	FacMagNet-l was developed in section4.2	DeepFacade(Liu et al., 2020)
February, 2021	Handsworth dataset was built	Pyramid ALKNet(Ma, Ma, Xu and Zha, 2020), Combo representation(Kong and Fan, 2020)
March, 2021	FacMagNet-s was re-developed in section4.3	FacMagNet-l(Dai et al., 2021)
September, 2021	Attention mechanism experiments in section4.4	No update
November, 2022	Thesis completion	DETR+PSPNet(Zhang et al., 2022), DeepFacade-v2(Liu et al., 2022)

1. a novel ensemble segmentation model tailored to handle facade images with inter-category size discrepancies e.g. window and wall and intra-category single-object class imbalance, e.g. due to perspective or capture distance, by incorporating a novel magnifier strategy;

2. the developed magnifier model is simplified to consider efficiency and achieved promising performance on the same dataset;

3. an exploration study using focal loss and attention mechanism on class-imbalance problem and testing an advantageous end-to-end model.

The next three sections describe the development of the residential building facade segmentation models. The first section describes the developed facade segmentation model - FacMagNet-l. The second section describes the simplified version of the re-developed FacMagNet, FacMagNet-s. The last section contains the exploration study of applying focal loss and various model architectures and attention mechanisms. After these, a cross-dataset adaptability test is designed to demonstrate the performance of the proposed magnifier strategy.

4.2 Residential Building Facade Segmentation

4.2.1 Categorical Semantic Segmentation Models

As reviewed in chapter 2, supervised deep learning-based semantic segmentation models were developed based on the fully convolutional neural network (FCN) (Long et al., 2015) with an encoder-decoder architecture. The spatial resolution of the feature maps, i.e. the outputs of each convolution layer, decreases throughout the feature extraction process i.e. the encoder network. This allows the learned feature maps to be more invariant to small translations of the inputs. Consequently, the ratio of the input image size to the output feature map size, known as downsampling rate, becomes a significant concern as redundant spatial resolution reductions will lead to target objects vanishing and insufficient resolution reduction may result in the model lacking sufficient translation invariance. Operations called skip connections were developed to concatenate feature maps at different levels, to help maintain the low-level information of the model, which is often lost in a linear convolution-deconvolution model (Long et al., 2015).

The multi-scale problem is a universal challenge in designing a CNN model for computer vision tasks. This problem means, in an image, the size of target objects varies in a large scope. Multiple techniques and model structures have been developed in this field including symmetric architecture (Badrinarayanan et al., 2017), feature pyramid (Zhao et al., 2017), and dilated convolution (Chen et al., 2018). These methods have shown deep learning-based models to have powerful capabilities to solve this problem. However, these approaches are designed to be a universal solution of urban scene segmentation, and, accordingly, lack refinement for a certain scenario, such as for facade segmentation.

U-Net model is another semantic segmentation model, based on the FCN, that was developed initially for medical images (Ronneberger et al., 2015). Its architecture has an efficient symmetric structure and is highly expandable. U-Net outperformed base FCN and related architectures, and the model structure has been applied in various fields, such as remote sensing (Chu et al., 2019). As introduced in chapter 2, the original U-Net comprises an encoder network with a standard CNN architecture, and a symmetric decoder network that recovers the spatial resolution of feature maps. Skip connections concatenate feature maps from the contracting path before doubling the number of feature channels to the symmetric feature maps in the expansive path. The design allows for features representing small object information to be transmitted to higher levels of the network. Compared with other multi-scale architectures, such as feature pyramids (Zhao et al., 2017), the symmetric architecture is able to better retain small object information. Because the images in the facade dataset contain a number of small objects, the benefits of the symmetric U-Net architecture are highly relevant to this problem.

Another benefit of using U-Net architecture in facade segmentation is its success on properties that are common in both facade images and medical images. For example, targets in medical images such as brain tumours usually have diffused and ambiguous boundaries which can make them difficult to segment (Havaei et al., 2017). Diffuse boundaries require low-level high-resolution edge information to refine the segmentation boundaries. In the captured facade image set, boundary ambiguity has been identified in all classes. Additionally, a degree of semantic information in the structure of a building has been identified, e.g. chimneys are typically located on roofs. This type of information is often found in medical images on which U-Net has proven to be effective, such as the human brain with a defined interior structure (Kermi et al., 2018). The high-level semantic information can support the detection of the target objects.

Employing the original U-Net architecture directly to the developed facade segmentation dataset is ill-considered. The original U-Net takes inputs with size 572 × 572 and has a downsampling rate of 16. The data captured by the vehicle-mounted system has a size of 2048 × 2048 and the intended input size for facade segmentation is 1024 × 1024. Therefore, data used here has a much greater size. In addition, the U-shape structure has been widely explored to fit it into different scenarios nowadays. As such, a new model was developed.

In the Crookesmoor dataset, most of the wall objects can occupy the majority area of an image and the roof objects are mainly slender shapes across the long-side of an image. However, the three smaller-sized categories, i.e. the window, the door and the chimney, have significant size differences because of elements including viewing perspective. By measuring the size of the minimum bounding rectangles (MBR) of the three smaller-size categories, the size distributions are plotted in Figure 4.1. These plots show that the objects in these three categories are distributed very widely and unevenly. An effective method of solving the high size discrepancy problem is to use different receptive fields aiming for different scales Hu and Ramanan (2017). Therefore, it was decided that the ensemble learning strategy would be adopted to build different models for different classes in this paper.

Ensemble learning is a common strategy in the machine learning community. The core of ensemble learning is to use multiple individual machine learning algorithms for a task and fuse predictions from them with a designated voting strategy to achieve better performance than using a single algorithm (Dong et al., 2020). In particular, random forest (Breiman, 2001) is an iconic representation of ensemble learning. Its variant, structured random forest

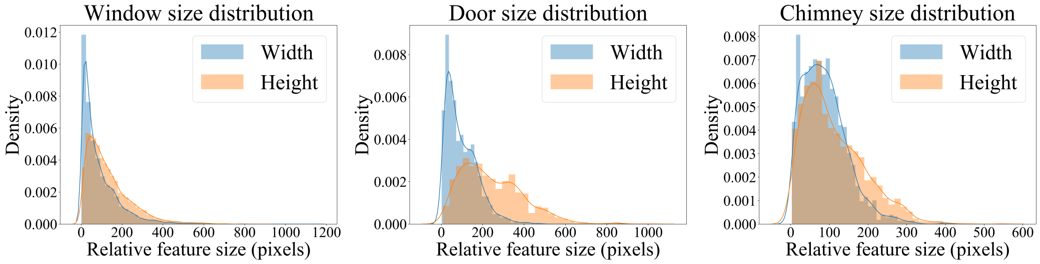


Figure 4.1: *Relative feature size statistics of window, door and chimney under the raw data image size; the plot shows the width and height distributions, the distributions show high varieties in sizes.*

has been used for facade segmentation in (Rahmani and Mayer, 2018). Ensemble learning has several different paradigms including bagging, boosting and stacking. Bagging is a parallel structure for which all individual algorithms generate predictions independently. Boosting is to improve model performance on handling difficulties. The strategy uses predictions from a base-model to adjust the dataset by applying weights and then training another base-model based on the weighted dataset. Stacking is a stage-wise approach which utilises a meta model to generate predictions based on inferences from base models.

To be more clear, bagging is a technique where multiple models are trained on different subsets of the training data, with replacement, to reduce variance and improve accuracy. The final prediction is obtained by averaging the predictions of all the models. Stacking is a more complex technique that involves training multiple models, or base models, on the same training data, but with different features or algorithms. The predictions of these base models are then used as input features for a higher-level model, called the meta-model, which learns to combine the base models to make the final prediction.

Three different downsampling rates were determined for this task. For the three smaller-size categories, the downsampling rate of 32 was selected to be 32, i.e. $log_2 32 = 5$ layers, which means the model will reduce the feature map size to 32 times smaller than the input size. The decision is in terms of the largest objects in the three categories occupying a significantly larger proportion than the target objects in medical images used in the original U-Net. For the roof model and the wall model, the downsampling rate is 64 and 128, respectively, because these two categories are both significantly larger than the smaller-size categories and thus require deeper models to extract semantic information.

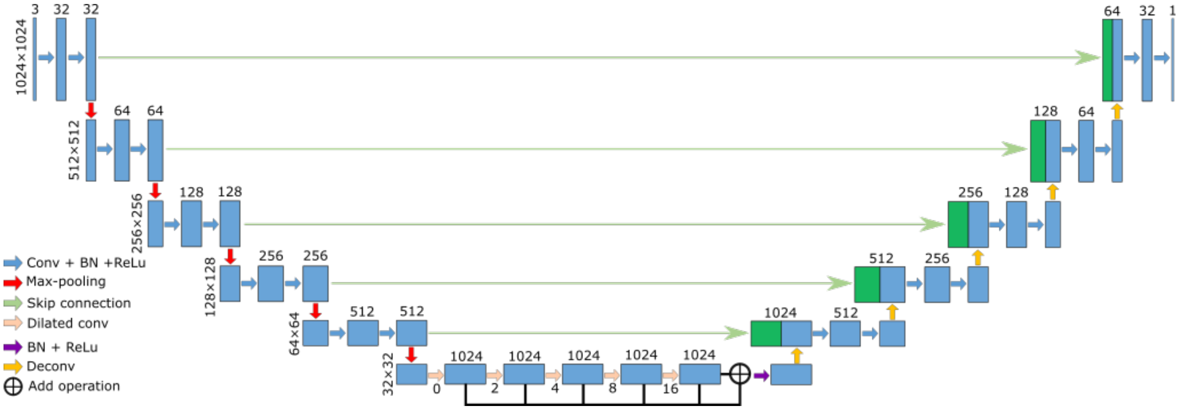


Figure 4.2: *The semantic segmentation model for window, door and chimney categories, the 'Conv' stands for the convolution operation and 'BN' is the abbreviation of batch normalisation which is a common way to prevent over-fitting. The numbers in the encoder-decoder network represent the channel number and numbers in the dilation block are the dilation rates. The feature maps from the dilation convolution layers are added at the end with subsequent batch normalisation and activation layer in the centre dilation block.*

Figure 4.2 shows an example of the model structure with a downsampling rate of 32. The black arrows indicate the skip connection which is the operation to concatenate the feature maps in the encoder network to the their symmetric ones in the decoder network. As the encoder network increases the translation invariance of the model, it loses detailed edge information and location information. The skip connection is included to combine low-level features with high-level features, which helps the model maintain information from different scales (Drozdzal et al., 2016). In the three smaller-sized categories, to prevent the small objects vanishing, a lower downsampling rate is selected. This benefits the detection of small objects in the image, however, there is a trade-off in the detection of larger objects. Inspired by the dilated convolution technique, which has been shown to extract richer semantic information, such as in the DeepLab and D-LinkNet models (Chen et al., 2017; Zhou et al., 2018), five dilated convolution layers with exponential growth dilation rates are utilised to replace the two convolution layers in the centre block. The five layers are concatenated using skip connections, as shown in Figure 4.2.

The detection of the roof and walls requires higher downsampling rates, and the convolution layers in the model are replaced with residual blocks(He et al., 2016*a*) to deal with the gradient vanishing problem: a common issue that occurs in this type of model architecture when detecting large objects. Residual blocks are built with a skip connection and two adjacent convolution layers to mitigate gradient vanishing. The centre dilation block is also replaced with two residual blocks for both the roof and the wall models.

The loss function is another crucial part of the model structure, that determines how effective the classification model is. A common loss function for the classification is the binary cross entropy, $\mathcal{L}_{bce}$, which represents the similarity between two distributions, and can be calculated as an average per-pixel loss. The dice loss, $\mathcal{L}_{dice}$ is another approach that represents the loss as a global function, i.e. it does not treat all pixels independently, like binary cross entropy. Dice loss is particularly useful for segmentation problems where there is class imbalance (Milletari et al., 2016).

In this work, a joint loss function is applied to combine the benefits of binary cross entropy and dice loss. For vectors containing true, $\mathbf{y}$, and predicted, $\hat{\mathbf{y}}$, pixel labels, the loss is defined:

$$\mathcal{L}(\mathbf{y},\hat{\mathbf{y}}) = \underbrace{-\frac{1}{N}\sum_{i}^{N}[y_i \log \hat{y}_i + (1-y_i)\log(1-\hat{y}_i)]}_{\mathcal{L}_{bce}} + \underbrace{1 - \frac{2\sum_i^N y_i\hat{y}_i}{\sum_i^N y_i^2 + \sum_i^N \hat{y}_i^2}}_{\mathcal{L}_{dice}},$$

where N is the number of pixels.

To combine the results of each semantic segmentation model, the output score maps are voted to find the most confident classification for each pixel.

Small object recognition is a common problem when using deep learning models. One of the reasons is that small objects often vanish in the down-sampling process. Symmetric model structure and ensemble learning strategy is adopted to solve this problem. A related issue is that small objects, by definition, only occupy a tiny area of an image. This can lead to severe class imbalance. As deep learning models are trained to learn gradients and minimise a loss function, class imbalance makes the model prone to classifying these pixels as background.

4.2.2 Using Object Detection as a Magnifier

In dealing with class imbalance as a result of differing object sizes in segmentation images, one approach was developed by cropping images into small tiles and feeding those into the model (Van Etten, 2018). However, using this approach directly can cause target objects to lose contextual shape information, which is essential in identification, especially in building facade images. Therefore, a new method is proposed: using an object detection model to extract objects from the image and applying a magnifying factor to balance the foreground and background. The magnification approach is only adopted for the three category models where small objects and class imbalance are observed in the data, specifically windows, doors and chimneys.

Mask-RCNN is an example of a model, designed for instance segmentation, that incorporates a joint object detection and semantic segmentation structure (He et al., 2017). However, as the design purpose is completely different, this model is not applicable in our task: the model uses only a single FCN model which, as discussed in the previous section, does not perform well in the multi-scale problems we are looking at. The Mask-RCNN model feeds the detected area directly into the semantic segmentation model, which does not balance input sizes to combat the intra-size discrepancy.

Object detection is an important topic in computer vision, the same as the semantic segmentation technique. The technique is designed to locate the target objects via bounding boxes. In previous work on building facade segmentation, object detection has been used as a shape refinement strategy (Rahmani and Mayer, 2018; Liu et al., 2020). As in the rectified frontal-parallel view facade images, objects such as windows and doors are in a rectangular shape. In the Crookesmoor dataset, it is not possible to use the technique as a shape refinement module. However, a potential is identified for integrating the technique to solve the class imbalance problem.

To use the object detection model, bounding box information is generated automatically by calculating the minimum bounding rectangles (MBR) of the pixel-wise annotations. For each annotation patch, its MBR coordinates are calculated first. As the MBR is not normally parallel to the axes, the coordinates of the minimum rectangle which covers the MBR, parallel to the axes, are calculated as the bounding box information.

An object detection model is trained to locate the bounding boxes of the three smaller-sized categories. Patches formed from the contents of the bounding boxes are expanded by a magnifying factor, based on their size. If the length of the bounding box's shorter side is fewer than 64 pixels, the area of the bounding box will be magnified by 25. When the short side is between 65 and 128 pixels, the magnification factor of 16, and all bounding boxes with short side larger than 128 pixels are magnified by 9. The magnified patches are tailored from the raw image and act as the input to their corresponding categorical semantic segmentation model. The output score maps of each patch are then recovered to their initial locations. The object detection model integration is shown in Figure 4.3.

To learn and predict bounding boxes, the Faster R-CNN model is used (Ren et al., 2015). In this model, a base CNN network is employed first to generate feature maps, similar to the encoder network in the semantic segmentation model. The outputs from this base network are fed into a region proposal network (RPN). The RPN proposes nine different anchor boxes for each point in the feature maps and determines if each of the anchor boxes contains a

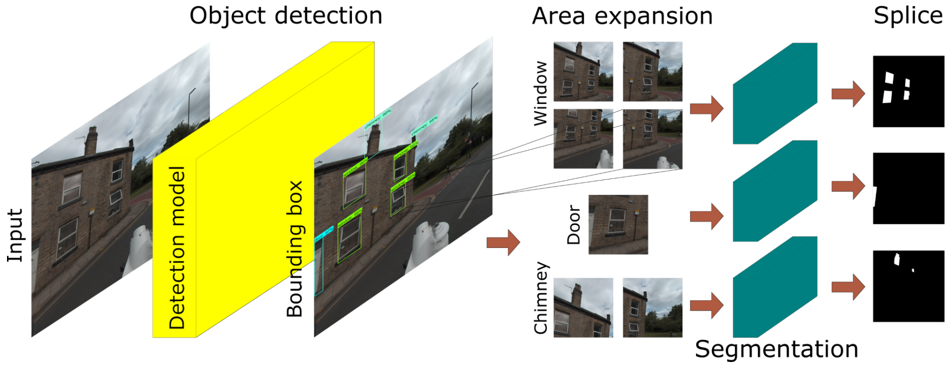


Figure 4.3: *Model workflow with object detection; the input image passes into the detection model to generate bounding boxes first; the interested areas are then expanded and extracted; the extracted patches are magnified to a unified size then fed into the corresponding semantic segmentation models; the output score maps are resized to the original size and spliced together in the end.*

target, along with their coordinates. The RPN uses the non-maximum suppression (NMS) to filter redundant anchor boxes. The technique determines a threshold, and any bounding boxes with an overlapping area larger than the threshold are removed. After the RPN, the classification and coordinates regression model will determine the category and refine the anchor box coordinates.

In this work, instead of using the VGG-16 model as the base network in the original paper of Faster R-CNN (Ren et al., 2015), the Inception ResNet-V2 (Huang, Rathod, Sun, Zhu, Korattikara, Fathi, Fischer, Wojna, Song, Guadarrama et al., 2017) is adopted. The anchor box ratios are fixed, as in the original paper, at 1:1, 1:2, and 2:1. The NMS threshold is fixed at 0.7.

The developed FacMagNet is demonstrated in Figure 4.4. The input image simultaneously passes into the detection model to generate bounding boxes and the two segmentation models for the wall and the roof class. In the detection branch, the interested areas are then expanded and extracted; the extracted patches are magnified to a unified size then fed into the corresponding semantic segmentation models; the output score maps are resized to the original size and spliced together at the end. At the end, the predictions from all the tailored segmentation models are merged together through voting. The voting strategy is made as if none of the scores from the tailored models in the same pixel are higher than 50%, the pixel will be set as background; otherwise, the category of the tailored model with the highest output score will be set.

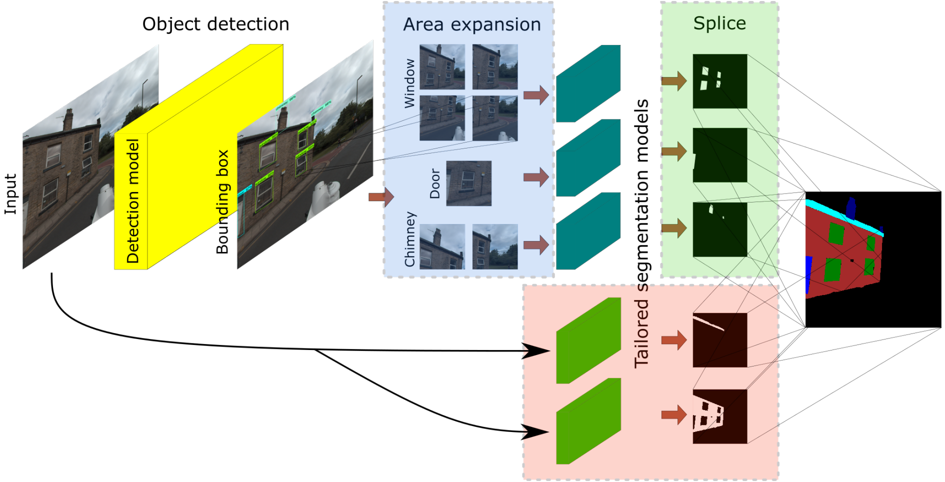


Figure 4.4: *The developed FacMagNet model workflow which combines the magnifier module.*

4.2.3 Experiments and Results

4.2.3.1 Training Strategy and Evaluation

Experiments are first conducted in each category to explore the best combination from various model choices discussed in section 4.2.1. A base U-Net model was built for the purpose of comparison. For the roof and wall category, a deeper U-Net model with a larger down-sampling rate, as well as a residual connection version of the deeper U-Net model was built. For the other three categories, the performance of the base U-Net model, the dilated version of the base U-Net and the object detection integration model was tested. Finally, the combined model was compared with existing state-of-the-art semantic segmentation models including DeepLab-v3plus (Chen et al., 2018), HRNet-v2 (Wang et al., 2020), PSPNet (Zhao et al., 2017), and SegNet (Badrinarayanan et al., 2017). DeepLab-v3plus and HRNet-v2 are two top performing methods in the Cityscapes semantic segmentation challenge (Cordts et al., 2016). PSPNet and SegNet were introduced in the literature review as two models used for urban scene segmentation tasks and widely used in the facade segmentation community.

To keep the detail of the images as high-quality as possible, and considering limitations to available computational resources, the input images were rescaled to 1024×1024 pixels. After the magnifier extracts image patches, each patch is re-scaled to 512×512 pixels before being fed into the smaller category models. A data augmentation technique was used during model training: geometric transformations and colour adjustments were applied to the base dataset to produce a larger training set. Horizontal mirroring, vertical and horizontal translations and small rotations were applied randomly to 50% of the data. The hue of the images was adjusted randomly by up to 10%.

The adaptive moment estimator (Adam) optimiser with a learning rate reduction strategy was used (Kingma and Ba, 2014). The minimum learning rate was set to 10^{-5}. To prevent overfitting, early-stopping is applied, stopping the training process if validation loss does not decrease for 30 epochs. The maximum number of training epochs was 500 for all models; typically, models took fewer than 200 epochs to train. All convolution layers in all segmentation models were initialised with Kaiming distribution (He et al., 2015).

Categorical models' performance is evaluated both qualitatively and quantitatively. Qualitative evaluation is based on visual inspection, and the quantitative evaluation includes the use of a confusion matrix and comparative evaluation metrics on component models. Accuracy, precision, recall, true negative rate (TNR), intersection-over-union (IoU) and the F1 score are used to indicate the quality of models. Each of these metrics relies on true positive, true negative, false positive and false negative numbers for each image. The true positive and negative represent the pixel quantities which are correctly predicted by the model, and the converse count incorrectly classified pixels. Accuracy denotes the percentage of correct classifications; precision measures the percentage of correct positive samples in all positive predictions; recall is a measure of the correct positive predictions over all positive samples; TNR measures a model's ability to correctly classify negative samples; and IoU measures the overlapping ratio of the positive predictions and the positive samples. Finally, the F1 score is widely used to measure the overall model performance by considering the impact of both the precision and recall values. The ensemble model is evaluated using the multi-class confusion matrix and visual inspection of the combined masks.

In this work, all code was written in Python, with all deep neural networks implemented with the TensorFlow library (Abadi, 2016). All models were trained on a workstation with Windows 10, 16GB RAM, an Intel Xeon E5-1620 v4 CPU and an NVIDIA Quadro P5000 GPU.

4.2.3.2 Object Detection Model Evaluation

The performance of the object detection model is assessed both quantitatively and qualitatively. The most common evaluation metric for an object detection task, average precision (AP50), is adopted here. The calculation of the metric varies in different guidelines (Everingham et al., 2010; Lin et al., 2014). This thesis employs the Microsoft COCO guideline (Lin et al., 2014) to calculate the AP50 metric. The AP50 values for the window, door and chimney categories have achieved 63.4%, 59.5% and 79.2%, respectively. The result is competitive in using YOLO-v3 model in a similar task (Kong and Fan, 2020). Qualitative assessment is demonstrated in Figure 4.5.

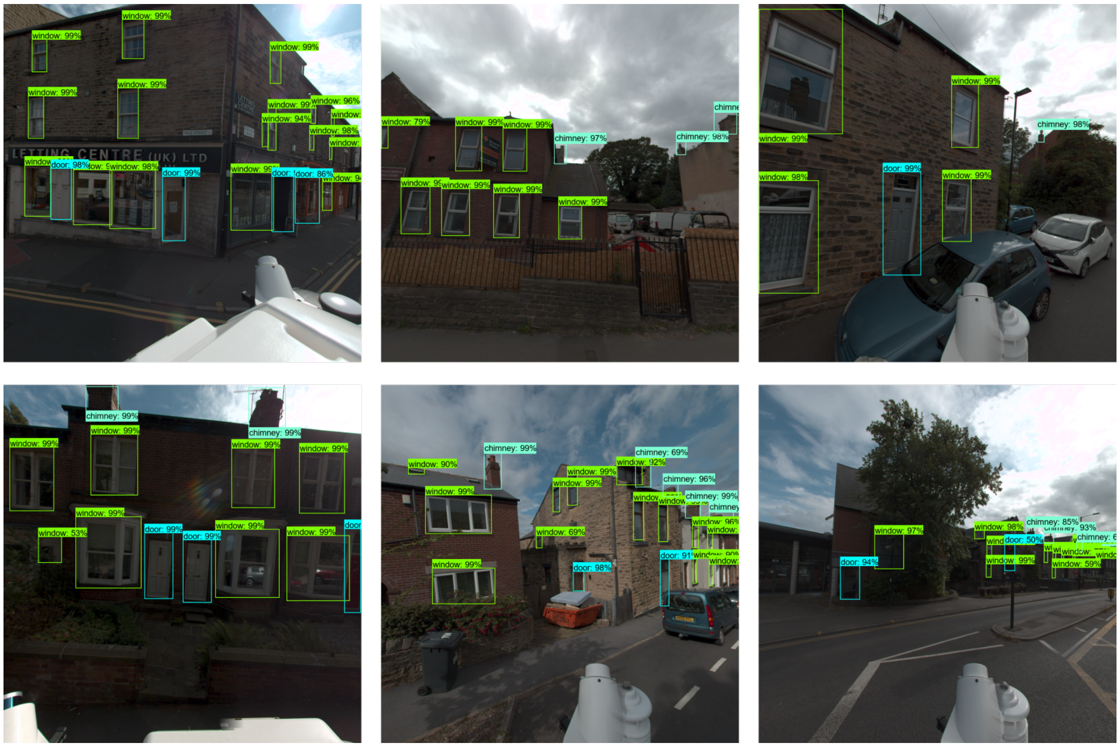


Figure 4.5: *The qualitative assessment examples of the object detection model. The figure clearly shows that target objects can be detected with high confidence levels.*

4.2.3.3 Categorical Model Evaluation

Metrics for the small component segmentation models are given in Table 4.2. Both the proposed model architectures outperform the base U-Net structure. Looking at the F1 score and IoU metrics, it is found that the proposed integrated magnifier model performs particularly highly for the chimney and door categories. The dilated centre block models, without magnification, tend to show higher precision value and lower recall value than with the magnifier. Since the denominator of the recall metric is a constant in predictions of the same image, this phenomenon indicates that the magnifier integration model tends to predict more positive pixels. Moreover, the results show that the accuracy and TNR metrics both have high values across different models due to the robust capabilities of all models of predicting negative samples and the highly imbalanced dataset. However, the two metrics are not suitable for comparing model performances in this task.

Table 4.2: *Smaller-sized categories' segmentation performance; The 'U-Net 32' is the base U-Net with a downsampling rate of 32. The 'Dilated U-Net 32' is the base U-Net with the dilation centre block and the 'With Magnifier' is to integrate the Faster-RCNN into the dilated U-Net 32 model.*

	Model	Accuracy	Precision	Recall	TNR	IoU	F1 score
Chimney	U-Net 32[%]	99.86	83.91	82.20	99.94	71.01	83.05
	Dilated U-Net 32[%]	99.89	**90.52**	81.24	**99.97**	74.87	85.63
	With Magnifier[%]	**99.90**	89.59	**85.12**	99.96	**77.46**	**87.30**
Door	U-Net 32[%]	99.53	82.65	64.83	99.87	57.06	72.66
	Dilated U-Net 32[%]	99.59	**89.61**	64.87	**99.93**	60.33	75.26
	With Magnifier[%]	**99.61**	81.93	**76.50**	99.84	**65.46**	**79.12**
Window	U-Net 32[%]	99.43	93.79	91.78	99.75	86.52	92.77
	Dilated U-Net 32[%]	**99.51**	**95.44**	92.18	**99.82**	**88.30**	**93.78**
	With Magnifier[%]	99.42	91.23	**94.60**	99.62	86.71	92.88

The qualitative analysis demonstrates the same overall results. Examples of segmentations are shown in Figure 4.6 and Figure 4.7 for the detection of doors and windows, respectively. The magnifier integration model generally shows better performance in handling boundaries and small objects.

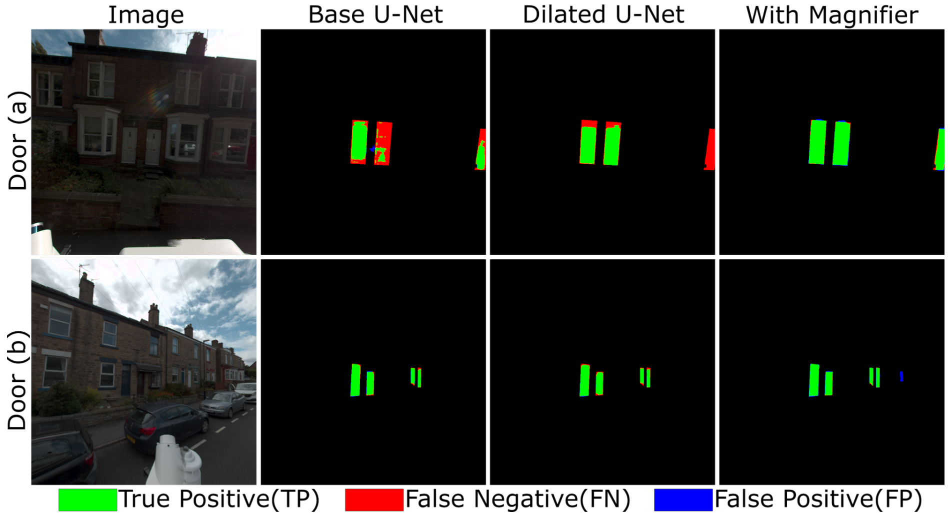


Figure 4.6: *Door qualitative examples; (a) clearly shows the performance improvements and (b) shows the object detection integration model predicting the object without annotation.*

For the window category, the F1 and IoU show distinct improvements in using the dilated centre block but only very minor refinements with the magnifier integration model compared to the base U-Net. Figure 4.7 demonstrates that the magnifier integration can generate more precise boundaries. However, as the model tends to classify glazing surfaces belonging to buildings as windows, such as the solar panel in Figure 4.7(b), and these kind of surfaces commonly exist on building facades, the tendency lowers its overall quantitative performance.

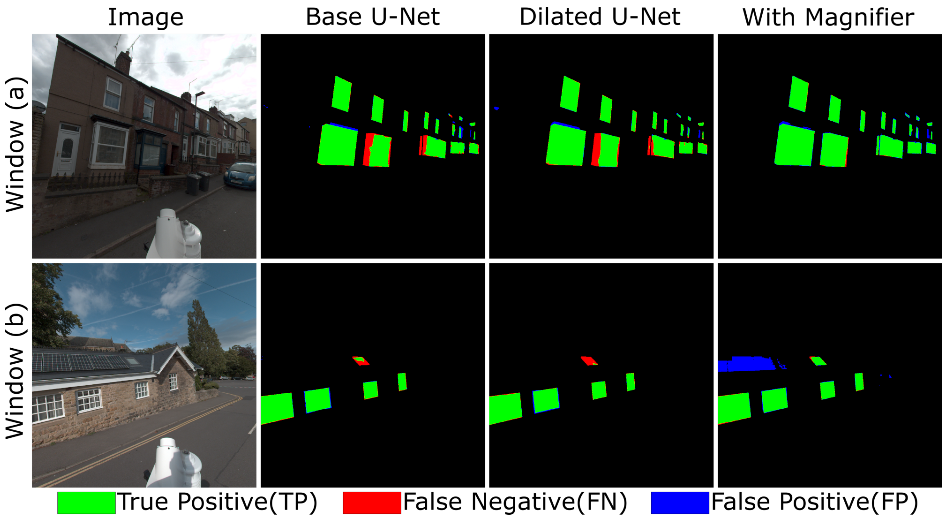


Figure 4.7: *Window qualitative examples; (a) shows the performance improvements and (b) shows the object integration model improving the performance but also recognises the solar panel as a window.*

The evaluation metrics for the roof and wall categories with different downsampling rates are shown in Table 4.3. The results for the roof category show that, with the higher downsampling, the overall performance drops. Although the use of residual blocks can improve the performance, the base U-Net model still performs highly. However, the results in wall classification show that the residual central block performs much better than the base U-Net, regardless of the downsampling rate. The quantitative analysis shows that the roof base U-Net model can produce more coherent predictions, and the residual model of the wall category is more friendly to boundary predictions. Visual examples of this are shown in Figure 4.8.

Table 4.3: *Roof & Wall segmentation performance; The 'U-Net' '64' and '128' means using the U-Net structure with downsampling rates of 64 and 128, respectively. 'Residual' means using the residual blocks across the model.*

	Model	Accuracy	Precision	Recall	TNR	IoU	F1 score
Roof	U-Net 32[%]	**99.48**	**92.74**	**89.65**	**99.78**	**83.77**	**91.17**
	U-Net 64[%]	99.30	91.10	84.88	99.74	78.39	87.88
	Residual U-Net 64[%]	99.42	92.20	88.14	99.77	82.02	90.12
Wall	U-Net 32[%]	97.16	93.00	**94.25**	97.98	88.01	93.62
	U-Net 128[%]	96.63	92.87	91.84	97.99	85.78	92.35
	Residual U-Net 128[%]	**97.60**	**95.31**	93.78	**98.69**	**89.64**	**94.54**

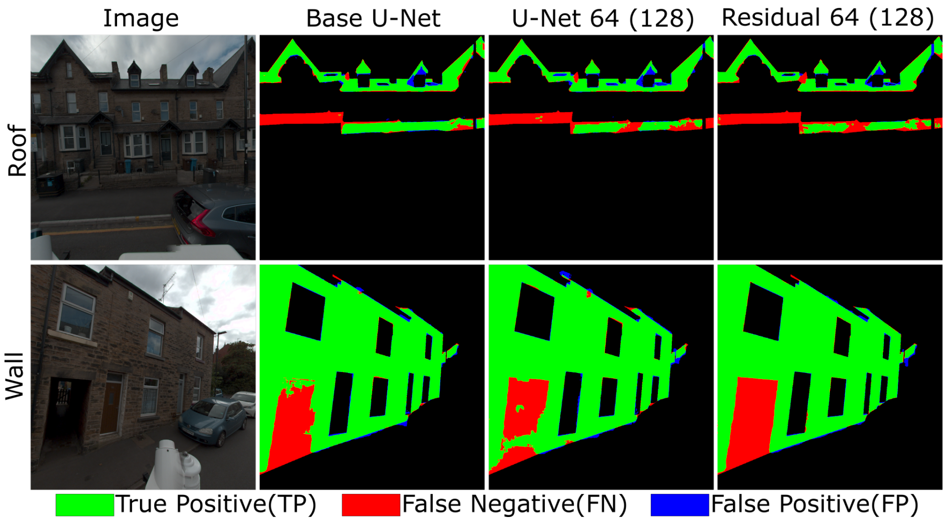


Figure 4.8: *Roof and Wall qualitative examples, the number in brackets indicates the down-sampling ratio of the wall category model; the roof category is more suitable for the base model and the residual model is more friendly to the wall category.*

4.2.3.4 Ensemble Model Evaluation

Based on the findings in the evaluation of each categorical model, the model ensemble, FacMagNet, uses the magnifier integration model for the three smaller-sized categories, because of its advantages in boundary and small-object predictions. The base U-Net for the roof category and the proposed residual U-Net 128 for the wall were selected due to sharing the highest IoU and F1 values in each category.

Figure 4.9 shows the produced multi-class confusion matrix of FacMagNet trained on the dataset. The confusion matrix shows that the window and the wall achieves the highest accuracy and the door is the lowest. Most of the considerable errors are caused by wrongly classifying pixels belonging to objects as background. Walls are the category to which the model will incorrectly assign pixels second-most often.

The proposed FacMagNet was compared using other models which are widely adopted in the semantic segmentation area, the categorical IoU values of each model are shown in Table 4.4. The mean IoU (mIoU) is calculated by computing the IoU average of all classes excluding the background.

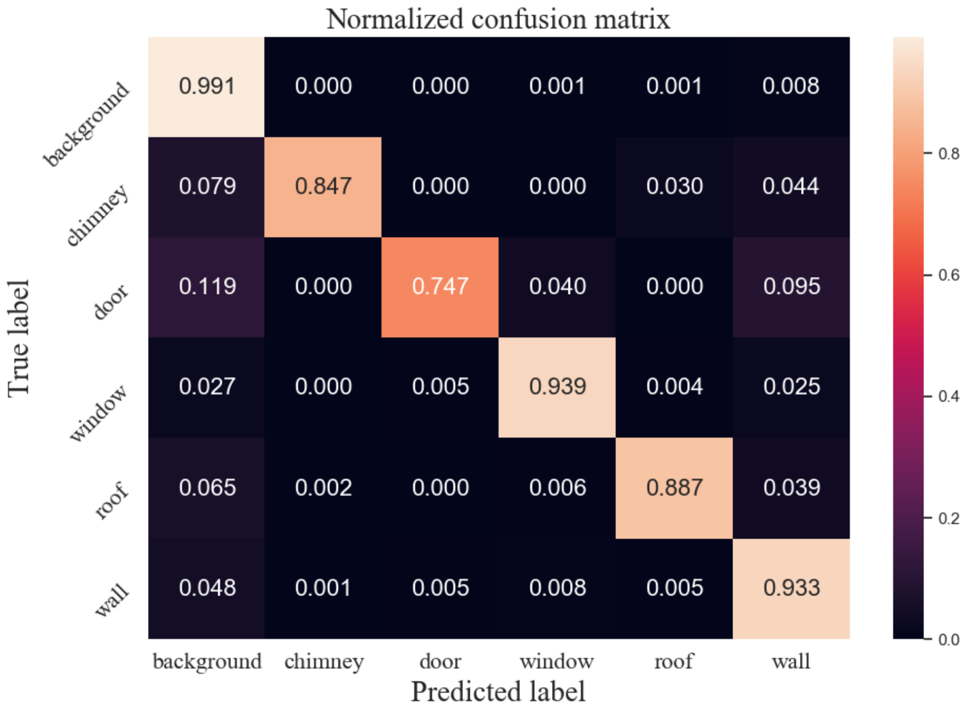


Figure 4.9: *Normalised ensemble model confusion matrix, the confusion matrix is normalised by dividing the sum of the ground-truth pixels in each category, the diagonal shows the percentages of the correctly predicted pixels over the sum of corresponding ground-truth pixels, i.e. precision values.*

Table 4.4: *Categorical IoU, the first row is the metrics using the developed FacMagNet, the second row is using the base U-Net with ensemble strategy. The third, fourth, fifth, sixth and seventh rows are using DeepLab-v3plus, HRNet, U-Net, PSPNet, and SegNet models, correspondingly, as multi-class classifiers. It is noted that DeepLab-v3plus is initialised by training on Cityscapes (Cordts et al., 2016) dataset. While other benchmark models are initialised by Kaiming initialisation (He et al., 2015).*

IoU% Model	Chimney	Door	Window	Roof	Wall	mIoU
FacMagNet[%]	**77.90**	**65.82**	**87.83**	**83.62**	**89.87**	**81.01**
Ensemble U-Net[%]	72.52	56.52	86.60	83.57	88.39	77.52
DeepLab-v3plus[%]	75.84	59.74	84.30	79.17	85.86	76.98
HRNet[%]	73.52	56.98	80.00	74.01	82.33	73.37
U-Net[%]	74.40	51.33	76.66	68.04	77.57	69.60
PSPNet[%]	59.85	48.67	73.46	66.86	72.53	64.27
SegNet[%]	54.01	39.97	57.67	36.58	65.32	50.71

From Table 4.4, it is clear that the designed FacMagNet model performed highest across all categories. FacMagNet's largest improvements were in the chimney and door categories, when compared to the other models. The table shows the benefits of applying the magnifier strategy and designing model structures for each facade component class: the mIoU of the FacMagNet is 3.49% higher than the ensemble U-Net model. Figure 4.10 shows that FacMagNet visually achieves a high performance segmentation, even when dealing with high-distortions, small-objects and obstacles. It can also be seen from Figure 4.10 that the designed approach can easily handle segmentation, even when the components are partially occluded by objects such as trees and fences.

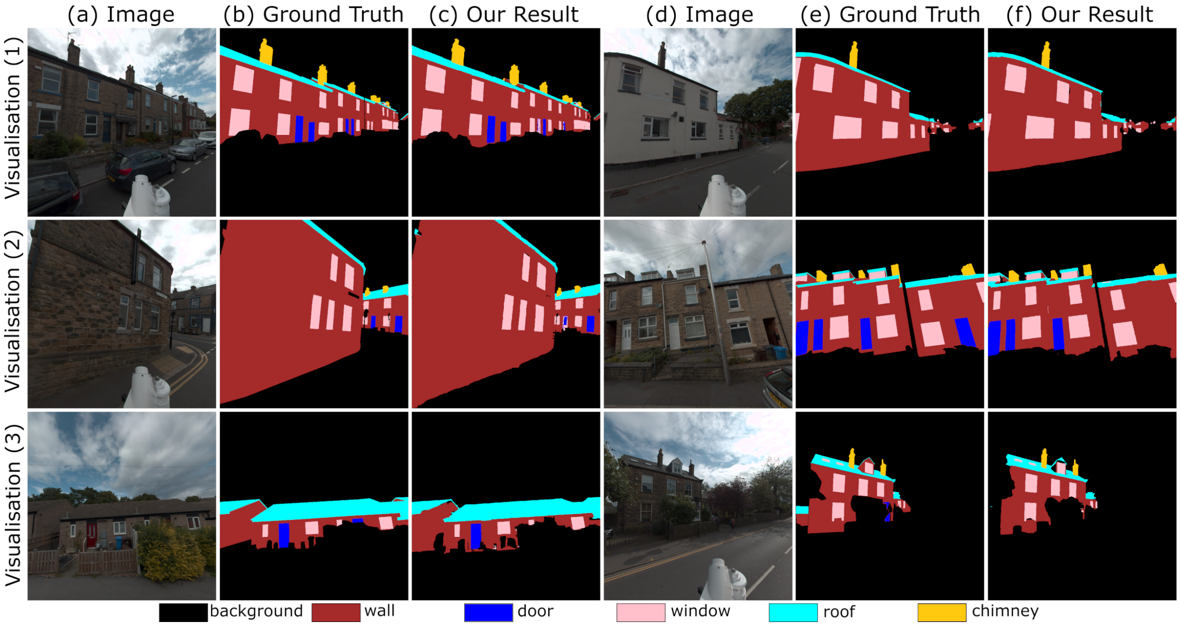


Figure 4.10: *Qualitative examples of the FacMagNet model; the visual results show the model has achieved high accuracy in large objects, and is friendly in handling small-object and occlusion problems.*

4.3 Rethinking the Residential Building Facade Segmentation

4.3.1 Rethinking the FacMagNet Model

This section re-develops the FacMagNet model by using the trained DeepLab-v3plus model as the stem and using the trained door model as a refinement branch. This can significantly reduce the complexity of the developed FacMagNet.

From the experiment results of the developed model FacMagNet, the model has achieved the SOTA performance on the built Crookesmoor dataset. However, the model has obvious drawbacks such as its efficiency and usability. The FacMagNet model is hard to train as it contains six individual models and thus it will take longer to predict an image in comparison with other end-to-end semantic segmentation models. In addition, training deep learning models takes a considerable amount of resources including electricity and time. The FacMagNet contains six individual deep learning-based models and thus requires a significant amount of resources during training. These drawbacks should not be heavily problematic as the model only needs training once and the inference stage is fully automatic. However, these drawbacks could limit the model to be promoted to large-sized datasets. Therefore, it is necessary to develop a more efficient residential building facade segmentation model.

Recalling the experiments conducted to develop the categorical semantic segmentation models for the three smaller-sized categories in Table 4.2, the magnifier strategy has increased 8.4% in IoU on the door category in comparison with a vanilla U-shape model. However, for the chimney and the window categories, implementing the magnifier strategy does not make a striking increase on the evaluation metrics. Therefore, it is unclear whether implementing the magnifier module on the window and the chimney categories is necessary , especially if a more efficient model is required.

The designed magnifier module is mainly based on an object detection model, Faster R-CNN (Ren et al., 2015). In the magnifier module, the raw images are fed into the Faster R-CNN first to detect the target objects with bounding boxes. The model contains a convolutional neural network which is to generate feature maps and a region proposed network (RPN) to propose anchor boxes. In the two-stage process, the output bounding boxes are scaled and adjusted to a square shape first. Raw images are sliced based on the adjusted bounding boxes, then the sliced image patches are fed into the semantic segmentation model to generate pixel-level predictions. At the end, patches are spliced together. The two-stage design makes the model become complicated in the training and inference stage.

As reviewed in chapter 2, attention mechanism in CNN is a method which simulates the human attention mechanism to improve the CNN performance. The CNN attention mechanisms have two different types: the hard attention (Mnih et al., 2014) and the soft attention (Woo et al., 2018). The hard attention mechanism is to give the desired region credits and the soft mechanism impacts on feature maps through the CNN learning process. In this attention mechanism facet, the magnifier module can be regarded as a hard attention mechanism method. The module uses the object detection model to find the target area in an image and then focuses on the target area to find the targets. Also, from the visualised experiment results such as in Figure 4.7 and Figure 4.6, two capabilities of the magnifier module are identified: 1. the module can refine the boundaries of detected target objects, 2. the module can detect some objects which failed to be found by directly applying the semantic segmentation models. These findings validate the success of the magnifier module as a hard attention mechanism.

In the contents above, the necessity of developing a light-weight building facade semantic segmentation model, the compromise which can be made to improve the efficiency of the FacMagNet and the essence of the magnifier module which is a type of hard attention mechanism are discussed. Based on these discussions, a reduced version of the proposed building facade semantic segmentation model, FacMagNet is proposed. The model is designated as FacMagNet-s; the 's' is the abbreviation of small and the developed full-size FacMagNet is renamed as FacMagNet-l. The FacMagNet-s is still a two-stage model but it only contains two deep learning-based semantic segmentation models. The FacMagNet-s is shown below in Figure 4.11. The model contains a multi-class facade segmentation model acting as a base model and a door enhancement module. The multi-class facade segmentation model predicts the raw image of all classes of objects first. In the benchmark test results shown in Table 4.4, the DeepLab-v3plus model has shown powerful performance in comparison with other single-stage models. Specifically, it is easy to find it performs much better than its single-stage competitors in the door prediction. Therefore, it is chosen here as the base model.

The door objects in the predicted mask are extracted and processed with morphological operations. The morphological operation is a series of image processing algorithms for removing noise. In the FacMagNet-s model, the open and the closed operations are applied. The morphological open operation is the dilation of the erosion operation (Dougherty, 1992). The erosion operation is to use a kernel scanning a binary image, and if the scanning area is completely contained by the foreground, the pixel in the centre of the scanning area will be kept, otherwise it will be deleted. The dilation operation will expand the foreground area if the locus of the kernel is inside the foreground. The operation of using the open then the closed operations helps to remove the noises of the predicted door masks.

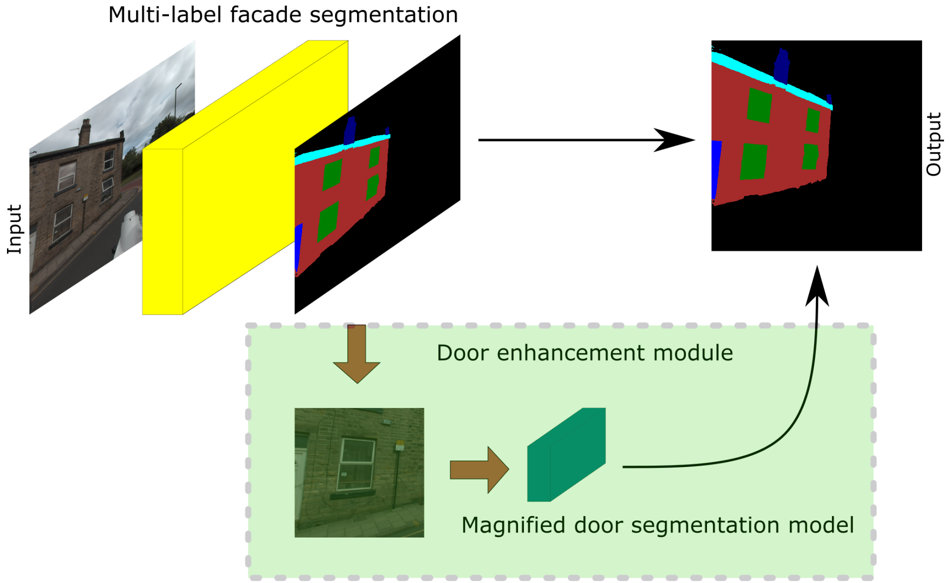

Multi-label facade segmentation
Input
Output
Door enhancement module
Magnified door segmentation model

Figure 4.11: *The developed FacMagNet-s model, the model first uses a single semantic segmentation model to generate multi-class prediction, then the door object is separated and processed with morphological operations. The bounding box of the processed mask is calculated and expanded as in the FacMagNet. The scaled bounding box area is then sliced and fed in to the door semantic segmentation model. The prediction from the door model is copied back to the multi-class prediction mask.*

The bounding boxes of the processed door masks are then calculated, expanded and reshaped as in the FacMagNet-l. The altered bounding box information is used to slice the image patches from the raw images. The process inherits the idea of the magnifier module but without the object detection model. Therefore, the module is potentially able to refine the predicted objects. Thus, the module is called the door enhancement module. The image patches are fed into the door model illustrated in Figure 4.2. The predictions of the door model are merged together with the prediction of the base model by an ordinary voting strategy to generate the final output.

4.3.2 Experiments and Results

FacMagNet-s contains two convolutional neural network models which need training. DeepLab-v3plus was chosen as the base model which is initialised by training it on the Cityscapes dataset (Cordts et al., 2016), and the categorical model developed in the FacMagNet-l was chosen as the enhancement model. Since the two models have already been trained on the same dataset, they are implanted in the FacMagNet-s. Therefore, the training process is exactly the same as in the FacMagNet-l. In the FacMagNet-s model, the morphological close and open operations are employed for noise removal. The structured element sizes of the operations are determined to be 30 and 5, respectively. The values are determined based on multiple tests and visual inspections.

Table 4.5 shows the quantitative comparisons of the FacMagNet-s model with FacMagNet-l model and other models in Table 4.4. The intersection-over-union (IoU) is still chosen as the evaluation metric as in Table 4.4. As shown in Table 4.5, FacMagNet-s model has achieved the second-best result, the most accurate model is still the FacMagNet-l. The model efficiency is analysed using FLOPs (Floating Point Operations), trainable parameters and training time per epoch are shown in Table 4.6 [1]. In comparison with FacMagNet-l and -s, FacMagNet-l model is 3.14% higher than -s in mIoU while it requires significantly more resources in training: nearly triple trainable parameters and significantly higher FLOPs. Therefore, considering the two models' performance, training and future deployment costs, FacMagNet-s is a more efficient choice than -l, which is the development target of FacMagNet-s. Ensemble U-Net achieves similar performance on mean IoU. However, ensemble U-Net contains five models which is also larger than FacMagNet-s in size and has higher a computational cost. Besides, Ensemble U-Net is 7.46% lower than FacMagNet-s in door predictions. By comparing the results of the FacMagNet-s and the DeepLab-v3plus which is acting as the base model of the FacMagNet-s, it is explicit that the door enhancement module has achieved excellent results. The door metric is raised by 4.24%. The enhancement module also slightly raises the performance of the DeepLab-v3plus on window and wall classes' predictions.

Table 4.5: *Categorical IoU, the first and the second row are the metrics of using our developed FacMagNet series model, the third row is using the base U-Net with ensemble strategy. The fourth row is using the DeepLab-v3plus as multi-class classifiers.*

IoU% / Model	Chimney	Door	Window	Roof	Wall	mIoU
FacMagNet-l[%]	**77.90**	**65.82**	**87.83**	**83.62**	**89.87**	**81.01**
FacMagNet-s[%]	75.84	63.98	84.40	79.17	85.99	77.87
Ensemble U-Net[%]	72.52	56.52	86.60	83.57	88.39	77.52
DeepLab-v3plus[%]	75.84	59.74	84.30	79.17	85.86	76.98

[1] TensorFlow Profiler tool is used for the trainable parameters and FLOPs analysis. However, the tool experienced a problem during processing Faster R-CNN model. Therefore, its statistics are estimated using its backbone network Inception ResNetv2 whose statistics are available at: `https://paperswithcode.com/model/inception-resnet-v2?variant=inception-resnet-v2-1`

Table 4.6: *Comparison between demanded computational resources of developed models. It is noted that, as FacMagNet series models utilise ensemble learning strategy, their trainable parameters and FLOPs are the sum of all their individual models. Their training time per epoch is also the sum of their individual models.*

Metric Model	Trainable parameters	FLOPs (floating point operations)	Approximate training time (minute/epoch)
FacMagNet-l	294.79M	17.5G	250
FacMagNet-s	100.47M	200.93M	63
Ensemble U-Net	240.51M	480.95M	53
DeepLab-v3plus	41.05M	82.10M	35
Dilated U-Net	59.42M	118.83M	13
U-Net32	31.11M	62.21M	7
ResU-Net128	31.13M	62.25M	7

Figure 4.12 shows the qualitative evaluation of FacMagNet-s. The figure shows the FacMagNet-s has achieved accurate predictions in all examples. In comparison with FacMagNet-l, FacMagNet-s sacrifices some boundary precision and consistency in wall and roof predictions. The same defect is also observed in window and door inference but is less obvious. The chimney qualitative prediction has achieved nearly the same performance as FacMagNet-l. In all, FacMagNet-s has achieved competitive accuracy in the building component inference task with higher efficiency than FacMagNet-l.

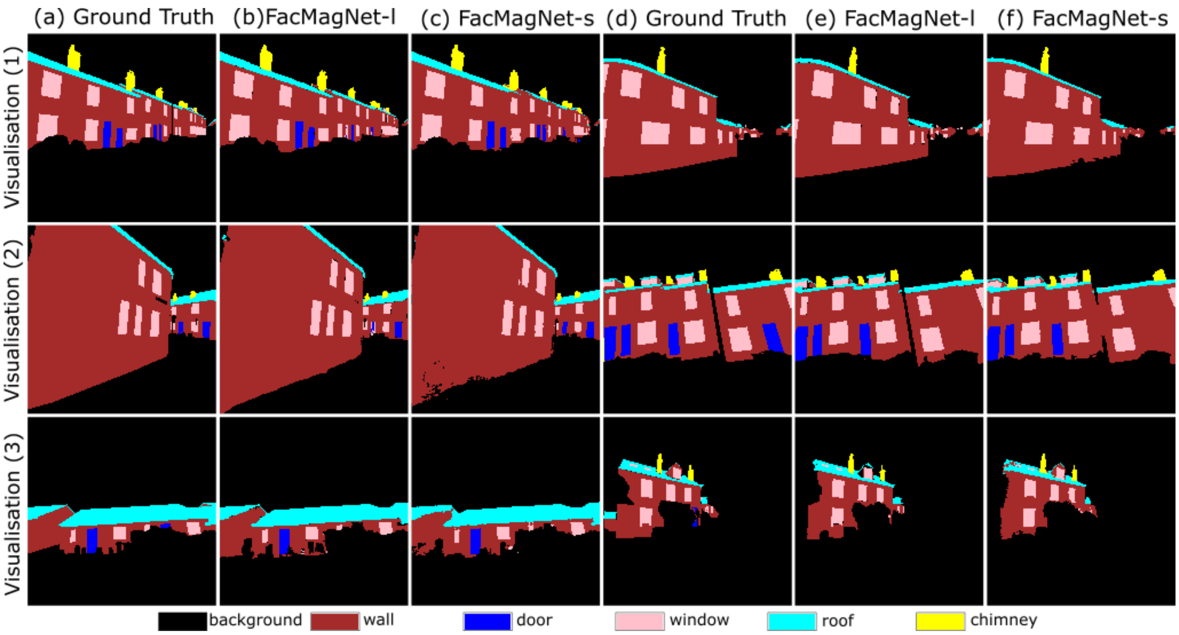


Figure 4.12: *The visualisation masks inferred by the FacMagNet-s. These examples are compared with the results from the FacMagNet-l and the ground truth masks.*

4.4 Loss Function and Model Architecture Studies

4.4.1 Residential Building Semantic Segmentation on Handsworth Dataset

In the previous sections of this chapter, two residential building facade semantic segmentation models were developed, which are FacMagNet-l and FacMagNet-s based on the Crookesmoor dataset. The key motivation of designing FacMagNet-l is that building components show both large inter- and intra- varieties on their sizes and shapes. FacMagNet-l exploits models built on U-Net (Ronneberger et al., 2015) structure with different numbers of down-sampling layers to counter the inter- size difference problem. A novel magnifier module is developed to counter the intra- size difference problem. The developed magnifier module acts as a hard attention mechanism. The main findings in developing the FacMagNet-l model include the feasibility of U-Net structure in the residential building recognition task and the designed magnifier strategy can lead to a performance rise.

FacMagNet-s is designed to tackle the door prediction difficulty and improve efficiency. The FacMagNet-s model focuses on reducing the computational cost of FacMagNet-l by shrinking the six-model-made FacMagNet-l to two models. DeepLab-v3plus is the foundation of FacMagNet-s which has also achieved the highest accuracy among other end-to-end models tested while developing FacMagNet models. FacMagNet-s employs door predictions from the end-to-end DeepLab-v3plus and a set of morphology operations to replace the object detection model. FacMagNet-s shows competitive results in comparison with FacMagNet-l while largely reducing the model's magnitude and computational cost. Although the U-Net structure has shown capability in binary building components classification in section 4.3, for example, Table 4.5 and 4.6 show that using smaller U-shape models, i.e. U-Net32 and ResU-Net128 on binary classification can achieve better performance than using a single larger model, i.e. DeepLab-v3plus, FacMagNet-s does not further explore the feasibility of the U-Net structure on the multi-class building components identification task.

The Handsworth dataset is more than five times larger than the Crookesmoor dataset and also has over ten times more door objects than the Crookesmoor dataset. Therefore, the development of an end-to-end semantic segmentation model is vital. The Handsworth dataset also has more complex scenes than the Crookesmoor dataset as accounted in table 3.7. In addition, annotations of the Handsworth dataset are commonly fragmented which deteriorates component shape features. In this case, the effect of magnifier strategy is unknown. In this study, three sets of experiments were designed to explore a suitable semantic segmentation model for the Handsworth dataset. Experiments were designed based on findings on developing FacMagNet series of models on the Crookesmoor dataset.

The first experiment was designed to explore the effect of focal loss (Lin, Goyal, Girshick, He and Dollár, 2017) on identifying building components. Focal loss is designed to add weight to categories which are hard to classify:

$$L_{focal} = -\frac{1}{N}\sum_{i}^{N}[(1-y_i)^{\gamma} y_i \log \hat{y}_i + y_i^{\gamma}(1-y_i)\log(1-\hat{y}_i)], \tag{4.1}$$

where $\gamma \geq 0$ and when $\gamma = 0$, the focal loss is equal to the cross entropy.

Based on the experiments taken on developing FacMagNet models, the five defined categories are not equally difficult to classify: door is the hardest category and wall and window are relatively simpler. Therefore, testing whether focal loss can improve inference performance on accuracy is valuable. In the first experiment, DeepLab-v3plus (Chen et al., 2018), U-Net (Ronneberger et al., 2015) and U-Net with residual connections (Chu et al., 2019) were adopted. The adopted U-Net is the original version developed in 2015. The residual connection can effectively resolve the gradient vanishing problem which is common in training a 'very deep' deep learning model (He et al., 2016*a*). The adopted ResU-Net is built based on the original version with two extra downsampling layers and each convolution layer in the original version U-Net is replaced by a residual module. The ResU-Net model structure is demonstrated in Figure 4.13. The first set of experiments is to find the answer to whether focal loss is more appropriate as a loss function. Furthermore, it also explores whether, in comparison with DeepLab-v3plus, the U-Net architecture is more suitable for the building component identification task.

The second experiment is to further explore a suitable model under the U-Net architecture for the building component identification task. In this set of experiments, various models under the ResU-Net architecture, which is used in the first experiment set, is built. This set first explores the optimum number of downsampling layers by testing models under the ResU-Net architecture with 4, 5, 6 and 7 max pooling layers. The number of downsampling layers is significant in designing deep learning models: insufficient downsampling layers may lead to inadequate feature extractions. In developing roof and wall models, the experiment in table 4.3 has shown an optimised number of downsampling layers can result in up to 5.38% of intersection-over-union (IoU) growth.

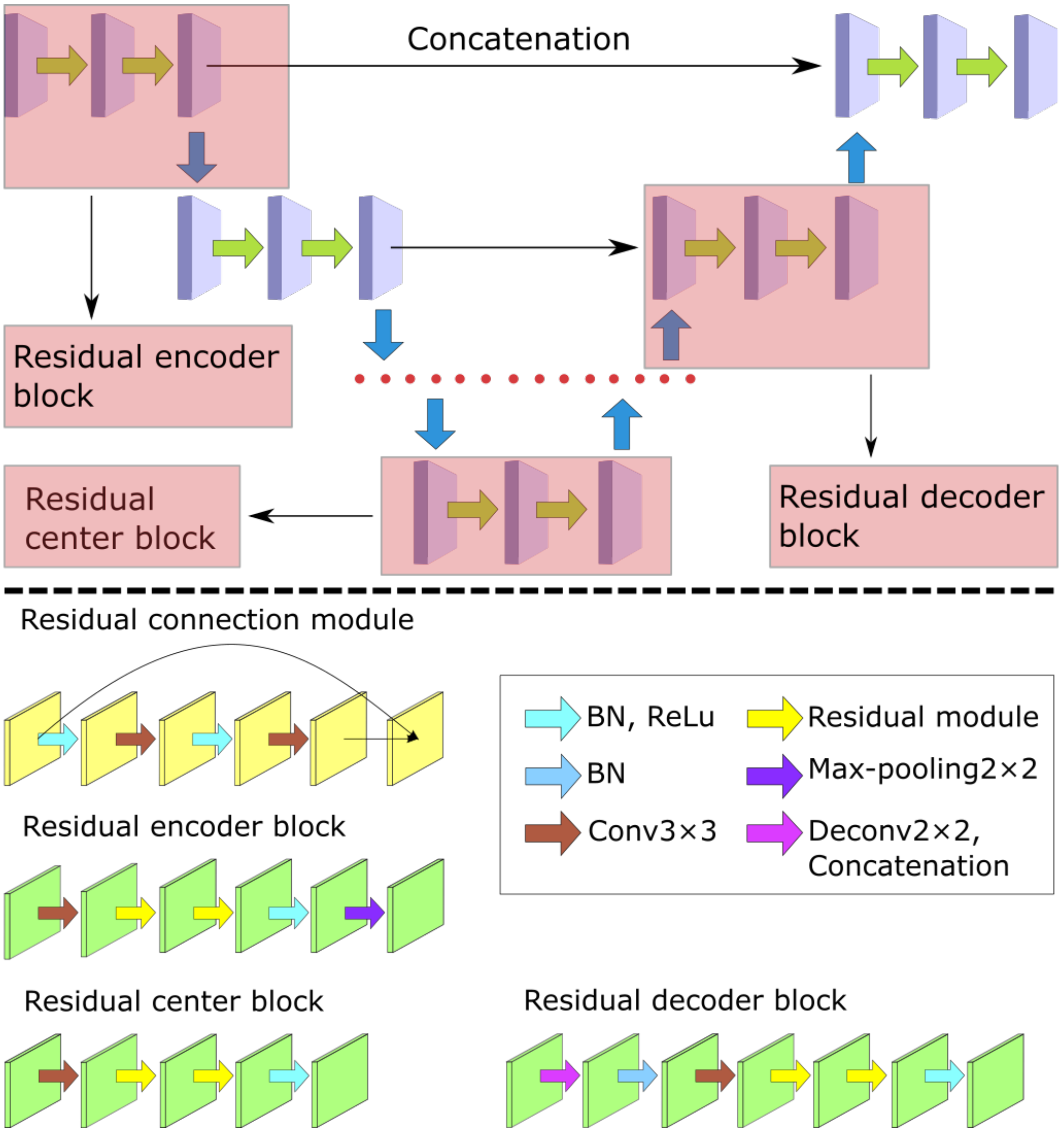


Figure 4.13: *The ResU-Net model architecture demonstration. The upper side which is separated by the long dashed line shows the architecture of the ResU-Net and the lower side shows the structure of every block and the residual connection. ResU-Net is to replace every convolution layer in a U-Net model with residual blocks. The residual block structure is shown in the figure. The structure contains a short connection between the first and last convolution layers of each block.*

As introduced in Chapter 2, attention mechanism in deep learning is inspired by the human vision system. Naturally, humans can effectively find salient regions in complex scenarios. The human vision system has two different attention mechanisms: the hard attention and the soft attention. In developing FacMagNet models, a hard attention mechanism module, magnifier, was proposed. The hard attention mechanism module has shown a distinct growth in detecting doors and other small objects in the Crookesmoor dataset. In boosting the deep learning model performance, soft attention has also become a focus over the past few years (Hu et al., 2018; Fu et al., 2019; Woo et al., 2018). The soft attention mechanism in the deep learning model is designed to be a dynamic and learnable weight adjustment process. Spatial-wise and channel-wise soft attention are two major soft attention types applied on visual tasks such as semantic segmentation and image classification (Guo et al., 2022). Spatial-wise attention is to generate attention masks on the spatial domain and emphasises salient regions. Channel-wise attention creates attention masks on the channel domain and select significant channels. The integration of both of the two soft attention types has also been widely studied. Soft attention modules are commonly used by inserting them directly into off-the-peg deep learning models. The flexibility makes the soft attention mechanism a more economic choice, rather than applying hard attention mechanisms. However, the soft attention mechanism has not yet been explored on the building component identification task in 2021.

In this study, two different soft attention modules, CBAM (Woo et al., 2018) and dual attention (Fu et al., 2019), were selected to test the feasibility of using soft attention mechanisms on the building component identification task. Both of these two attention modules are spatial-channel integrated. The CBAM modules have been implemented in models with the U-shape architecture previously used for medical image segmentation (Zhao et al., 2021). The CBAM module contains a cascading structure in which the channel attention mask is generated and applied to the input feature maps. Then the output feature maps are used to produce the spatial attention masks and applied. Unlike the CBAM module, the dual attention module has a parallel structure in which the spatial and channel attention masks are produced using the same input feature maps and then summarised together.

The position where the attention module is inserted is also critical. The CBAM module is initially implemented to refine feature maps from every convolution block in ResNet, however, this would lead to a high computational cost, especially as the Handsworth dataset has an image size 1024 × 1024. Therefore, in this task, two different implementations were designed: the first one is inserting the CBAM module after each concatenation layer shown in Figure 4.13; the second one is implementing the module after every last convolution layer in the encoder path. The second implementation approach has achieved success for tissue segmentation in medical imaging (Khanh et al., 2020). The fist method is because it is argued that the concatenation operation would lead to a substantial redundancy of feature maps, the attention module may filter these redundancies. The dual attention module is inserted in the centre block of the built ResU-Net model. The dual attention module can capture contextual information at long distance. Therefore, installing the module at the centre block may improve the model capability for identifying individual objects, since high-level feature maps contain a substantial amount of semantic information.

All experiments were implemented using TensorFlow (Abadi, 2016) library and trained on a workstation with Windows 10, 16 GB RAM, an Intel Xeon E5-1620 v4 CPU and an NVIDIA Quadro P5000 GPU. The adaptive moment estimator (Adam) (Kingma and Ba, 2014) optimiser was adopted with a learning rate reduction strategy starting from 10^{-3}, the minimum learning rate is set to 10^{-5}. To prevent overfitting, an early-stopping strategy is applied which will stop the training process if validation loss does not decrease in 10 epochs. The maximum epochs are set to 200 across all models. All convolution layers are initialised with Kaiming distribution (He et al., 2015).

4.4.2 Results and Discussion

Table 4.7 shows the experimental results of focal loss and model architecture tests. The models trained with the focal loss are advantageous over their reproductions trained with the cross entropy loss except for the vanilla U-Net. DeepLab-v3plus has achieved the highest scores among all end-to-end segmentation models including U-Net in the Crookesmoor dataset, however, it was observed that the U-Net model achieved an equivalent performance with the DeepLab-v3plus model in both cross entropy and focal loss experiments. The ResU-Net model with the focal loss has achieved the highest intersection-over-union (IoU) scores in every category, while not distinct. Thus, whether or not focal loss can contribute to this task is still ambiguous.

Model name	Loss	mIoU	Wall	Roof	Chimney	Window	Door
DeepLab_{v3}^{+}	CE	77.70%	83.92%	78.49%	79.90%	80.91%	63.76%
ResU-Net_{64}	CE	80.66%	86.58%	82.68%	84.37%	84.06%	65.58%
U-Net	CE	79.80%	85.78%	81.44%	83.22%	84.14%	64.41%
DeepLab_{v3}^{+}	FL	78.68%	85.21%	80.66%	82.37%	81.87%	63.31%
ResU-Net_{64}	FL	**81.24%**	**86.91%**	**83.13%**	**84.87%**	**84.33%**	**66.95%**
U-Net	FL	78.55%	84.90%	82.01%	83.57%	81.88%	60.39%

Table 4.7: *The focal loss feasibility test results. Models trained with focal loss have achieved more accurate results than their repetitions trained with cross entropy loss, except for the plain U-Net. Using focal loss has only shown inconspicuous benefits in this experiment.*

Figure 4.14 shows the qualitative comparisons of using the cross entropy loss and the focal loss. The figure also shows the qualitative comparisons of various model architectures. Five examples are selected in this figure. Example 0 is a sample of a complex scene including buildings orientated in different directions; example 1 is a sample of a semi-detached building; example 2 is a terraced building with occlusions; examples 3 and 4 are street scenes with less light over-exposures. Over-exposure is a serious problem in photography which will cause a region of an image taken in extremely bright light to lose information or detail. The over-exposure is common in the Handsworth dataset. The raw images in example 0, 1 and 2 were taken with strong over-exposures. However, there is no distinct evidence observed in the inference masks of all models applied that over-exposure would affect the model precision. Comparing the results of line 3 and 4, and 5 and 6, models trained with focal loss would generate predictions with more precise boundaries and less noise in these cases. However, U-Net and ResU-Net models can handle occlusions better than DeepLab-v3plus. This is very obvious in example 2, DeepLab-v3plus tends to predict the whole area with tree branches as occlusions which leads to a large region of the building area being avoided.

A bizarre situation is shown in example 3, when DeepLab-v3plus is trained by the cross entropy loss, the left-hand side of the building is ignored by the model. The situation is also observed in other inference examples in the test set. However, the reason is still unknown. Looking at the last three lines in example 1 of the figure, the U-Net model has achieved a better result predicting roofs and the ResU-Net trained with focal loss achieves the second best. This comparison shows that an appropriate number of down-sampling layers is essential in designing deep learning models for the building component identification task. In the Crookesmoor dataset, identifying small objects is a considerable problem and the strategy named magnifier is developed to tackle the problem. However, in all examples shown in Figure 4.14, small objects can be effectively recognised across all different models, even though the Handsworth dataset has more small objects. The only explanation of why the same model would achieve a better performance in the Handsworth dataset is the dataset scope. The Handsworth dataset has over five times more data than the Crookesmoor dataset which could provide a deep learning model more experience on learning features of small objects.

The model architecture experiments show ResU-Net is an appropriate model architecture in the building component identification task. However, there are still other numerous adjustments which can be made to refine a model, such as the model depth, model width and off-the-peg refinement strategies, for example attention modules. Table 4.8 shows the experimental results of exploring a predominant model refinement structure with the ResU-Net architecture.

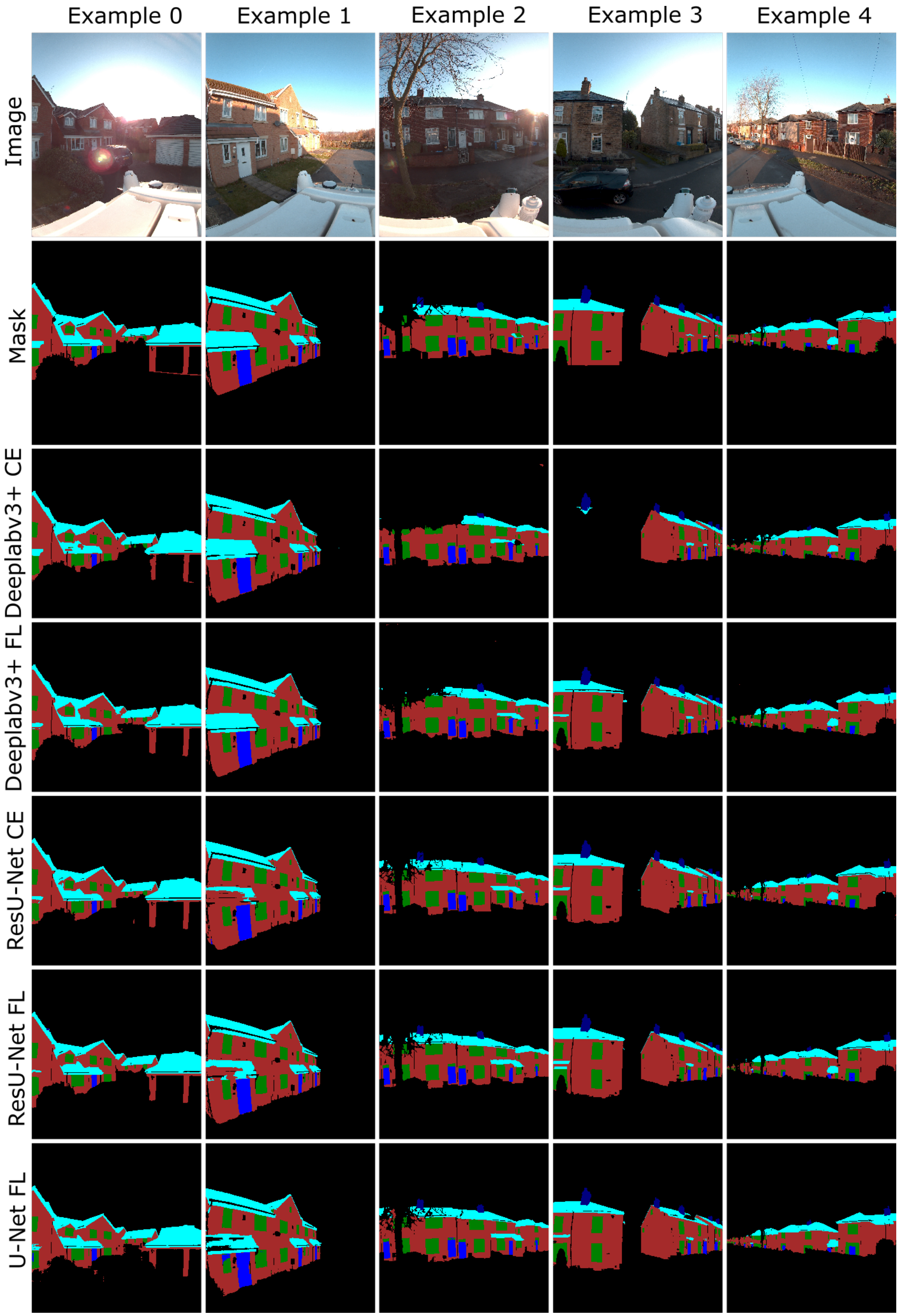


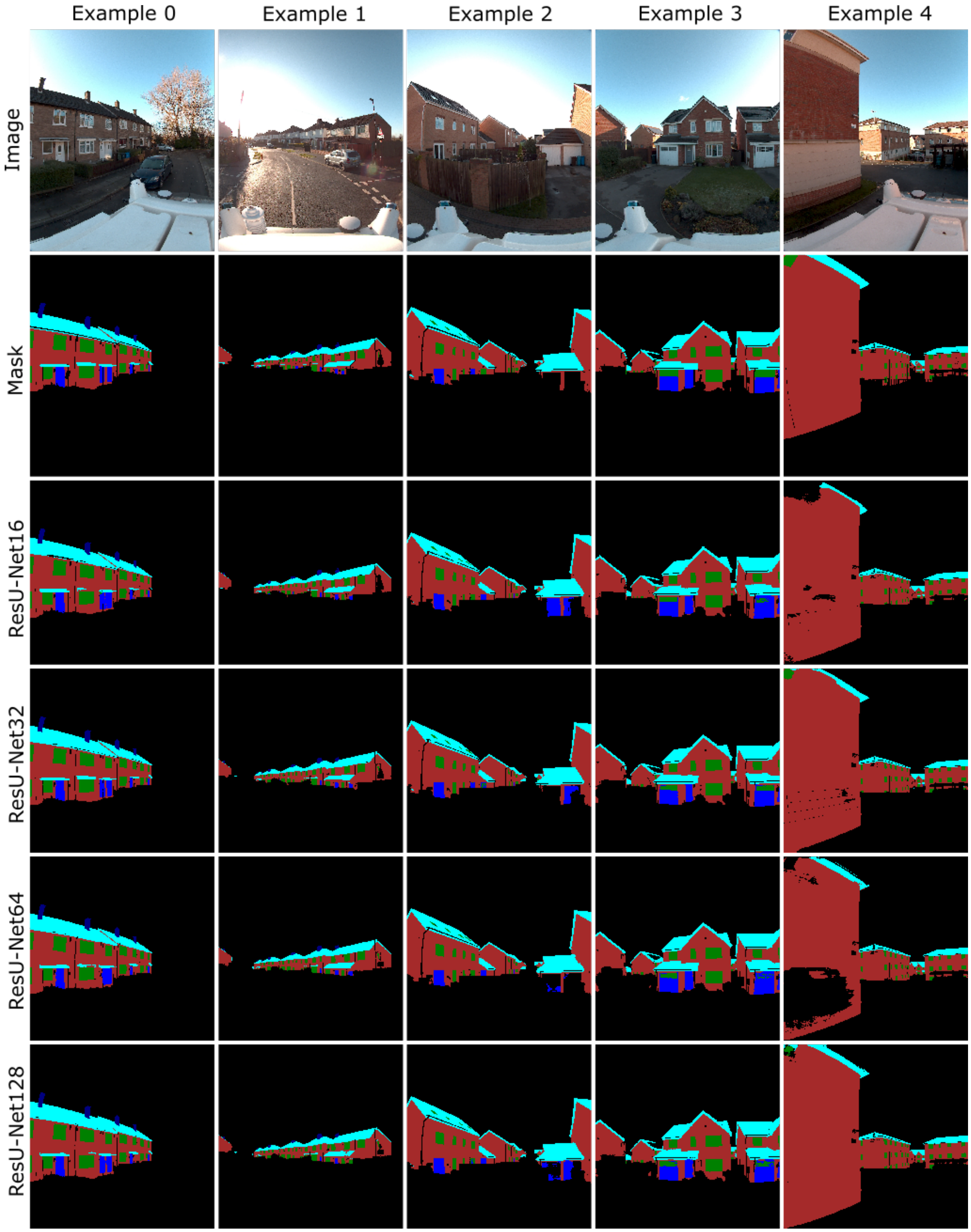
Figure 4.14: *Qualitative analysis of the model architecture and loss function experiments. Five examples with different scenarios and light conditions were selected for demonstration. The abbreviation CE (cross entropy) and FL (focal loss) indicates which loss function has been applied.*

Model name	mIoU	Wall	Roof	Chimney	Window	Door
ResU-Net$_{16}$	81.96%	87.82%	**83.83%**	85.86%	85.18%	67.12%
ResU-Net$_{32}$	**82.15%**	**88.19%**	83.48%	**86.03%**	**85.25%**	67.79%
ResU-Net$_{32}$+CBAM$_L$	81.90%	88.14%	83.65%	85.55%	85.21%	66.98%
ResU-Net$_{32}$+CBAM$_R$	81.99%	88.03%	83.69%	85.24%	85.04%	**67.95%**
ResU-Net$_{32}$+DA$_{CAM}$	81.99%	88.05%	83.76%	85.79%	85.18%	67.19
ResU-Net$_{64}$	81.24%	86.91%	83.13%	84.87%	84.33%	66.95%
ResU-Net$_{64}$+CBAM$_R$	81.04%	86.90%	82.65%	84.62%	84.29%	66.74%
ResU-Net$_{64}$+DA	80.96%	86.99%	82.39%	84.13%	84.16%	67.10%
ResU-Net$_{128}$	81.09%	86.68%	82.81%	84.36%	84.43%	67.14%

Table 4.8: *The ResU-Net based model refinement test. CBAM and DA represent models adding these two attention modules. The L and R show that the CBAM module is inserted on the encoder or decoder path of the Res-UNet, respectively.*

The quantitative results have shown that the ResU-Net models with five deconvolution layers are slightly more advantageous than other candidates. Two of the top three mean IoU values are observed in models with five deconvolution layers. Among the results of the top-three models, ResU-Net32 has shown the best performance: it has achieved the highest mean IoU and also gained the highest in three over five categorical IoU(s). Moreover, although attention modules have achieved great success in many other tasks and datasets, attention modules are only observed to deteriorate the model performance in the Handsworth dataset. The mIoU would decrease c.0.2% once an attention module is applied.

Figure 4.15 shows the selected qualitative results of the model structure refinement experiment. Observing the five examples on the first four models, ResU-Net32 has shown a more comprehensive and competitive performance over the other candidates: ResU-Net32 has generated the most smooth and integrated roof prediction in example 0; in example 1, it has produced the most accurate occlusion boundary; in examples 3 and 4, it can properly infer the out-of-the-ordinary windows which are on the garage doors (example 3) and highly distorted on the wall top (example 4). Other models have shown better predictions in some facets, for example, ResU-Net128 has achieved the most smooth prediction in walls on example 4. However, ResU-Net32 is still the most balanced choice among the four models. ResU-Net32-CBAM$_R$ has clearly deteriorated the prediction performance. It has generated a plurality of noises in examples 2 and 3. ResU-Net32-CBAM$_L$ does not show distinct benefits or defects compared to ResU-Net32. In the qualitative analysis study, ResU-Net32 is slightly more superior to other candidates. In addition, although the CBAM and dual-attention modules were reported to be beneficial to model inference capability, there is no evidence in this building component identification task on the Handsworth dataset showing that using attention modules is advantageous.


Example 0
Example 1
Example 2
Example 3
Example 4
Image
Mask
ResU-Net16
ResU-Net32
ResU-Net64
ResU-Net128

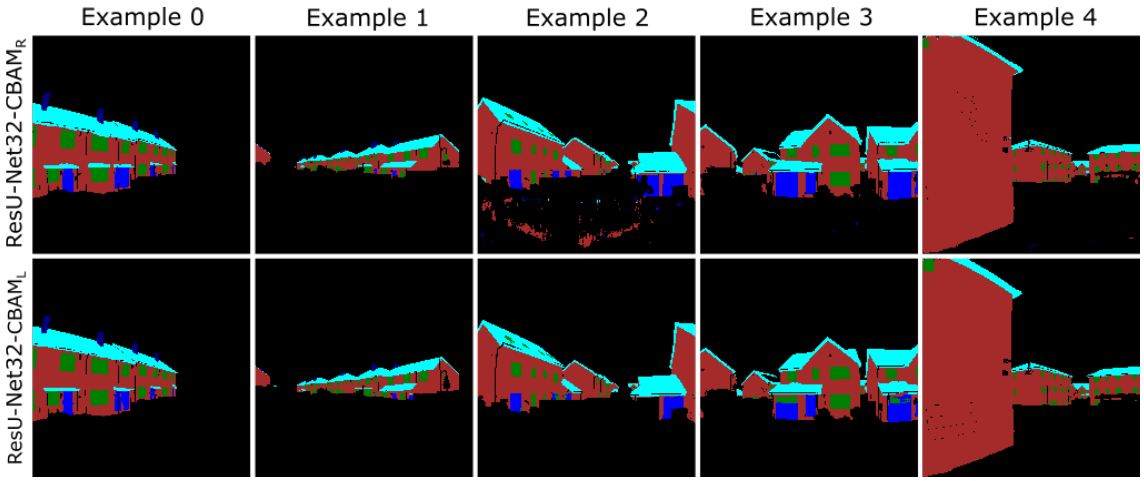


Figure 4.15: *Qualitative analysis of the model architecture refinement experiment. Five samples are selected here to compare qualitative performances of the six different models.*

4.5 Cross-dataset Adaptability Test

4.5.1 Test Aims and Experiment Design

The generalisation of a trained model and the representativeness of a dataset is critical. In section 3.3.3, the generalisation of a portion of the Crookesmoor dataset is primarily validated. However, this test only counts the number of predicted objects while the generalisation performance on pixel-level accuracy of this dataset has not yet been explored much. In this section, a series of experiments are proposed to validate 1. the representativeness of both the Crookesmoor and Handsworth datasets and 2. the generalisation of the designed FacMagNet-s model using the latest Oxford RobotCar-Facade dataset.

Two model architectures are selected as the base models. One is DeepLab-v3plus which has achieved the best result of all end-to-end models on the Crookesmoor dataset, as shown in Table 4.4. The other is the ResU-Net16 which has shown an equivalent competitiveness on the Handsworth dataset as shown in Table 4.8. The door magnifier model is initialised using weights of the door segmentation model of the trained FacMagNet-s on the Crookesmoor dataset. The designed test contains eight experiments including:

1. ***model generalisation test:*** Experiments applying the DeepLab-v3plus and the ResU-Net16 which have been trained on the Crookesmoor and the Handsworth datasets, respectively, directly on the test sets of the Oxford dataset;

2. ***transfer learning test:*** Experiments re-training the two models in 1. using the training set of the Oxford dataset. In each model, their last convolution layer has been

removed and replaced with a new head which is initialised by the Kaiming initialisation;

3. ***transfer learning necessity test:*** Experiments training a pre-trained DeepLabv3plus with weights initialised by the Cityscapes dataset and training a ResU-Net16 with weights initialised by the Kaiming initialisation;

4. ***magnifier strategy adaptability test:*** Experiments adding door magnifier modules on the best re-trained models for each architecture in test 2. and 3.

All experiments are implemented using TensorFlow (Abadi, 2016) library and trained on a workstation with Windows 10, 16 GB RAM, an Intel Xeon E5-1620 v4 CPU and an NVIDIA Quadro P5000 GPU. The adaptive moment estimator (Adam) (Kingma and Ba, 2014) optimiser is adopted with a learning rate decay strategy starting from 10^{-4}, the minimum learning rate is set to 10^{-6}. To prevent overfitting, an early-stopping strategy is applied which will stop the training process if validation precision does not increase in 10 epochs. All models are trained with categorical cross-entropy loss. Data augmentation setting is the same as in experiments on Crookesmoor dataset. The door magnifier module is trained using Adam optimiser with a learning rate decay strategy from 10^{-5} to 10^{-6} for 200 epochs, and the joint loss function defined in equation 4.2.1 is adopted here.

4.5.2 Oxford RobotCar-Facade Dataset

Oxford RobotCar-Facade dataset was announced in May, 2022 (Wang et al., 2022) which is the latest facade segmentation dataset publicly available. The dataset has been briefly reviewed in section 2.3.2. The dataset is a re-development of the Oxford RobotCar dataset (Maddern et al., 2017) which is based on a certain route in Oxford, UK. The dataset contains six categories which are background, facade, window, door, shop and balcony. The annotation strategy purpose is unknown. The class shop is not labelled as in other facade segmentation datasets such as ECP as an integration of billboards and show windows. Only billboards are annotated. Balconies are railing-made platforms.

A key characteristic of this dataset is that the camera used for capturing data only points towards the front. Therefore, the majority of data collected shows a street-canyon view. Street-canyon images were used for measuring radiation view factors previously by Gong et al. (2018). An ambiguity is how the authors annotated objects with occlusions. In some cases, objects with very sparse occlusions are not annotated. In comparison with region-wise annotation datasets such as CMP and ECP, Oxford RobotCar-Facade performs pixels-level annotation which provides more accurate ground-truth. However, annotation errors are still observed in this dataset, especially for missing elements. Whether or not the annotation

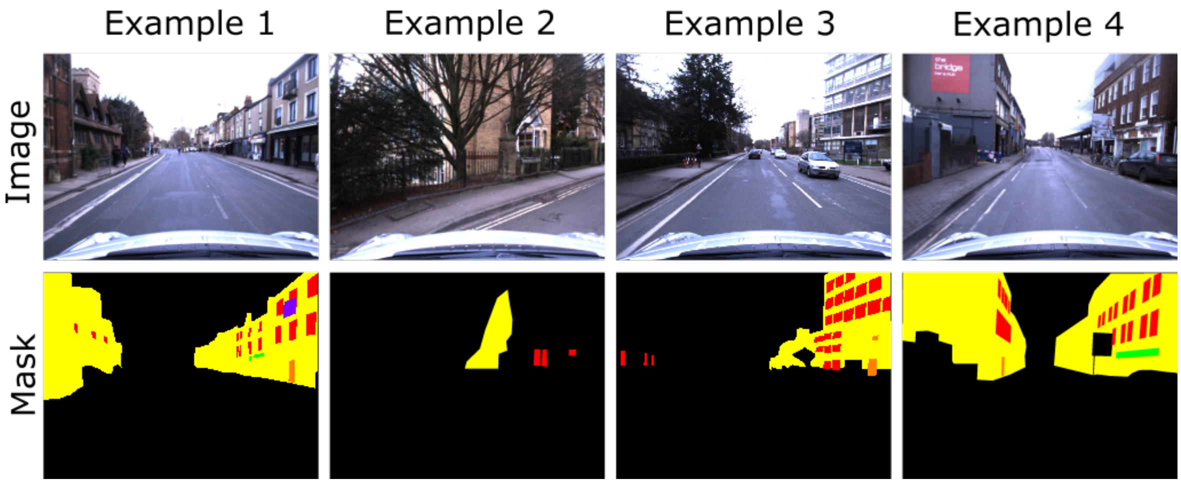


Figure 4.16: *Annotation examples of the Oxford RobotCar-Facade dataset. Example 1, annotation with all categories; Example 2, annotation with sparse occlusions; Example 3, annotation with missing elements, the top and third floor have missing windows, the ground floor has missing windows and incorrectly labelled doors; Example 4, annotation with shop instances.*

process has a quality control is still unknown. Some examples are shown in Figure 4.16.

Another problem of the Oxford dataset is its image quality although its image size is the highest among other publicly available facade segmentation datasets, as shown in table 2.2. However, its data has more noise which may be because of their sensor or camera settings such as exposure and shutter speed. From the images, it is suspected that they have stretched the image contrast while objects are still blurry in many cases.

4.5.3 Results and Discussion

4.5.3.1 Test Results

As the annotation rule of the Oxford RobotCar-Facade dataset is different to the proposed rule in section 3.2. The following changes were made first on the Oxford dataset:

1. The classes 'balcony' and 'shop' were merged with the class 'facade'. This decision is feasible and potentially it will change component features at a minimum as the 'shop' class is actually billboards. However, it seems like only shop headers have been annotated while some large billboards are not labelled as in figure 4.16 Example 4. The 'balcony' is not merged with the 'window' class since it is observed that the boundary of a 'balcony' object will commonly exceed the boundary of a window.

2. The raw Oxford data has a fixed image size of 1280 × 960, however, the data used for models trained in this thesis needs to be a square shape. Therefore, a raw image is split into two images, each with a fixed size of 960 × 960 and is resized to 1024 × 1024 to fit the input image size of the trained models. Then, during the inference stage, the predicted split twin is merged through a regular voting strategy.

Table 4.9 demonstrates the outcome of the eight tests described above. Considering the state-of-the-art results on this dataset is 53.8% in mIoU (Wang et al., 2022), models trained on the Crookesmoor and Handsworth datasets can generalise properly in the Oxford dataset, 40.87% and 34.32%, respectively, in mIoU. After training on the Oxford dataset, their performance reaches the same level c.57%. In comparison with using a random initialisation, using pre-trained weights leads to a 3% increase in the ResU-Net16 model. However, it is observed that using weights trained on the Cityscapes dataset leads to a better result than using weights trained on the Crookesmoor dataset. Applying the magnifier strategy improves the DeepLab-v3plus's performance by 0.03% while it deteriorates ResU-Net16's performance by 0.28%.

Table 4.9: *The mIoU results of the cross-dataset adaptability test on the Oxford RobotCar Facade dataset.*

Base model	DeepLab$^{+}_{v3}$				ResU-Net$_{16}$			
Weight initialisation	Crookesmoor		Cityscapes		Handsworth			Kaiming
Method / Cat.%	Trained	Transfer	Transfer	Magnifier	Trained	Transfer	Magnifier	Random
Background	84.50	92.54	93.46	**93.47**	79.08	91.75	91.65	91.27
Window	28.65	46.80	51.67	**51.70**	28.66	50.05	49.86	47.41
Door	4.11	17.27	26.11	**26.40**	8.60	15.19	14.72	8.36
Wall	46.24	72.85	75.39	**75.39**	20.90	71.19	70.86	68.98
mIoU	40.87	57.37	61.71	**61.74**	34.31	57.05	56.77	54.00

Figure 4.17 shows two examples of the qualitative results of each experiment. Comparing predictions (c) and (d) of ResU-Net16 and (b) and (d) of DeepLab-v3plus in Figure 4.17, the magnifier module can help to refine the door boundaries. However, it will also refine falsely detected doors, e.g. Example 1(d) using the DeepLabv3plus model. Oxford RobotCar-Facade is a challenging dataset. A substantial amount of images in this dataset contain a considerable amount of noises. This leads to object boundaries being ambiguous in many cases, especially for doors as seen in the dataset. Therefore, without sufficient features, wrong features could be learned from the noised data and that is potentially why the biased predictions are not found on the Crookesmoor dataset as shown in Figure 4.12.

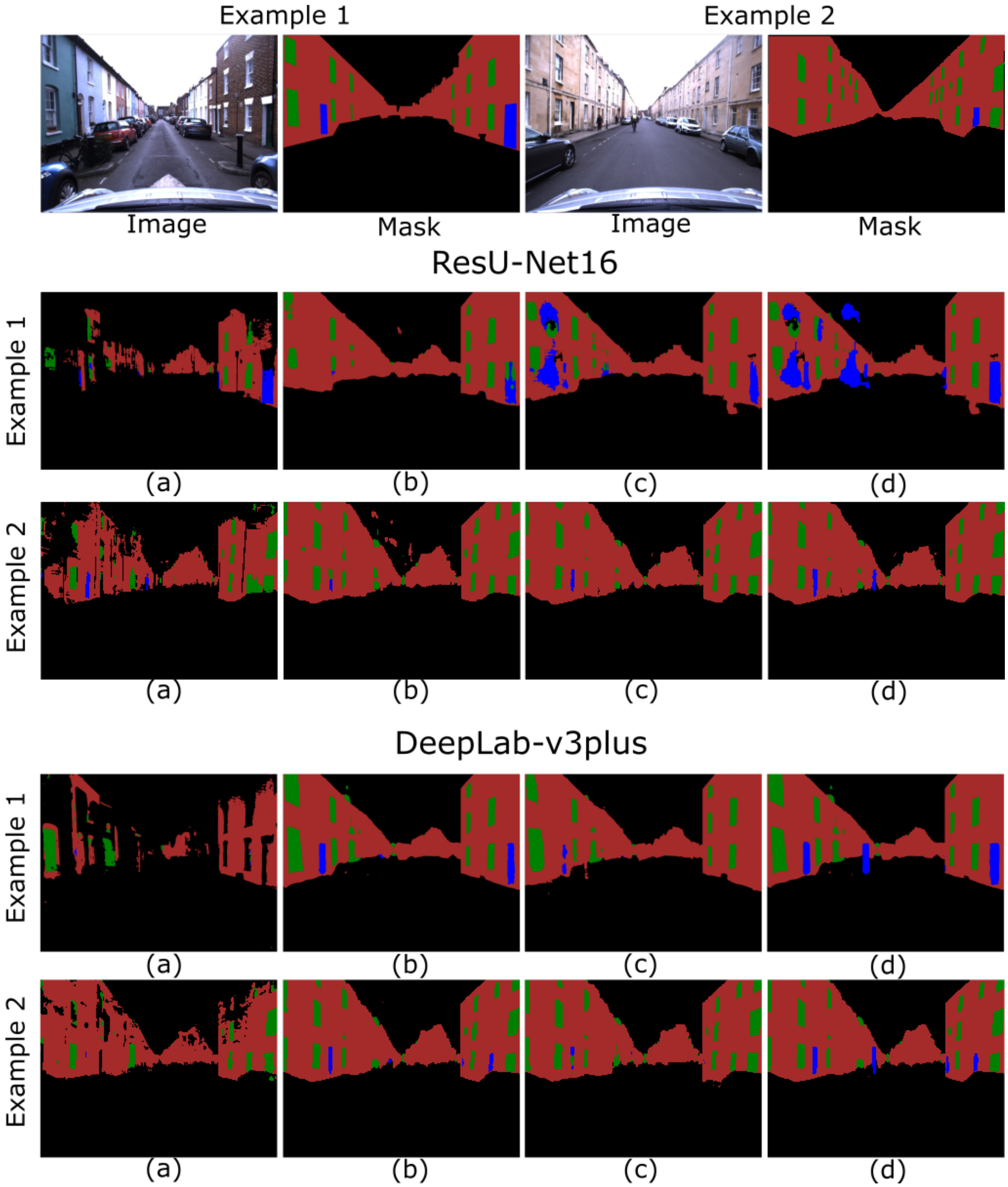


Figure 4.17: *Qualitative analysis of the Oxford RobotCar-Facade dataset benchmark test. (a) is the results of using pre-trained models; (b) is the results of using random initialisation in ResUnet16 model and Cityscapes weights in DeepLab-v3plus model; (c) is the results of using fine-tuning training and (d) is the results of adding a magnifier module.*

In order to quantitatively measure the boundary refinement capability of the magnifier module, the recall metric results of the test are reported in table 4.10. The recall metric measures the ratio of true positives over the ground-truth. Therefore, the metric can well represent how object boundaries have been refined. The results have shown that the proposed magnifier strategy can help both models increase their recall values. Their recall values have been raised by the magnifier module for 15.1% and 15.5% for the Cityscapes-initialised DeepLab-v3plus and the Handsworth-initialised ResU-Net16 models, respectively.

Table 4.10: *The recall results of the cross-dataset adaptability test on the Oxford RobotCar Facade dataset. 'Before' sections represent recall results without integrating the magnifier module; 'After' sections are results after adding the magnifier module.*

Category Model		Background	Wall	Window	Door
DeepLab_{v3}^{+}	Before	0.975	0.848	0.622	0.330
	After	0.974	0.845	0.621	0.481(+0.151)
ResU-Net_{16}	Before	0.965	0.821	0.584	0.253
	After	0.964	0.816	0.580	0.408(+0.155)

4.5.3.2 Full-Category Test Results

The results which have been achieved above have shown that the developed magnifier strategy can improve door prediction performance. The highest score of mIoU which has been achieved in the Oxford dataset is 53.8% (Wang et al., 2022). Therefore, FacMagNet-s may be also competitive in the state-of-the-art works although it was proposed over a year ago and only aimed to enhance door predictions. However, as shop and balcony classes are merged into the wall category in this experiment to fit proposed labelling definitions in chapter 3, how FacMagNet-s will perform in the initial six-class classification is still unknown. Therefore, the declaration that FacMagNet-s is state-of-the-art cannot be made here.

In order to investigate the performance of the developed FacMagNet-s compared to other facade segmentation approaches, an extra experiment is conducted which includes all the six categories in the Oxford dataset. A DeepLab-v3plus is adopted as the stem segmentation model and the magnifier module is the same as in the previous magnifier strategy adaptability test. The DeepLab-v3plus is initialised by training on the Cityscapes dataset and the loss function is changed to focal loss. Other experiment settings are the same as the previous 4-category experiments. Table 4.11 demonstrates quantitative comparisons of SOTA approaches of facade segmentation on the Oxford RobotCar-Facade dataset.

Table 4.11: *Benchmark results on the Oxford RobotCar-Facade dataset. Wang et al. (2022) also train their DeepLab$^{+}_{v3}$ on this dataset. Their result is included in the bracket. Benchmark models include DeepFacade (Liu et al., 2020), Pyramid ALKNet (Ma, Ma, Xu and Zha, 2020) and the Facade R-CNN (Wang et al., 2022) which are SOTA in facade segmentation and have all been reviewed in section 2.3.3.*

Model	DeepFacade	Pyramid ALKN	Facade R-CNN	FacMagNet-s	DeepLab$^{+}_{v3}$
mIoU(%)	47.31	51.22	**53.8**	50.22	49.96 (50.33)

Overall, FacMagNet has shown a comparable result among state-of-the-art facade segmentation models. The magnifier module has helped the DeepLab-v3plus to increase 0.26% on mIoU. Wang et al. (2022) do not include categorical IoUs in their article. However, here the categorical IoU results are included in table 4.12. The table has shown that the door magnifier module has led to a 1.71% door IoU increase while it leads to an IoU decrease in wall and background categories.

Table 4.12: *Benchmark results on the Oxford RobotCar-Facade dataset with full categories included.*

Cat.% / Model	mIoU	Background	Wall	Window	Door	Balcony	Shop
DeepLab-v3plus	49.96	92.99	74.67	49.70	22.01	23.41	36.98
FacMagNet-s	50.22	92.97	74.51	49.71	23.72	23.41	37.00

4.5.3.3 Section Discussion and Conclusion

The developed magnifier strategy does not lead to a significant door prediction increase on the Oxford dataset. The major potential reason is that the data quality of the Oxford dataset is not satisfactory. Most of the door instances are highly distorted due to street canyon views and a significant portion of door instances have blurred boundaries and appearances. This may cause these instances to not contain sufficient features for deep learning models to learn. Figure 4.18 compares door instances from the Oxford dataset and the built Crookesmoor and Handsworth datasets. The figure clearly shows that door instances in the Crookesmoor and Handsworth datasets contain much more details than ones in the Oxford dataset and have sharper boundaries. The data quality problems are less severe for window and shop instances. Therefore, from table 4.12, it can be seen that window and shop can achieve higher scores. Besides, the Oxford dataset contains 8820 windows but only has 331 doors. The limited quantity of door instances also leads to insufficient features.

Labelling quality is another potential reason of why the magnifier strategy does not achieve distinct quantitative performance refinement on the Oxford dataset. As an example, in figure 4.17 example 2, there is a highly distorted door on the left-hand side of building blocks while it is not annotated. Nearly all models can recognise this instance and magnifier models have given confident shapes. However, as it is not labelled, in calculating IoU, predictions of this instance will be accounted as false positives rather than true positives. It is unknown whether this is a labelling error or whether the builders of the Oxford dataset have defined specific rules to avoid annotating ambiguous instances. In comparison, the 'labelling all visible' rule defined in section 3.2.2 clearly shows its advantage in more precisely measuring model performances.

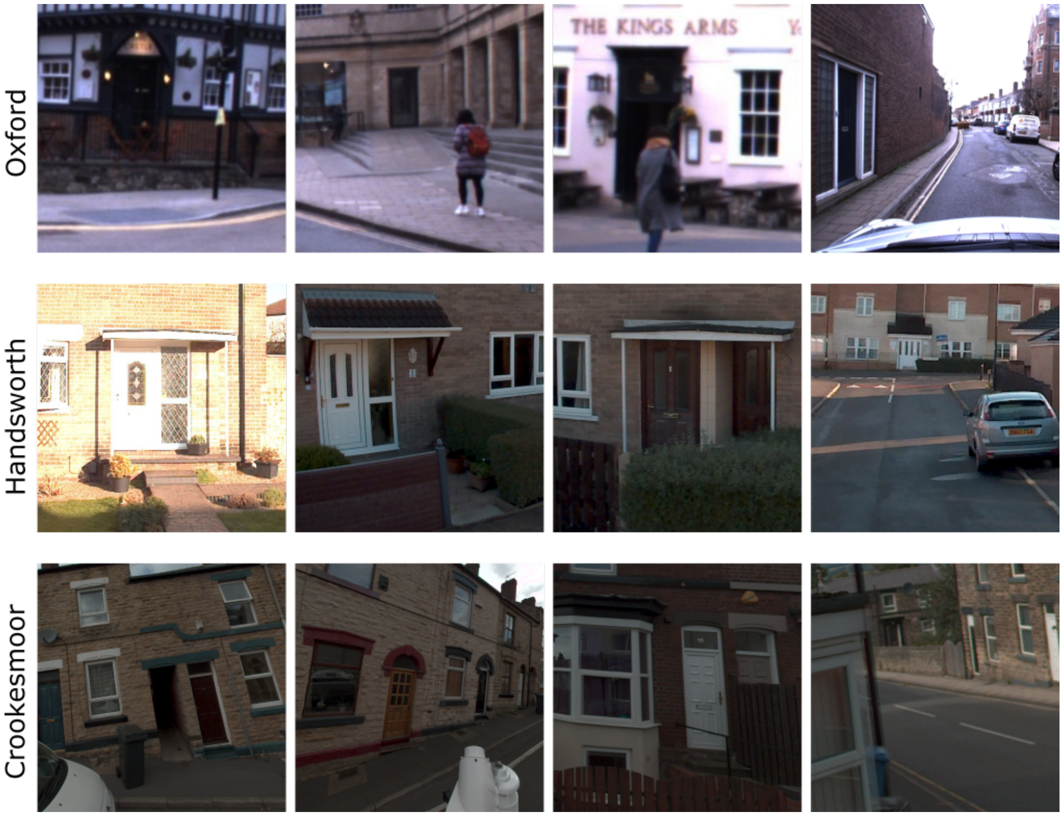


Figure 4.18: *Door instance comparison across three datasets.*

4.6 Chapter Discussion and Conclusions

4.6.1 Discussion

In this chapter, a building facade semantic segmentation model, FacMagNet-l, is first developed to recognise residential building facades at component level. The model has been carefully designed to detect features of residential buildings from street-level imaging. A key characteristic of the data used is that there is a high size discrepancy both between different classes and within the same class. The proposed model employs a symmetric structure, dilated convolution and a bagging ensemble learning strategy, as well as a magnifier to handle intra-class imbalance. The results presented in section 4.2.3.4 demonstrate the prediction performance of the developed model on facade segmentation against contemporary semantic segmentation models. The results have shown the proposed model exceeds the DeepLab-v3plus model for 4.03% in mIoU. The FacMagNet-l model is further explored and shrunken for efficiency purposes. The designed FacMagNet-s model has achieved the second-best accuracy performance over the state-of-the-art models while it has a smaller size and lower computational cost than the FacMagNet-l model.

A critical drawback of the proposed models, due to the ensemble nature of the two models, is the high time and computational resource requirements, both for the training and more importantly for prediction. As six individual models are integrated with differing architectures in FacMagNet-l and two individual models are cooperated in FacMagNet-s, the designed models require more resources compared with using an end-to-end model, such as the DeepLab-v3plus (Chen et al., 2018). Table 4.6 shows that FacMagNet-s exceeds the DeepLab-v3plus model for over 118M FLOPs and has 59M more trainable parameters. Although, in comparison with some other SOTA works in the computer vision community such as the latest vision transformer (Dosovitskiy et al., 2021) which contains 87M trainable parameters and 67G FLOPs, the designed FacMagNet-s is still within a reasonable level. The two-stage design would still make the training and inference more complicated, however, because the use case of the model, for example within the retrofit pipeline, is unlikely to need real-time execution, this is not likely to impact the usefulness of the proposed models.

Soft attention mechanism and focal loss were tested on the Handsworth dataset, however, unlike they are validated in corresponding papers, there is no obvious evidence showing that they are beneficial to facade segmentation at this point. The computational resource that the PhD project owned also limits validating more attention models. A cross-dataset adaptability test was designed in this chapter using the most recent Oxford RobotCar-Facade dataset. Results show that the proposed FacMagNet-s is a comparable state-of-the-art approach in facade segmentation using street-view images. Besides that, the generalisation capability of trained models on the built Crookesmoor and Handsworth dataset has been quantitatively and qualitatively validated. Corresponding experiments show that the model trained on the Crookesmoor dataset is more beneficial on the Handsworth dataset.

4.6.2 Conclusions

The data collected by street-level capture contains a significant amount of environmental noise irrelevant to buildings. To extract building information such as building geometry and thermal characteristics, from the data collected, a facade segmentation model with accuracy priority is essential. The state-of-the-art related works in the areas of facade segmentation and urban scene segmentation show the dominant position of the convolutional neural network technology. The state-of-the-art works in the facade segmentation area show the significance of refining the boundaries of smaller facade components such as windows and doors to improve the segmentation results. Smaller facade components are also key components which require high accuracy in assisting building retrofit. The state-of-the-art works in the urban scene segmentation make many contributions to the multi-scale challenge. However, these works lack component-level consideration in segmenting buildings. Before 2021, the state-of-the-art facade segmentation methods commonly focus on heavily pre-processed data according to the literature review in chapter 2.

In this chapter, two models are presented for the semantic segmentation task of building facade components, known as FacMagNet-l and FacMagNet-s. These models were purposely built for the task, utilising contemporary deep learning architectures and utilising ensemble learning strategies to categorise each object. The results have demonstrated that the accuracy of these two models on labelling images is at a promising level. The cross-dataset test demonstrates that the developed FacMagNet-s has achieved a comparable result among other state-of-the-art models for the given task.

Along with the development of the model and evaluation of urban street-level data, clear motivation has been identified for this approach in the pathway to scalable residential retrofit. By incorporating multispectral capture, the localised building features will be able to directly contribute to automating the current building energy analysis and building material stock modelling, for use by stakeholders such as local government authorities.

Chapter 5

Discussion and Conclusions

5.1 Discussion

This section provides a comprehensive discussion of the thesis. The section is arranged by individual topics in each subsection.

5.1.1 Dataset Construction and Scope

5.1.1.1 Experience on Building Datasets for Facade Segmentation

Building the dataset is the first step of any data-driven task. The pipeline of dataset annotation is common in the machine learning community: an annotation guide should be proposed first, which should meet needs of a specific task; then annotators should be appointed to label the collected data. In this thesis an annotation pipeline is proposed with annotation rules defined to fit building retrofit needs. Two datasets were built based on the determined rules with different annotators.

A critical problem which may be raised during constructing datasets is how to manage annotation quality. Ideally, annotators should be experienced in annotation tasks and have expertise on the target domain of the task. Experienced annotators are more efficient in using labelling tools. The candidate has independently annotated a 300-image window-and-door dataset. The dataset is introduced in Appendix A. In the window-and-door annotation task, the labelling speed can be reduced from 30 min to 10 min per-image once the annotator becomes familiar with the annotation tool and rules. Having knowledge of the task target can help annotators make more accurate decisions when making annotations. However, experienced and knowledgeable annotators are not always available. Therefore, a preliminary annotation guide was produced for constructing the Crookesmoor dataset. In table 3.3, the evaluation shows that, even with preliminary guidance, a significant number of annotation errors are still observed. Therefore, apart from a preliminary interactive annotation training, a quality inspection process is also necessary to build a high-quality dataset.

5.1.1.2 The Representativeness of Built Datasets

The scope of the dataset determines the upper bounds of the generalisation of trained models to the real-world. The raw images in the Crookesmoor and Handsworth datasets were all collected in the Sheffield region. However, due to the architectural similarities of houses to those in other parts of the country, this does not mean the built dataset can only represent houses in the Sheffield region. In section 4.5, trained models are applied on the Oxford RobotCar-Facade dataset directly and have shown satisfactory results. The DeepLab-v3plus model which was trained on the Crookesmoor dataset shows 40.87% in mIoU and the performance of the re-trained model is 57.37%. Training the same model on the Oxford dataset only raise its mIoU for 16.5%. In June, 2022, a batch of new data was captured by MARVEL in Merthyr Tydfil, Wales. Although these data are not labelled, they can still be used to qualitatively test, i.e. by visual inspection, the generalisation of created datasets in different regions. As this database consisted of distorted raw images as in the Handsworth data, the trained ResU-Net16 was adopted as in section 4.5. Figure 5.1 shows the qualitative test results. All examples were randomly selected. The quantitative test on the Oxford dataset and the qualitative test on the Merthyr Tydfil database validate that the built datasets can be populated against data captured in other regions, at least in the three designated areas in the UK.

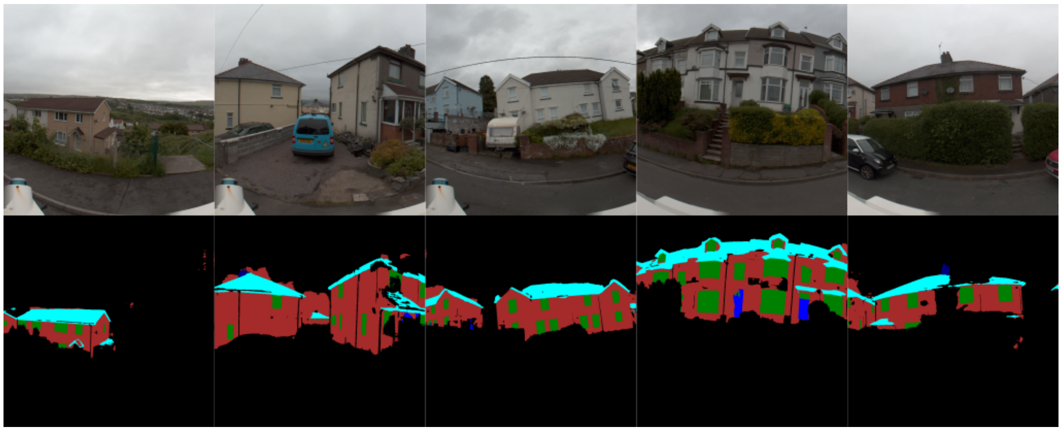


Figure 5.1: *Qualitative analysis of generalisation capability of model trained on Sheffield houses using Merthyr Tydfil data.*

5.1.2 Facade Segmentation Using Street View Images

Prior to 2019, approaches developed for facade segmentation were dominantly based on using rectified frontal-parallel images. This type of image is beneficial for building procedural modelling for computer graphic applications such as video games. However, from environmental modelling perspectives, cropped frontal-parallel facade images do not contain much required information for retrofit such as multi-perspectives. RueMonge 2014 (Riemenschneider et al., 2014) is the first dataset developed for structure-from-motion based building reconstruction. However, this dataset has cropped views and labelling only on buildings on the data-capture street. The drivers, i.e. what applications will be used for, and enablers, i.e. what datasets are available to use, in the facade segmentation community both limited the progress of this research domain. Street view building images contain more information such as surrounding environment and multi-perspectives which is more beneficial to built environment modelling. From 2021, two street view image based facade segmentation datasets became publicly available (Orhei et al., 2021; Wang et al., 2022). In the coming years, it could be expected that more attention could be paid to street view facade segmentation since more street view datasets become available.

A significant consideration of using street view images on facade segmentation is what segmentation methods to use. Among state-of-the-art approaches for facade segmentation using frontal-parallel facade images, prior knowledge of facade is widely-exploited and is critical to improve segmentation results. However, building facades in street view images will lose many critical architectural features such as regularity and symmetry. Therefore, segmentation approaches developed based on building priors such as regularity and symmetry will not theoretically bring benefits to facade segmentation using street view images. As shown in Wang et al. (2022), a vanilla DeepLab-v3plus exceeds the state-of-the-art DeepFacade model by 3.02% in mIoU on the Oxford RobotCar-Facade dataset. The result is also demonstrated in table 4.11.

Whether or not an advantageous model architecture like U-Net in medical image segmentation exists for facade segmentation is unknown. This is very possible, as a building is an organised structure like organs and cells even without rectification. Facade segmentation is a branch of the general image segmentation which is a dynamic and fast-growing research domain. Many off-the-peg algorithms have been developed for image segmentation, although they do not consider the uniqueness of buildings as a man-made structure and have not yet been benchmarked in facade segmentation. In section 4.2 and 4.3, contemporary models have been tested including DeepLab-v3plus, PSPNet, SegNet, U-Net, attention modules. These experiments have shown that different model architectures may vary distinctively. For example, PSPNet only reaches 64.27% in mIoU in the Crookesmoor dataset, while it has achieved success in rectified facade segmentation (Zhang et al., 2022). It is unknown whether this is due to different training strategies or simply because of model structure fitness. DeepLab-v3plus and the U-shape structure model both have achieved promising results in this thesis. In the Crookesmoor dataset, DeepLab-v3plus has achieved 76.98% and U-Net has achieved 69.6% in mIoU.

Soft attention mechanism does not show performance improvement on the Handsworth dataset. In the best case, adding attention mechanism only increases door prediction performance for 0.16% while it decreases the mIoU for 0.16%. However, attention mechanism should be effective for facade segmentation as each category is not equally hard to predict. There could be multiple reasons leading to failed attention mechanism outcome such as training strategies, hyper-parameters, base model's architectures, etc. Apart from this, the stability of attention mechanism might be another reason. As introduced in chapter 2, the attention mechanism is proposed in the computer vision community but is then progressed in the natural language processing community of the Query-Key-Value paradigm before being populated back to the computer vision community. A common and plausible explanation of attention mechanism feasibility is that the human vision system also has an attention mechanism. However, prior to 2021, research on attention mechanism in the computer vision community were based on plug-and-play modules, to the best of the candidate's knowledge. Whether or not these modules can be globally effective like the fully convolution architecture has not been sufficiently studied. In 2021, a novel attention-based architecture was developed called Transformer (Liu et al., 2021) which may be useful for facade segmentation.

How data scope will impact on results and generalisation of facade segmentation models is still not clear, especially for facade segmentation using street view images. The Crookesmoor dataset has 997 images and the Handsworth dataset has 5905, however, the same DeepLab-v3plus model has achieved equivalent performance on these two datasets: 76.98% and 77.70%, respectively, in mIoU. Therefore, it is unknown why model performance does not increase with significantly more samples. If c.80% is the ceiling of British house recognition using DeepLab-v3plus it means that it is unnecessary to build any larger datasets for this task. One plausible reason could be that objects in these datasets are not equivalently hard to predict: according to table 3.7, objects in the Handsworth dataset are smaller than those in the Crookesmoor dataset with the same image size. However, it is still significant to know how much data would be required in order to provide sufficient information for a facade segmentation model.

The scope of the dataset will affect the model generalisation performance. In conventional facade segmentation research that uses frontal-parallel facade images, the top-three frequently-used facade segmentation datasets, all have sizes smaller than, or approximate to one-hundred images which is extremely small in this deep learning era. Also, works discussing whether models trained on these datasets can be applied or populated to other scenes is currently lacking. Kelly et al. (2017) built an 800-image private dataset to compensate for the data scarcity for their application which was scalable building reconstruction. Compared to a rectified facade image, a street view image contains scenes from different perspectives and environmental information. Therefore, generalisation will be more challenging for models trained on street view images which makes knowing the number of sufficient scenes critical to a real-world application.

Door prediction is a difficulty in facade segmentation. Table 5.1 lists the state-of-the-art door prediction performance in three street view image datasets and it also includes results from the ECP dataset which is the most frequently used dataset in conventional facade segmentation domain. In all four datasets, door prediction performance is significantly lower than their average IoUs. The proposed magnifier strategy can refine door predictions although it it still cannot fully address the door prediction weakness.

Table 5.1: *Comparisons of building component prediction performance across different datasets.*

Dataset Category%	Crookesmoor	Handsworth	Oxford	ECP
Door	65.82	67.79	26.40	66.2
Window	87.83	82.25	51.70	81.6
mIoU	81.01	82.15	61.74	81.9
Source	Dai et al. (2021)	Table 4.8	Table 4.9	Zhang et al. (2022)

5.1.3 The Magnifier Strategy and its Limitations

The magnifier strategy is proposed in this thesis, which aims to reduce the intra-class data imbalance problem. The strategy crops an area of raw images containing target objects to balance the foreground and background. Using an individual object detection model to generate object patches has validated that it can refine object boundaries and detect missing objects on the Handsworth dataset. The magnifier strategy has shown distinct capability in improving door predictions. FacMagNet-l has contributed to a 9.3% mIoU increase and FacMagNet-s has led to a 4.24% increment. FacMagNet-s was then applied to the Oxford RobotCar-Facade dataset. Using DeepLab-v3plus as the base model contributes to a 0.03%

increment. However, using the Handsworth dataset lead to a 0.47% performance drop. Although the magnifier strategy does not show prominent capability in the mIoU metric on the Oxford dataset, it has raised c.15% in the recall metric.

Overall, as a strategy to improve hard case predictions, i.e. door, the magnifier module has shown promising and effective performance. However, as a data-driven approach, the magnifier strategy is still heavily relied on the quality of dataset. This limitation is very obvious in experiments conducted on the Oxford dataset: falsely detected objects from stem segmentation models and ambiguous boundaries due to dataset quality will cause the magnifier model even to be negative to the mIoU performance. Although, thanks to the advancement of camera sensors, high-quality images become moor achievable than a decade before. They are not still always available. How to reduce negative impacts caused by data quality becomes critical towards further improving the magnifier strategy. An new idea from the candidate is to exploit vanishing points of images, but it still needs time to explore and validate.

Using an object detection model to extract objects will significantly increase model sizes and computational cost as shown in table 4.6. However, comparing results of FacMagNet-l and FacMagNet-s in table 4.5, an object detection model might still be necessary. Although ensemble learning is widely used in SOTA approaches of conventional facade segmentation, reducing computational cost is still essential to make the training process more feasible. As an example, Zhang et al. (2022) use four NVIDIA GFORCE 2080Ti GPUs to train their model which requires approximately £6000 at 2022 price according to Amazon.co.uk. In order to reduce computational cost, FacMagNet-l can be reduced to three models i.e. an object detection model, a three-class segmentation model for window, door and chimney and a two-class segmentation model for wall and roof.

5.1.4 The Overall Target of Assisting Building Retrofit

The thesis target is building a facade segmentation pipeline to assist large-scale building retrofit. As discussed in section 1.1.1, an efficient and flexible data collection approach with automated data analysis methods is vital in delivering building retrofit at scale. On-street data capture, such as MARVEL (Meyers et al., 2019), is designed to scale the data collection of urban environmental data. It is impractical, even infeasible, to manually extract information at this scale. Automating feature extraction leads to efficient strategies for urban data analysis.

A vehicle-mounted remote sensing system realises a highly efficient solution to multi-spectral urban environmental data capture. Localisation of properties in space can help build a portrait of a building, for entry into further modelling such as building energy models, which are vital in building retrofit solutions. Combining localised properties with co-captured data from thermal and hyperspectral cameras can provide high fidelity representations of different features. When incorporating hyperspectral information, it may be possible to characterise the material properties of the wall that may not be possible with visible light only due to, for example, painted surfaces. Likewise, thermal images can be divided and the thermal bridges of different components can be assessed independently. This is useful for fault detection in glazing, or determining the nature of the cavity in a wall. Both of these features are important when building high quality models for facade recognition.

The developed facade segmentation pipeline demonstrates an accurate process for segmenting features on building facades, thus inferences can be made on the properties such as some of those outlined in Table 3.1, e.g. number of storeys and number of windows. Performing this in a scalable manner is the first step towards automating or semi-automating the development of retrofit solutions for residential buildings. The developed model has been shown to give high quality segmentations of all defined components which are determined in section 1.1.1, in a wide range of building types, with little noise from unrelated structures or objects.

The integration of the proposed building facade segmentation approach and other types of data collected by vehicle-mounted data capture system will provide higher-level understandings of buildings in the selected area. The 3D point-cloud urban environment models can be generated by the LiDAR units. Through integrating the facade segmentation results with the 3D models, the buildings and their components can be identified from the urban environmental models. The integrated model will then provide a comprehensive understanding of the regional buildings. As the point cloud data contains real-world space information, building volume can be calculated. The building components' quantities in an area, e.g. window and door, can also be counted automatically. Other important building metrics such as the glazing ratio, which is a significant parameter in evaluating the building energy behaviour, can also be calculated by the point cloud-semantic segmentation integration model. This ratio can be used to evaluate the building's natural illumination and ventilation conditions. For example, a building with a low glazing ratio might need the introduction of synthetic ventilation, such as the THEX (Total Heat EXchanger) (Fukami and Okamoto, 1984) during retrofit.

The thermal images and hyperspectral images determine the thermal performance and material types of buildings, respectively. The use of integration of point-cloud and infrared thermography data to detect thermal leakage was previously explored in (Hoegner and Stilla, 2015). With the integration of the facade segmentation approach, certain building components of thermal leakages can be determined. This will contribute to a more precise retrofitting plan in terms of material replacement or improvement. The spectral information can be used to identify building materials (Ilehag et al., 2019). Incorporating the hyperspectral data with the integrated 3D model could be used to gather statistics of building material stock. Compared to the traditional methods, this method has fewer constraints since it does not require historical records. In addition, the method is also potentially more accurate since it does not need to define archetypes to approximate the building types in an area.

To identify the potential superiority of using hyper-spectral imaging in identifying building construction materials in the future, a wall construction type dataset is built by the candidate and introduced in Appendix B using Google Street View images and energy performance certificates (EPCs). In the UK, EPC contains building type information which relates to heat transmittance. The dataset is built by registering GSV building images with corresponding EPC building types using address information through the Google Street View API. The GSV API will return the nearest image based on the input address. Therefore, the returned images may not be captured perpendicular to target building facades. Particularly for the designated area, Merthyr Tydfil, it was found that the density of the data capture point is rather sparse. In some cases, only one data capture point is observed and thus the registered dataset needs to be manually filtered first. After filtering, the dataset contains three categories: cavity, stone and solid brick with each containing 4412, 4208 and 1182 images. A preliminary experiment is conducted by applying a vanilla ResNet50 (He et al., 2016*a*) onto this dataset which has achieved 85% in overall accuracy.

5.2 Conclusion

The thesis has developed two supervised deep learning-based models to detect the building facades and their components from visual images. The work builds the foundation of the building retrofit pipeline. The supervised learning-based approaches have shown promising performance in detecting building facade components in both the accuracy and the generalisation facets. The major research contributions can be summarised as follows:

- The thesis provides a pipeline of automatically characterising English house-building facades using the deep learning technique. The pipeline includes an annotation protocol and a building ontology which is proposed to fit building retrofit needs; five key elements, window, door, chimney, wall and roof, for building retrofit were identified. The protocol can be conveniently extended if a building facade requires more sophisticated annotation, e.g. window needs to be further segmented into frame and glass. Two datasets were constructed based on the proposed protocol. Their raw data is captured in the Crookesmoor and Handsworth areas in Sheffield, UK. Compared to existing facade segmentation datasets, these two datasets have the highest image sizes, and more diverse scenes and they are the first proper facade segmentation datasets using street view images. The process of constructing the two datasets is carefully designed and dataset quality is inspected. It is found that an interactive training process could reduce labelling errors and a quality inspection procedure is important.

- A deep learning technique-based strategy, magnifier, is proposed in this thesis. The strategy is developed to balance foreground and background pixels. Implementing the strategy has two different formats. The first, FacMagNet-l is to use an object detection model to extract patches containing target objects. The second, FacMagNet-s is to use predictions from a multi-class semantic segmentation model and morphology operations to extract target objects' patches. Results have demonstrated that both of the implementation approaches can contribute to performance improvement. However, the implementation with an individual object detection model shows a 4.03% higher result in mIoU in the Crookemoor dataset. The FacMagNet models are the state-of-the-art in the Crookesmoor dataset. The FacMagNet-s is applied to the latest Oxford RobotCar-Facade dataset, it also achieved a comparable state-of-the-art. The generalisations of trained models and representativeness of constructed datasets were tested using the Oxford dataset and raw data captured in Merthyr Tydfil by MARVEL. The results have shown that without re-training, models trained on developed datasets can be directly used for segmenting facades on images captured by MARVEL and data captured by different rigs. In addition, they have also shown competitive results on the dataset with different scenes and quality.

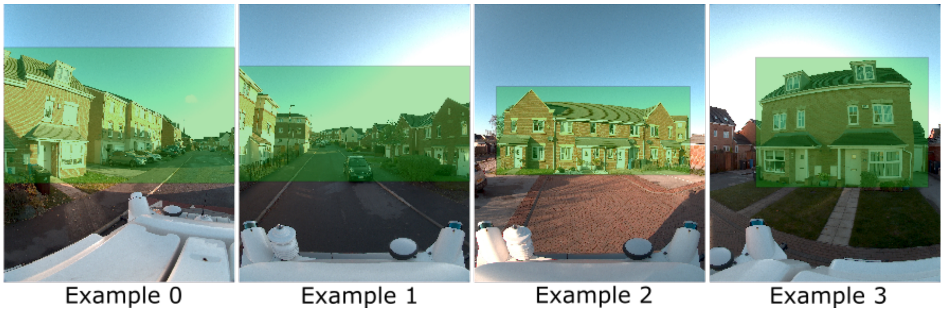


Figure 5.2: *Examples of using bounding boxes to cover building area. The green area represents the minimum bounding box cover. Terraced houses commonly show a long rectangular shape and detached houses normally approach a square shape. Large viewing angle would commonly lead to low occupation ratio and the opposite would result in high occupation ratio.*

5.3 Recommendations of Future work

The residential building is a highly-organised artificial structure which contains substantial prior knowledge. Although using the prior knowledge of buildings on the facade detection has been widely explored, the state-of-the-art approaches only focus on exploiting structure pattern features such as the grid outlines by using the rectified images (Femiani et al., 2018). However, for the street view residential building images, the buildings will not appear with structured patterns but with distortions instead. During the data annotation process, it was observed that although the facade patterns are destroyed due to image distortions, the building envelope scenes still have weak regularities. For example, if using the minimum bounding boxes to cover the detached and semi-detached houses in the captured visual images, the bounding box commonly approximates a short rectangular shape and building envelopes will normally occupy the majority of the bounding box areas. The occupied area ratio, which can be defined as the area of the building, envelops over the bounding box area. Terraced houses will commonly show a long rectangular appearance and their occupation can be estimated fluctuating in a range. Some examples are shown in Figure 5.2. The examples have shown the building types and orientations can potentially be estimated by the building box shapes and the occupation ratio.

The estimated building type and orientation information can be treated as a valuable prior in improving the performance of a semantic segmentation model. For example, if the bounding box has a long shape and occupation ratio is relatively low, the hypothesis can be made that the building inside the bounding box could be terraced with an inclined view. Then the box area can be sliced vertically to generate building patches. One of the difficulties in the facade component detection is the small object problem. The developed FacMagNet models utilise the component-level bounding boxes with the magnifier strategy to try to address the problem. The sliced building patches can then be magnified and inferred as in the developed magnifier module in FacMagNet models. As the magnifier strategy has shown clear advancement in improving the accuracy level, if using the strategy in magnified building patches, accuracy level may also rise.

The magnifier strategy has shown a distinct advantage in improving the door prediction accuracy level. However, using the component-level bounding boxes will inevitably introduce additional deep learning models whether an object detection model (in the FacMagNet-l) or a door semantic segmentation model (in the FacMagNet-s) or in the proposed building patch magnifier aforementioned. The unsupervised learning technology has been greatly explored in the past few years in the semantic segmentation area (Zou et al., 2018). If using an unsupervised learning model to roughly segment the building envelope from the image data first, the minimum bounding box can be used to extract the building locations and dimensions, in pixels. Then, without introducing an extra segmentation or detection model, the small-scale buildings can be magnified as in the FacMagNet model. The accuracy can then be potentially improved.

Deep learning-based semantic segmentation techniques have been widely explored in this thesis to identify building components. The exploration study shows that U-Net architecture-based models have outstanding performance in identifying building components. Besides designing a more advanced model architecture, action can be taken to optimise the model training process. The training strategies applied are still very restricted. Relevant literature has shown that various training tactics such as a different optimiser, starting learning rate and its reduction scheme choices would impact on inference results variously. Therefore, experiments on different training strategies are still required.

References

Abadi, M. (2016), TensorFlow: learning functions at scale, *in* 'Proceedings of the 21st ACM SIGPLAN International Conference on Functional Programming', pp. 1–1.

Adelson, E. H., Anderson, C. H., Bergen, J. R., Burt, P. J. and Ogden, J. M. (1984), 'Pyramid methods in image processing', *RCA Engineer* **29**(6), 33–41.

AgEagle (2022), 'Drone cameras available for eBee X'. Available at: `https://ageagle.com/drones/ebee-x/` (Accessed: 06 December 22).

Alom, M. Z., Hasan, M., Yakopcic, C., Taha, T. M. and Asari, V. K. (2018), 'Recurrent Residual Convolutional Neural Network based on U-Net (R2U-Net) for Medical Image Segmentation', *arXiv preprint arXiv:1802.06955* . Available at: `https://arxiv.org/abs/1802.06955` (Accessed: 06 December 22).

Anguelov, D., Dulong, C., Filip, D., Frueh, C., Lafon, S., Lyon, R., Ogale, A., Vincent, L. and Weaver, J. (2010), 'Google Street View: Capturing the World at Street Level', *Computer* **43**(6), 32–38.

Arbabi, H., Lanau, M., Li, X., Meyers, G., Dai, M., Mayfield, M. and Densley Tingley, D. (2022), 'A scalable data collection, characterization, and accounting framework for urban material stocks', *Journal of Industrial Ecology* **26**(1), 58–71.

Badrinarayanan, V., Kendall, A. and Cipolla, R. (2017), 'SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **39**(12), 2481–2495.

Boon, M. A., Drijfhout, A. P. and Tesfamichael, S. (2017), 'Comparison of a Fixed-Wing and Multi-Rotor UAV for Environmental Mapping Applications: a Case Study', *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* **42W6**, 47–54.

Breiman, L. (2001), 'Random Forests', *Machine Learning* **45**(1), 5–32.

Brostow, G. J., Fauqueur, J. and Cipolla, R. (2009), 'Semantic object classes in video: A high-definition ground truth database', *Pattern Recognition Letters* **30**(2), 88–97.

Brust, C.-A., Sickert, S., Simon, M., Rodner, E. and Denzler, J. (2015), Convolutional patch networks with spatial prior for road detection and urban scene understanding, *in* 'CVPR Scene Understanding Workshop'.

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A. and Zagoruyko, S. (2020), End-to-End Oobject Detection with Transformers, *in* 'European conference on computer vision', Springer, pp. 213–229.

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K. and Yuille, A. L. (2017), 'DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **40**(4), 834–848.

Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F. and Adam, H. (2018), Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation, *in* 'Proceedings of the European Conference on Computer Vision (ECCV)', pp. 833–851.

Chen, X., Yao, L. and Zhang, Y. (2020), Residual Attention U-Net for Automated Multi-Class Segmentation of Covid-19 Chest CT Images, *in* 'Proceedings of the Medical Image Computing and Computer Assisted Intervention–MICCA', pp. 405–414.

Choi, S., Kim, J. T. and Choo, J. (2020), Cars Can't Fly up in the Sky: Improving Urban-Scene Segmentation via Height-Driven Attention Networks, *in* 'Proceedings of the IEEE/CVF conference on computer vision and pattern recognition', pp. 9373–9383.

Choi, Y., Kim, N., Hwang, S., Park, K., Yoon, J. S., An, K. and Kweon, I. S. (2018), 'KAIST Multi-Spectral Day/Night Data Set for Autonomous and Assisted Driving', *IEEE Transactions on Intelligent Transportation Systems* **19**(3), 934–948.

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

Chu, Z., Tian, T., Feng, R. and Wang, L. (2019), Sea-Land Segmentation With Res-UNet And Fully Connected CRF, *in* 'IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium', IEEE, pp. 3840–3843.

Clark, D. R., Meffert, C., Baggili, I. and Breitinger, F. (2017), 'DROP (DRone Open source Parser) your drone: Forensic analysis of the DJI Phantom III', *Digital Investigation* **22**, S3–S14.

Cohen, A., Oswald, M. R., Liu, Y. and Pollefeys, M. (2017), Symmetry-Aware Facade Parsing with Occlusions, *in* '2017 International Conference on 3D Vision (3DV)', IEEE, pp. 393–401.

Cohen, A., Schwing, A. G. and Pollefeys, M. (2014), Efficient Structured Parsing of Facades Using Dynamic Programming, *in* 'Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition', pp. 3206–3213.

Commitee on Climate Change (2019), 'UK housing: Fit for the future?'. Available at: `https://www.theccc.org.uk/publication/uk-housing-fit-for-the-future/` (Accessed: 06 December 22).

Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S. and Schiele, B. (2016), The Cityscapes Dataset for Semantic Urban Scene Understanding, *in* 'Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition', pp. 3213–3223.

Dai, M., Ward, W. O., Meyers, G., Tingley, D. D. and Mayfield, M. (2021), 'Residential building facade segmentation in the urban environment', *Building and Environment* **199**, 107921.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K. and Fei-Fei, L. (2009), ImageNet: A large-scale hierarchical image database, *in* '2009 IEEE Conference on Computer Vision and Pattern Recognition', IEEE, pp. 248–255.

Densley Tingley, D. (2022), 'Embed circular economy thinking into building retrofit', *Communications Engineering* **1**(1), 1–4.

Department for Business, Energy & Industrial Strategy (2020), 'UK energy in brief 2020'. Available at: `https://www.gov.uk/government/statistics/uk-energy-in-brief-2020` (Accessed: 06 December 22).

Department for Business, Energy & Industrial Strategy (2021), 'Final UK greenhouse gas emissions national statistics: 1990 to 2019'. Available at: `https://www.gov.uk/government/statistics/final-uk-greenhouse-gas-emissions-national-statistics-1990-to-2019` (Accessed: 06 December 22).

Department of Transport (1996), 'The Rules of the Air Regulations 1996'. Available at: `https://www.legislation.gov.uk/uksi/1996/1393/made` (Accessed: 06 December 22).

Diakogiannis, F. I., Waldner, F., Caccetta, P. and Wu, C. (2020), 'ResUNet-a: A deep learning framework for semantic segmentation of remotely sensed data', *ISPRS Journal of Photogrammetry and Remote Sensing* **162**, 94–114.

Dimitrov, A. and Golparvar-Fard, M. (2014), 'Vision-based material recognition for automated monitoring of construction progress and generating building information modeling from unordered site image collections', *Advanced Engineering Informatics* **28**(1), 37–49.

Ding, X., Guo, Y., Ding, G. and Han, J. (2019), ACNet: Strengthening the Kernel Skeletons for Powerful CNN via Asymmetric Convolution Blocks, *in* 'Proceedings of the IEEE/CVF International Conference on Computer Vision', pp. 1911–1920.

DJI (2022), 'Zenmuse H20 Series – Unleash the Power of One - DJI'. Available at: `https://www.dji.com/cn/zenmuse-h20-series` (Accessed: 06 December 22).

Dong, X., Yu, Z., Cao, W., Shi, Y. and Ma, Q. (2020), 'A survey on ensemble learning', *Frontiers of Computer Science* **14**(2), 241–258.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S. et al. (2021), An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, *in* 'International Conference on Learning Representations'.

Dougherty, E. (1992), *An Introduction to Morphological Image Processing*, Books in the Spie Tutorial Texts Series, SPIE Optical Engineering Press.

Drozdzal, M., Vorontsov, E., Chartrand, G., Kadoury, S. and Pal, C. (2016), The Importance of Skip Connections in Biomedical Image Segmentation, *in* 'Deep Learning and Data Labeling for Medical Applications', Springer, pp. 179–187.

D'Ulizia, A., Ferri, F. and Grifoni, P. (2011), 'A survey of grammatical inference methods for natural language learning', *Artificial Intelligence Review* **36**(1), 1–27.

European Commission (2020), 'Energy performance of buildings directive'. Available at: `https://energy.ec.europa.eu/topics/energy-efficiency/energy-efficient-buildings/energy-performance-buildings-directive_en` (Accessed: 06 December 22).

Everingham, M., Van Gool, L., Williams, C. K., Winn, J. and Zisserman, A. (2010), 'The pascal visual object classes (voc) challenge', *International journal of computer vision* **88**, 303–338.

Femiani, J., Para, W. R., Mitra, N. and Wonka, P. (2018), 'Facade Segmentation in the Wild', *arXiv preprint arXiv:1805.08634* .

Fewins, C. (2004), 'The pros and cons of different construction systems', *Home Building & Renovating* pp. 1–11.

Fröhlich, B., Rodner, E. and Denzler, J. (2010), A fast Approach for Pixelwise Labeling of Facade Images, *in* '2010 20th International Conference on Pattern Recognition', IEEE, pp. 3029–3032.

Fröhlich, B., Rodner, E. and Denzler, J. (2012), Semantic Segmentation with Millions of Features: Integrating Multiple Cues in a Combined Random Forest Approach, *in* 'Asian Conference on Computer Vision', Springer, pp. 218–231.

Fröhlich, B., Rodner, E., Kemmler, M. and Denzler, J. (2013), 'Large-scale gaussian process multi-class classification for semantic segmentation and facade recognition', *Machine Vision and Applications* **24**(5), 1043–1053.

Fu, J., Liu, J., Tian, H., Li, Y., Bao, Y., Fang, Z. and Lu, H. (2019), Dual Attention Network for Scene Segmentation, *in* 'Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition', pp. 3146–3154.

Fukami, A. and Okamoto, K. (1984), 'Total heat exchanger'. US Patent 4460388A, Available at: `https://patents.google.com/patent/US4460388A/en` (Accessed: 06 December 22).

Gadde, R., Jampani, V., Marlet, R. and Gehler, P. V. (2017), 'Efficient 2D and 3D Facade Segmentation Using Auto-Context', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **40**(5), 1273–1280.

Gadde, R., Marlet, R. and Paragios, N. (2016), 'Learning Grammars for Architecture-Specific Facade Parsing', *International Journal of Computer Vision* **117**(3), 290–316.

Geiger, A., Lenz, P., Stiller, C. and Urtasun, R. (2013), 'Vision Meets Robotics: The KITTI Dataset', *The International Journal of Robotics Research* **32**(11), 1231–1237.

Geyer, J., Kassahun, Y., Mahmudi, M., Ricou, X., Durgesh, R., Chung, A. S., Hauswald, L., Pham, V. H., Mühlegg, M., Dorn, S. et al. (2020), 'A2d2: Audi Autonomous Driving Dataset', *arXiv preprint arXiv:2004.06320* .

Gong, F.-Y., Zeng, Z.-C., Zhang, F., Li, X., Ng, E. and Norford, L. K. (2018), 'Mapping sky, tree, and building view factors of street canyons in a high-density urban environment', *Building and Environment* **134**, 155–167.

Goodfellow, I., Bengio, Y. and Courville, A. (2016), *Deep learning*, Cambridge, MA: MIT press.

Great Britain (2019), *The Climate Change Act 2008 (2050 Target Amendment) Order 2019*, Statutory Instruments Series, Stationery Office. Available at: `https://www.legislation.gov.uk/ukdsi/2019/9780111187654/article/2` (Accessed: 06 December 22).

Guo, M.-H., Xu, T.-X., Liu, J.-J., Liu, Z.-N., Jiang, P.-T., Mu, T.-J., Zhang, S.-H., Martin, R. R., Cheng, M.-M. and Hu, S.-M. (2022), 'Attention mechanisms in computer vision: A survey', *Computational Visual Media* pp. 1–38.

Hao, S., Zhou, Y. and Guo, Y. (2020), 'A Brief Survey on Semantic Segmentation with Deep Learning', *Neurocomputing* **406**, 302–321.

Havaei, M., Davy, A., Warde-Farley, D., Biard, A., Courville, A., Bengio, Y., Pal, C., Jodoin, P.-M. and Larochelle, H. (2017), 'Brain tumor segmentation with Deep Neural Networks', *Medical Image Analysis* **35**, 18–31.

He, K., Gkioxari, G., Dollár, P. and Girshick, R. (2017), Mask R-CNN, *in* 'Proceedings of the IEEE International Conference on Computer Vision', pp. 2961–2969.

He, K., Zhang, X., Ren, S. and Sun, J. (2015), Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification, *in* 'Proceedings of the IEEE International Conference on Computer Vision', pp. 1026–1034.

He, K., Zhang, X., Ren, S. and Sun, J. (2016*a*), Deep Residual Learning for Image Recognition, *in* 'Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition', pp. 770–778.

He, K., Zhang, X., Ren, S. and Sun, J. (2016*b*), Identity Mappings in Deep Residual Networks, *in* 'European Conference on Computer Vision', Springer, pp. 630–645.

Hoegner, L. and Stilla, U. (2015), 'BUILDING FACADE OBJECT DETECTION FROM TERRESTRIAL THERMAL INFRARED IMAGE SEQUENCES COMBINING DIFFERENT VIEWS', *ISPRS Annals of Photogrammetry, Remote Sensing & Spatial Information Sciences* **2**(3/W4), 55–62.

Hong, T., Chen, Y., Luo, X., Luo, N. and Lee, S. H. (2020), 'Ten questions on urban building energy modeling', *Building and Environment* **168**, 106508.

Hornik, K., Stinchcombe, M. and White, H. (1989), 'Multilayer feedforward networks are universal approximators', *Neural Networks* **2**(5), 359–366.

Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M. and Adam, H. (2017), 'Mobilenets: Efficient Convolutional Neural Networks for Mobile Vision Applications', *arXiv preprint arXiv:1704.04861* .

Hu, J., Shen, L. and Sun, G. (2018), Squeeze-and-Excitation Networks, *in* 'Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition', pp. 7132–7141.

Hu, P. and Ramanan, D. (2017), Finding Tiny Faces, *in* 'Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition', pp. 1522–1530.

Hu, Q., Yang, B., Khalid, S., Xiao, W., Trigoni, N. and Markham, A. (2022), 'Sensaturban: Learning semantics from urban-scale photogrammetric point clouds', *International Journal of Computer Vision* **130**(2), 316–343.

Huang, G., Liu, Z., Van Der Maaten, L. and Weinberger, K. Q. (2017), Densely Connected Convolutional Networks, *in* 'Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition', pp. 4700–4708.

Huang, J., Rathod, V., Sun, C., Zhu, M., Korattikara, A., Fathi, A., Fischer, I., Wojna, Z., Song, Y., Guadarrama, S. et al. (2017), Speed/Accuracy Trade-Offs for Modern Convolutional Object Detectors, *in* 'Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition', pp. 3296–3297.

Huang, X., Cheng, X., Geng, Q., Cao, B., Zhou, D., Wang, P., Lin, Y. and Yang, R. (2018), The Apolloscape Dataset for Autonomous Driving, *in* 'Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops', pp. 954–960.

Ilehag, R., Schenk, A., Huang, Y. and Hinz, S. (2019), 'KLUM: An Urban VNIR and SWIR Spectral Library Consisting of Building Materials', *Remote Sensing* **11**(18), 2149.

Jampani, V., Gadde, R. and Gehler, P. V. (2015), Efficient Facade Segmentation Using Auto-Context, *in* '2015 IEEE Winter Conference on Applications of Computer Vision', IEEE, pp. 1038–1045.

Jin, Q., Meng, Z., Pham, T. D., Chen, Q., Wei, L. and Su, R. (2019), 'DUNet: A deformable network for retinal vessel segmentation', *Knowledge-Based Systems* **178**, 149–162.

Kelly, T., Femiani, J., Wonka, P. and Mitra, N. J. (2017), 'BigSUR: large-scale structured urban reconstruction', *ACM Transactions on Graphics* **36**(6).

Kendall, A., Badrinarayanan, V. and Cipolla, R. (2017), Bayesian segnet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding, *in* 'Proceedings of the British Machine Vision Conference (BMVC)', BMVA Press, pp. 57.1–57.12.

Kermi, A., Mahmoudi, I. and Khadir, M. T. (2018), Deep Convolutional Neural Networks Using U-Net for Automatic Brain Tumor Segmentation in Multimodal MRI Volumes, *in* 'International MICCAI Brainlesion Workshop', Springer, pp. 37–48.

Khanh, T. L. B., Dao, D.-P., Ho, N.-H., Yang, H.-J., Baek, E.-T., Lee, G., Kim, S.-H. and Yoo, S. B. (2020), 'Enhancing U-Net with Spatial-Channel Attention Gate for Abnormal Tissue Segmentation in Medical Imaging', *Applied Sciences* **10**(17), 5729.

Kingma, D. P. and Ba, J. (2014), 'Adam: A Method for Stochastic Optimization', *arXiv preprint arXiv:1412.6980* .

Kirillov, A., He, K., Girshick, R., Rother, C. and Dollár, P. (2019), Panoptic segmentation, *in* 'Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition', pp. 9404–9413.

Kong, G. and Fan, H. (2020), 'Enhanced Facade Parsing for Street-Level Images Using Convolutional Neural Networks', *IEEE Transactions on Geoscience and Remote Sensing* **59**(12), 10519–10531.

Kontschieder, P., Bulo, S. R., Bischof, H. and Pelillo, M. (2011), Structured class-labels in random forests for semantic image labelling, *in* '2011 International Conference on Computer Vision', IEEE, pp. 2190–2197.

Korč, F. and Förstner, W. (2009), eTRIMS Image Database for Interpreting Images of Man-Made Scenes, Technical Report TR-IGG-P-2009-01, Dept. of Photogrammetry, University of Bonn.

Kozinski, M., Gadde, R., Zagoruyko, S., Obozinski, G. and Marlet, R. (2015), A MRF shape prior for facade parsing with occlusions, *in* 'Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition', pp. 2820–2828.

Koziński, M., Obozinski, G. and Marlet, R. (2014), Beyond Procedural Facade Parsing: Bidirectional Alignment via Linear Programming, *in* 'Asian Conference on Computer Vision', Springer, pp. 79–94.

Krizhevsky, A., Sutskever, I. and Hinton, G. E. (2017), 'ImageNet classification with deep convolutional neural networks', *Communications of the ACM* **60**(6), 84–90.

Lambers, M. (2020), 'Survey of cube mapping methods in interactive computer graphics', *The Visual Computer* **36**(5), 1043–1051.

Li, J., Wei, Y., Liang, X., Dong, J., Xu, T., Feng, J. and Yan, S. (2016), 'Attentive Contexts for Object Detection', *IEEE Transactions on Multimedia* **19**(5), 944–954.

Li, X. and Tingley, D. D. (2021), 'Solid wall insulation of the Victorian house stock in England: A whole life carbon perspective', *Building and Environment* **191**, 107595.

Lin, G., Milan, A., Shen, C. and Reid, I. (2017), Refinenet: Multi-path Refinement Networks for High-Resolution Semantic Segmentation, *in* 'Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition', pp. 5168–5177.

Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B. and Belongie, S. (2017), Feature Pyramid Networks for Object Detection, *in* 'Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition', pp. 2117–2125.

Lin, T.-Y., Goyal, P., Girshick, R., He, K. and Dollár, P. (2017), Focal Loss for Dense Object Detection, *in* 'Proceedings of the IEEE International Conference on Computer Vision', pp. 2980–2988.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P. and Zitnick, C. L. (2014), Microsoft COCO: Common Objects in Context, *in* 'European Conference on Computer Vision', Springer, pp. 740–755.

Litwiller, D. (2001), 'CCD vs. CMOS', *Photonics spectra* **35**(1), 154–158.

Liu, H., Li, W. and Zhu, J. (2022), 'Translational Symmetry-Aware Facade Parsing for 3D Building Reconstruction', *IEEE MultiMedia* pp. 1 – 11.

Liu, H., Xu, Y., Zhang, J., Zhu, J., Li, Y. and Hoi, S. C. (2020), 'DeepFacade: A Deep Learning Approach to Facade Parsing with Symmetric Loss', *IEEE Transactions on Multimedia* **22**(12), 3153–3165.

Liu, X., Deng, Z. and Yang, Y. (2019), 'Recent progress in semantic image segmentation', *Artificial Intelligence Review* **52**(2), 1089–1106.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S. and Guo, B. (2021), Swin Transformer: Hierarchical Vision Transformer using Shifted Windows, *in* 'Proceedings of the IEEE/CVF International Conference on Computer Vision', pp. 9992–10002.

Loga, T., Stein, B. and Diefenbach, N. (2016), 'TABULA building typologies in 20 European countries—Making energy-related features of residential building stocks comparable', *Energy and Buildings* **132**, 4–12.

Long, J., Shelhamer, E. and Darrell, T. (2015), Fully convolutional networks for semantic segmentation, *in* 'Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition', pp. 3431–3440.

Lucon, O., Ürge-Vorsatz, D., Ahmed, A. Z., Akbari, H., Bertoldi, P., Cabeza, L. F., Eyre, N., Gadgil, A., Harvey, L. D., Jiang, Y. et al. (2014), Buildings, *in* 'Climate Change 2014: Mitigation of Climate Change. IPCC Working Group III Contribution to AR5', Cambridge University Press. Available at: `https://www.ipcc.ch/report/ar5/wg3/buildings/` (Accessed: 06 December 22).

Ma, W., Ma, W., Xu, S. and Zha, H. (2020), 'Pyramid ALKNet for Semantic Parsing of Building Facade Image', *IEEE Geoscience and Remote Sensing Letters* **18**(6), 1009–1013.

Ma, W., Xu, S., Ma, W. and Zha, H. (2020), 'Multiview Feature Aggregation for Facade Parsing', *IEEE Geoscience and Remote Sensing Letters* **19**.

Ma, W., Xu, S., Ma, W., Zhang, X. and Zha, H. (2022), 'Progressive Feature Learning for Facade Parsing With Occlusions', *IEEE Transactions on Image Processing* **31**, 2081–2093.

Ma, Z., Cooper, P., Daly, D. and Ledo, L. (2012), 'Existing building retrofits: Methodology and state-of-the-art', *Energy and Buildings* **55**, 889–902.

Maddern, W., Pascoe, G., Linegar, C. and Newman, P. (2017), '1 year, 1000 km: The Oxford RobotCar dataset', *The International Journal of Robotics Research* **36**(1), 3–15.

Mao, Z., Huang, X., Gong, Y., Xiang, H. and Zhang, F. (2022), 'A Dataset and Ensemble Model for Glass Façade Segmentation in Oblique Aerial Images', *IEEE Geoscience and Remote Sensing Letters* **19**, 1–5.

Martinović, A., Mathias, M., Weissenberg, J. and Gool, L. V. (2012), A three-layered approach to facade parsing, *in* 'Proceedings of the 12th European Conference on Computer Vision', Springer, pp. 416–429.

Martinovic, A. and Van Gool, L. (2013), Bayesian Grammar Learning for Inverse Procedural modeling, *in* 'Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition', pp. 201–208.

Mathias, M., Martinović, A. and Van Gool, L. (2016), 'ATLAS: A Three-Layered Approach to Facade Parsing', *International Journal of Computer Vision* **118**(1), 22–48.

Meyers, G., Zhu, C., Mayfield, M., Tingley, D. D., Willmott, J. and Coca, D. (2019), Designing a Vehicle Mounted High Resolution Multi-Spectral 3D Scanner: Concept Design, *in* 'Proceedings of the 2nd Workshop on Data Acquisition to Analysis', pp. 16–21.

Milletari, F., Navab, N. and Ahmadi, S.-A. (2016), V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation, *in* '2016 Fourth International Conference on 3D Vision (3DV)', IEEE, pp. 565–571.

Ministry of Housing, Communities & Local Government (2018), 'English Housing Survey, Households Report, 2017-18'. Available at: `https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/817286/EHS_2017-18_Households_Report.pdf` (Accessed: 06 December 22).

Mitchell, T. M. (1997), 'Machine learning', **1**(9). New York: McGraw-hill.

Mnih, V., Heess, N., Graves, A. and Kavukcuoglu, K. (2014), 'Recurrent Models of Visual Attention', *Advances in Neural Information Processing Systems* **27**.

Müller, R., Kornblith, S. and Hinton, G. E. (2019), 'When does label smoothing help?', *Advances in neural information processing systems* **32**.

Neuhausen, M. and König, M. (2018), 'Automatic window detection in facade images', *Automation in Construction* **96**, 527–539.

Neuhold, G., Ollmann, T., Rota Bulo, S. and Kontschieder, P. (2017), The Mapillary Vistas Dataset for Semantic Understanding of Street Scenes, *in* 'Proceedings of the IEEE International Conference on Computer Vision', pp. 4990–4999.

Ni, Z.-L., Bian, G.-B., Zhou, X.-H., Hou, Z.-G., Xie, X.-L., Wang, C., Zhou, Y.-J., Li, R.-Q. and Li, Z. (2019), RAUNet: Residual Attention U-Net for Semantic Segmentation of Cataract Surgical Instruments, *in* 'International Conference on Neural Information Processing', Springer, pp. 139–149.

Nixon, M. and Aguado, A. (2019), *Feature extraction and image processing for computer vision*, 4 edn, Academic press.

Noh, H., Hong, S. and Han, B. (2015), Learning Deconvolution Network for Semantic Segmentation, *in* 'Proceedings of the IEEE International Conference on Computer Vision', pp. 1520–1528.

Ohta, Y.-i., Kanade, T. and Sakai, T. (1978), An Analysis System for Scenes Containing objects with Substructures, *in* 'Proceedings of the Fourth International Joint Conference on Pattern Recognitions', pp. 752–754.

Oktay, O., Schlemper, J., Folgoc, L. L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N. Y., Kainz, B. et al. (2018), 'Attention U-Net: Learning Where to Look for the Pancreas', *arXiv preprint arXiv:1804.03999* .

Orhei, C., Vert, S., Mocofan, M. and Vasiu, R. (2021), TMBuD: a dataset for urban scene building detection, *in* 'International Conference on Information and Software Technologies', Springer, pp. 251–262.

O'Mahony, N., Campbell, S., Carvalho, A., Harapanahalli, S., Hernandez, G. V., Krpalkova, L., Riordan, D. and Walsh, J. (2019), Deep learning vs. traditional computer vision, *in* 'Science and information conference', Springer, pp. 128–144.

Peng, C., Zhang, X., Yu, G., Luo, G. and Sun, J. (2017), Large Kernel Matters–Improve Semantic Segmentation by Global Convolutional Network, *in* 'Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition', pp. 4353–4361.

Piddington, J., Nicol, S., Garrett, H. and Custard, M. (2020), 'The Housing Stock of the United Kingdom', *BRE Trust: Watford, UK* . Available at: `https://files.bregroup.com/bretrust/The-Housing-Stock-of-the-United-Kingdom_Report_BRE-Trust.pdf` (Accessed: 06 December 22).

Rahmani, K., Huang, H. and Mayer, H. (2017), 'FACADE SEGMENTATION WITH A STRUCTURED RANDOM FOREST', *ISPRS Annals of Photogrammetry, Remote Sensing & Spatial Information Sciences* **4**.

Rahmani, K. and Mayer, H. (2018), 'HIGH QUALITY FACADE SEGMENTATION BASED ON STRUCTURED RANDOM FOREST, REGION PROPOSAL NETWORK AND RECTANGULAR FITTING', *ISPRS Annals of Photogrammetry, Remote Sensing & Spatial Information Sciences* **4**(2).

Redmon, J. and Farhadi, A. (2018), 'YOLOv3: An incremental improvement', *arXiv preprint arXiv:1804.02767* .

Ren, S., He, K., Girshick, R. and Sun, J. (2015), 'Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks', *Advances in Neural Information Processing Systems* **28**.

RICS (2016), *Surveys of residential property, 3rd edition*, Royal Institution of Chartered Surveyors professional guidance, UK. Available at: `https://www.rics.org/globalassets/rics-website/media/upholding-professional-standards/sector-standards/building-surveying/surveys-of-residential-property-3rd-edition-reissue-rics.pdf` (Accessed: 06 December 22).

RICS (2021), *Earth observation and aerial surveys, 6th edition*, Royal Institution of Chartered Surveyors professional guidance, UK. Available at: `https://www.rics.org/uk/upholding-professional-standards/sector-standards/land/earth-observation-and-aerial-surveys-6th-edition-global-guidance-note/` (Accessed: 06 December 22).

Riemenschneider, H., Bódis-Szomorú, A., Weissenberg, J. and Van Gool, L. (2014), Learning Where to Classify in Multi-View Semantic Segmentation, *in* 'European Conference on Computer Vision', Springer, pp. 516–532.

Riemenschneider, H., Krispel, U., Thaller, W., Donoser, M., Havemann, S., Fellner, D. and Bischof, H. (2012), Irregular lattices for complex shape grammar facade parsing, *in* '2012 IEEE Conference on Computer Vision and Pattern Recognition', IEEE, pp. 1640–1647.

Robinson, R., Oktay, O., Bai, W., Valindria, V. V., Sanghvi, M. M., Aung, N., Paiva, J. M., Zemrak, F., Fung, K., Lukaschuk, E. et al. (2018), Real-Time Prediction of Segmentation Quality, *in* 'International Conference on Medical Image Computing and Computer-Assisted Intervention', Springer, pp. 578–585.

Ronneberger, O., Fischer, P. and Brox, T. (2015), U-Net: Convolutional Networks for Biomedical Image Segmentation, *in* 'International Conference on Medical Image Computing and Computer-Assisted Intervention', Springer, pp. 234–241.

Russell, B. C., Torralba, A., Murphy, K. P. and Freeman, W. T. (2008), 'LabelMe: a database and web-based tool for image annotation', *International Journal of Computer Vision* **77**(1), 157–173.

Schmitz, M. and Mayer, H. (2016), 'A convolutional network for semantic facade segmentation and interpretation', *The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences* **41**, 709.

Schroff, F., Criminisi, A. and Zisserman, A. (2008), Object Class Segmentation using Random Forests, *in* 'Proceedings of the British Machine Vision Conference (BMVC)', pp. 1–10.

Shao, H., Svoboda, T. and Van Gool, L. (2003), 'ZuBud-Zurich buildings database for image based recognition', *Technical report, Computer Vision Lab, Swiss Federal Institute of Technology, Switzerland* **260**(20), 6.

Simonyan, K. and Zisserman, A. (2015), Very Deep Convolutional Networks for Large-Scale Image Recognition, *in* 'International Conference on Learning Representations'.

Smith, R. (2011), *Surveying and assessing dwellings for low carbon retrofit*, Retrofit academy. Available at: `https://www.retrofitacademy.org/wp-content/uploads/2020/05/2-Surveying-and-Assessing-Dwellings-1.pdf` (Accessed: 06 December 22).

Sun, Y. and Gu, Z. (2022), 'Using computer vision to recognize construction material: A trustworthy dataset perspective', *Resources, Conservation and Recycling* **183**, 106362.

Syrris, V., Pesek, O. and Soille, P. (2020), 'Satimnet: Structured and harmonised training data for enhanced satellite imagery classification', *Remote Sensing* **12**(20), 3358.

Szegedy, C., Ioffe, S., Vanhoucke, V. and Alemi, A. A. (2017), Inception-v4, inception-ResNet and the impact of residual connections on learning, *in* 'Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence', pp. 4278–4284.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V. and Rabinovich, A. (2015), Going deeper with convolutions, *in* 'Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition', pp. 1–9.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. and Wojna, Z. (2016), Rethinking the Inception Architecture for Computer Vision, *in* 'Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition', pp. 2818–2826.

Tan, M. and Le, Q. V. (2019), MixConv: Mixed Depthwise Convolutional Kernels, *in* 'The British Machine Vision Conference (BMVC)'.

Tao, Y., Zhang, Y.-T. and Chen, X.-J. (2022), 'Element-Arrangement Context Network for Facade Parsing', *Journal of Computer Science and Technology* **37**(3), 652–665.

Teboul, O., Kokkinos, I., Simon, L., Koutsourakis, P. and Paragios, N. (2011), Shape grammar parsing via reinforcement learning, *in* 'CVPR 2011', IEEE, pp. 2273–2280.

Teboul, O., Kokkinos, I., Simon, L., Koutsourakis, P. and Paragios, N. (2012), 'Parsing Facades with Shape Grammars and Reinforcement Learning', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **35**(7), 1744–1756.

Teboul, O., Simon, L., Koutsourakis, P. and Paragios, N. (2010), Segmentation of building facades using procedural shape priors, *in* '2010 IEEE Computer Vision and Pattern Recognition', IEEE, pp. 3105–3112.

TELEDYNE FLIR (2021), 'Key differences between CCD and CMOS imaging sensors'. Available at: `https://www.flir.eu/support-center/iis/machine-vision/knowledge-base/key-differences-between-ccd-and-cmos-imaging-sensors/` (Accessed: 06 December 22).

Treue, S. (2003), 'Visual attention: the where, what, how and why of saliency', *Current Opinion in Neurobiology* **13**(4), 428–432.

Tu, Z. and Bai, X. (2009), 'Auto-Context and its Application to High-Level Vision Tasks and 3D Brain Image Segmentation', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **32**(10), 1744–1757.

Tyleček, R. and Šára, R. (2013), Spatial Pattern Templates for Recognition of Objects with Regular Structure, *in* 'Proceedgins of the German Conference on Pattern Recognition (GCPR)', Saarbrucken, Germany.

United Nations (2016), 'The Paris Agreement'. Available at: `https://unfccc.int/process-and-meetings/the-paris-agreement/the-paris-agreement` (Accessed: 06 December 22).

Van Etten, A. (2018), 'You only look twice: Rapid multi-scale object detection in satellite imagery', *arXiv preprint arXiv:1805.09512* .

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. and Polosukhin, I. (2017), Attention is all you need, *in* 'Proceedings of the 31st International Conference on Neural Information Processing Systems'.

Wang, J., Sun, K., Cheng, T., Jiang, B., Deng, C., Zhao, Y., Liu, D., Mu, Y., Tan, M., Wang, X. et al. (2020), 'Deep high-resolution representation learning for visual recognition', *IEEE transactions on pattern analysis and machine intelligence* **43**(10), 3349–3364.

Wang, S., Kang, Q., She, R., Tay, W. P., Navarro, D. N. and Hartmannsgruber, A. (2022), 'Building facade parsing R-CNN', *arXiv preprint arXiv:2205.05912* .

Woo, S., Park, J., Lee, J.-Y. and Kweon, I. S. (2018), CBAM: Convolutional Block Attention Module, *in* 'Proceedings of the European Conference on Computer Vision (ECCV)', pp. 3–19.

Xia, Y., Yabuki, N. and Fukuda, T. (2021), 'Development of a system for assessing the quality of urban street-level greenery using street view images and deep learning', *Urban Forestry & Urban Greening* **59**, 126995.

Yu, C., Wang, J., Gao, C., Yu, G., Shen, C. and Sang, N. (2020), Context Prior for Scene Segmentation, *in* 'Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern recognition', pp. 12416–12425.

Zeiler, M. D. and Fergus, R. (2014), Visualizing and Understanding Convolutional Networks, *in* 'European Conference on Computer Vision', Springer, pp. 818–833.

Zeppelzauer, M., Despotovic, M., Sakeena, M., Koch, D. and Döller, M. (2018), Automatic Prediction of Building Age from Photographs, *in* 'Proceedings of the 2018 ACM International Conference on Multimedia Retrieval', pp. 126–134.

Zhang, G., Pan, Y. and Zhang, L. (2022), 'Deep learning for detecting building façade elements from images considering prior knowledge', *Automation in Construction* **133**, 104016.

Zhang, Z., Wu, C., Coleman, S. and Kerr, D. (2020), 'DENSE-INception U-net for medical image segmentation', *Computer Methods and Programs in Biomedicine* **192**, 105395.

Zhao, H., Shi, J., Qi, X., Wang, X. and Jia, J. (2017), Pyramid Scene Parsing Network, *in* 'Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition', pp. 2881–2890.

Zhao, Z., Chen, K. and Yamane, S. (2021), CBAM-Unet++: easier to find the target with the attention module "CBAM", *in* '2021 IEEE 10th Global Conference on Consumer Electronics (GCCE)', IEEE, pp. 655–657.

Zheng, S., Lu, J., Zhao, H., Zhu, X., Luo, Z., Wang, Y., Fu, Y., Feng, J., Xiang, T., Torr, P. H. et al. (2021), Rethinking Semantic Segmentation from a Sequence-to-Sequence Perspective with Transformers, *in* 'Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition', pp. 6881–6890.

Zhou, L., Zhang, C. and Wu, M. (2018), D-LinkNet: LinkNet with Pretrained Encoder and Dilated Convolution for High Resolution Satellite Imagery Road Extraction, *in* 'Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops', pp. 182–186.

Zoph, B., Vasudevan, V., Shlens, J. and Le, Q. V. (2018), Learning transferable architectures for scalable image recognition, *in* 'Proceedings of the IEEE conference on computer vision and pattern recognition', pp. 8697–8710.

Zou, Y., Yu, Z., Kumar, B. and Wang, J. (2018), Unsupervised Domain Adaptation for Semantic Segmentation via Class-Balanced Self-Training, *in* 'Proceedings of the European Conference on Computer Vision (ECCV)', pp. 289–305.

Appendices

Google Street View Window-and-Door Dataset

Dataset

This appendix introduces a high-quality dataset labelled by the candidate to identify key building components focusing on windows and doors. The dataset was constructed in the year 2019. The dataset contains 300 annotated images and the annotation process takes about 20 days at a daily eight-hour working basis. The raw images were collected using the Google Street View service (Anguelov et al., 2010) in the Sheffield region. However, instead of using their API to download images automatically, the raw images collected are screen shots which are manually adjusted for views to be roughly perpendicular to facades and scaled to fit the majority of an image. The dataset will be made available online with the name 'Sheffield window and door' once the thesis is submitted to WhiteRose e-Theses online. Sample images are shown in Figure 3. The dataset is randomly split into 80% training and 20% validation.

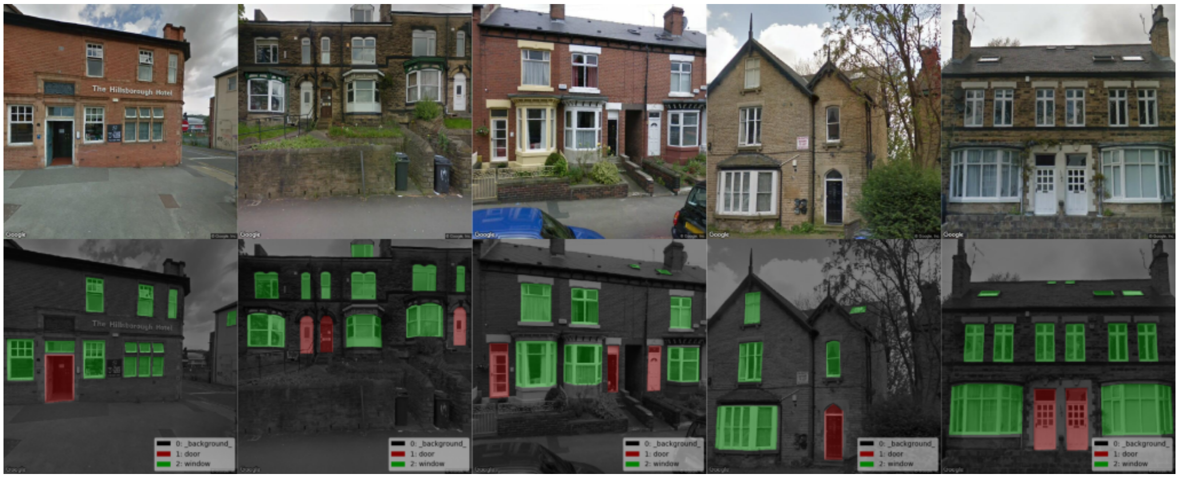


Figure 3: *The demonstration of 'Sheffield window and door' dataset.*

Benchmark Test

A benchmark test was conducted using the same model structure as in figure 3.14. Data augmentation setup includes a 50% chance of horizontal flip, 50% chance of 10% shift and 0.1 hue adjusting. The learning rate is set as 10^{-3}, the input resolution is 256×256, batch size is set to be 6. The training set is trained for 30 epochs on a laptop with 8GB RAM, an Intel i7-7700HQ CPU and an Nvidia GTX 1050 GPU. The benchmark test results are shown in table 2.

Table 2: *'Sheffield window and door' benchmark test results.*

Metrics / Category	Accuracy	Precision	TPR	TNR	F1 score
Window	0.979	0.933	0.908	0.990	0.921
Door	0.958	0.670	0.650	0.979	0.660

Summary

In summary, this appendix introduces a dataset manually labelled by the candidate. Building the dataset has three purposes:

1. Exploring annotation rules by viewing various window and door instances;
2. Estimating the time expenses of building datasets for facade segmentation. The time expense of building this dataset was used to evaluate whether an outsourcer was necessary for this PhD project;
3. Estimating the feasibility of using deep learning techniques on recognising building components. Therefore, great care was taken in annotating this dataset to ensure it is a high-quality dataset.

Results have shown that to build a facade segmentation dataset of over 1000 images, hiring outsourcers is essential. Using deep learning techniques on facade segmentation is seeming feasible. However, care should be taken in improving door prediction performance.

Google Street View Wall Construction Type Dataset

Introduction

This Appendix introduces a dataset created by the candidate for wall construction type classification. The dataset will be made available online with the name 'Merthyr Tydfil wall type classification' once the thesis is submitted to WhiteRose e-Theses online.

Wall construction types closely correlate to building energy cost and construction material stock. For example, cavity walls commonly have thermal insulation between the outer and inner leaves which provides better thermal performance than solid wall buildings and leads to smaller U-values. If wall construction types can be automatically categorised, the U-value of a specific building can then be inferred and the building's material stock can be more accurately predicted. The British energy performance certificates (EPCs) categorise wall types into five groups, 1) cavity, 2) solid brick, 3) sedimentary rock, 4) igneous rock and 5) timber frame. Theoretically, the five wall types can be simply differentiated by their visual features: stone walls should have different aesthetic features to brick walls, solid brick walls have different brick layout to cavity walls, etc. In previous works, efforts have been made on using image patches to identify construction materials (Dimitrov and Golparvar-Fard, 2014; Sun and Gu, 2022). However, in reality, the situation becomes more complicated of using image patches for material recognition.

The hardest problem is outer wall painting. If a wall is painted, its texture feature, which is the critical characteristic to be used for identifying wall types, could be largely covered. Therefore, whether other features on facades can be used to infer building wall types becomes critical to investigate. If without other features that could be used to identify wall construction types, raw images cannot be correctly labelled on painted walls. Although registering wall type data of EPCs with Google Street View facade images using geo-location information can produce an annotated wall type dataset, explainable machine learning is still significant to real-world applications and thus more evidence of whether using images to identify wall types need to be justified.

From a civil engineering perspective, wall construction types may possibly be correlated to building ages. For example, solid walls are common in Victorian buildings. Cavity walls started to become widespread from the 1920s (Fewins, 2004). Using facade images for building age detection has been studied and validated by Zeppelzauer et al. (2018). Therefore, if building wall construction types can be validated to have a correlation with building ages, using facade images on wall type recognition will stand at a more solid ground.

In this study, the residential EPCs of Merthyr Tydfil, a county borough in Wales, are collected and analysed. After filtering records without wall types, 21,207 records are achieved. Their wall type occupations of the whole stock and by age band are plotted in figure 4.

The whole stock occupation figure has shown that cavity walls dominate the Merthyr Tydfil building stock and timber frames are the least common. The occupation by age band figure has shown that solid walls including solid rock and brick were the dominant types before 1930, after which cavity walls became the majority. Therefore, based on the building age band distribution analysis, at least cavity walls and solid walls can potentially be recognised without fine wall fabric texture information.

Dataset Annotation and Benchmark Test Design

Next the collected EPCs are registered with Google Street View images using geo-location information. The resulted dataset is quite coarse. The first problem is that data capture points of Google Street View are sparse in some cases. It has been found that in some streets, only one image is captured and, therefore, all properties on that street are labelled the same. The second problem is that Google Street View API will return the nearest image of the designated address which does not ensure the nearest image is captured in front of target properties. It has been found in many cases, especially for rural areas, that Street View images do not contain any architecture. In order to avoid such duplication and capture location floating errors, all images in the dataset were manually inspected by the candidate.

During the manual dataset inspection process, it was found that the resulted timber frame buildings are duplicated significantly and their appearances show that they have cavity walls. Therefore, timber frame buildings are not included in the final dataset. Sedimentary and igneous rocks can be easily recognised in a museum since they have different aesthetic

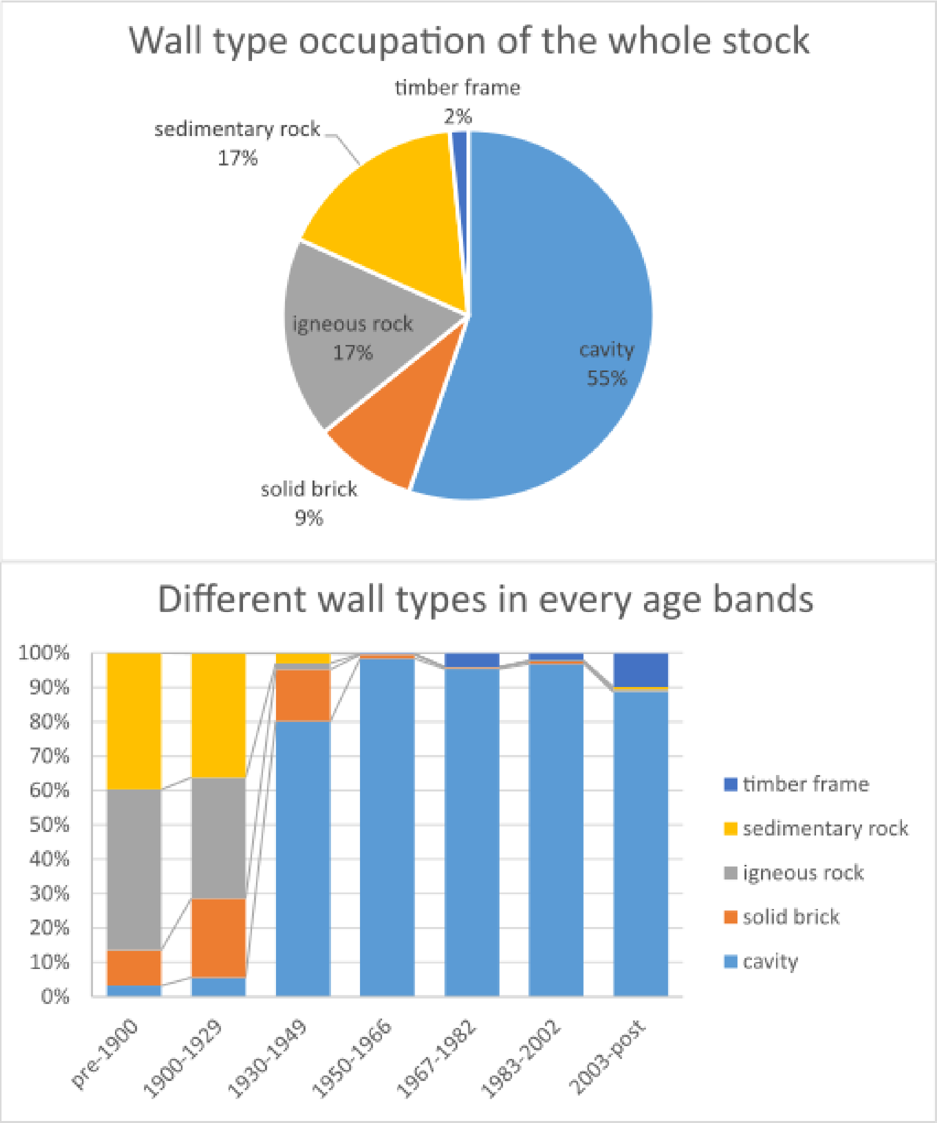


Figure 4: *Analysis of wall construction types of building stock in Merthyr Tydfil.*

appearances, however, using Street View images cannot properly differentiate them in all cases, at least by the candidate. Therefore, an extra label 'suspicious rock' is added for those images with less confident labels. The three rock wall classes were given the same summary label 'natural stone' as well. Cavity walls are the dominant wall types, to balance the number of cavity wall building images with other categories, the final version dataset does not include all inspected cavity wall house images. The statistics of the final version of the Merthyr Tydfil wall type dataset are shown in table 3. The dataset is randomly split into 80% training, 10% validation and 10% test, which results in a 7842-image training set, a 981-image validation set and a 980-image test set.

Table 3: *Merthyr Tydfil wall type classification task dataset statistics.*

Category	**Cavity**	**Igneous**	**Sedimentary**	**Suspicious rock**	**Solid brick**	**Total**
Num. of images	4412	2148	1968	92	1182	9802

A benchmark test was designed using four commonly-used classification models including a ResNet50 (He et al., 2016*a*), a DenseNet121 (Huang, Liu, Van Der Maaten and Weinberger, 2017), a NASNet-Large (Zoph et al., 2018), a VGG19 (Simonyan and Zisserman, 2015) and an Xception (Chollet, 2017). The test uses three classes: cavity, natural stone and solid brick. The test configurations are shown in table 4. A label smoothing strategy (Müller et al., 2019) is applied to labels in this test to prevent overfitting. The test is implemented using TensorFlow (Abadi, 2016) library and trained on a workstation with Windows 10, 16 GB RAM, an Intel Xeon E5-1620 v4 CPU and an NVIDIA Quadro P5000 GPU.

Table 4: *Test settings of Merthyr Tydfil wall type classification using chosen classification models.*

Data augmentation	**General settings**	**Weight Initialisation and loss function**	**Learning rate strategy**
50% chance of shifting horizontally or vertically with 10% range, 50% chance of horizontal flip, 0.1 hue adjusting	Adam optimiser with an input resolution 640 × 640, batch size=12 while 8 for DenseNet and trained 100 epochs	ImageNet weight initialisation and categorical cross entropy with label smoothing=0.1	starting at 10^{-3} and weight decay 10^{-5}, no early stop setting

Results and Discussion

The test has achieved the highest macro accuracy result in the ResNet50 and the NASNet-Large for 85% while the DenseNet and Xception have achieved similar results: 83% and 84%, respectively. The VGG19 does not learn anything from the dataset. The quantitative benchmark test results for all models are shown in table 5. These results have shown that all trained models can recognise cavity and stone walls with high confidence while they have difficulty in predicting solid brick walls.

Table 5: *'Merthyr Tydfil wall construction type classification' benchmark test results. The numbers in the second row represent different models: [1] is the ResNet50, [2] is the NASNet-Large, [3] is the DenseNet121, [4] is the Xception.*

Metrics Category	**Recall**				**Precision**				**F1 score**			
Model	[1]	[2]	[3]	[4]	[1]	[2]	[3]	[4]	[1]	[2]	[3]	[4]
Cavity	0.93	0.93	0.92	0.90	0.93	0.95	0.91	0.96	0.93	0.94	0.92	0.93
Nature stone	0.89	0.93	0.88	0.95	0.83	0.81	0.81	0.78	0.86	0.86	0.84	0.85
Solid brick	0.39	0.28	0.32	0.25	0.50	0.56	0.50	0.51	0.44	0.38	0.39	0.34
Mean	0.75	0.72	0.71	0.70	0.73	0.77	0.74	0.75	0.74	0.73	0.72	0.71

Figure 5 demonstrates confusion matrices of the benchmark test. These confusion matrices have shown that solid brick walls are frequently confused with stone walls. One reason could be a share of solid brick walls have stone-like appearances which might be faux-stone coverings. Furthermore, whether or not they are really solid brick-wall buildings is suspicious. EPCs might have survey errors and the registered Street View images might not be the right properties. Another reason could be what features a deep learning classification model was using to recognise wall types. If the model only uses features related to age predictions for wall type recognition, the model is likely to get confused predicting solid brick walls as the construction ages of solid wall is mixed with both stone and cavity walls.

As examples, the ResNet50 prediction results of buildings which are tagged as having a solid wall but with stone features are visualised in figure 6. The trained model predicts all examples of the first row and first two examples on the second row as 'natural stone'. However, even though the last two examples on the second row have stone textures, they are still categorised as 'solid brick'.

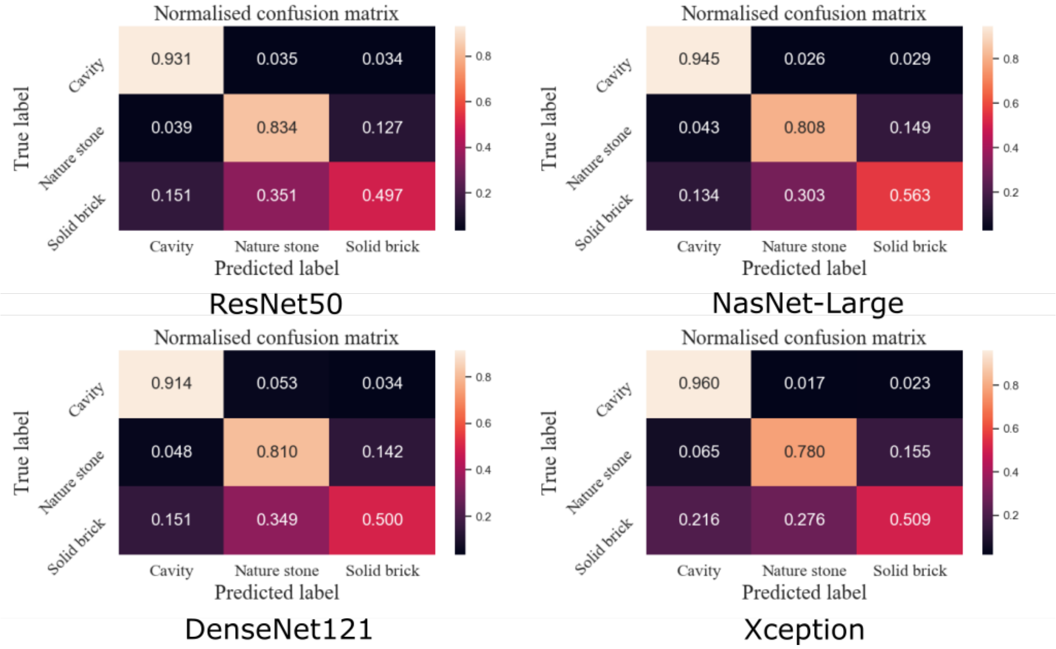


Figure 5: *The normalised confusion matrices of the benchmark test using a ResNet50, a NASNet-Large, a DenseNet121 and a Xception models.*

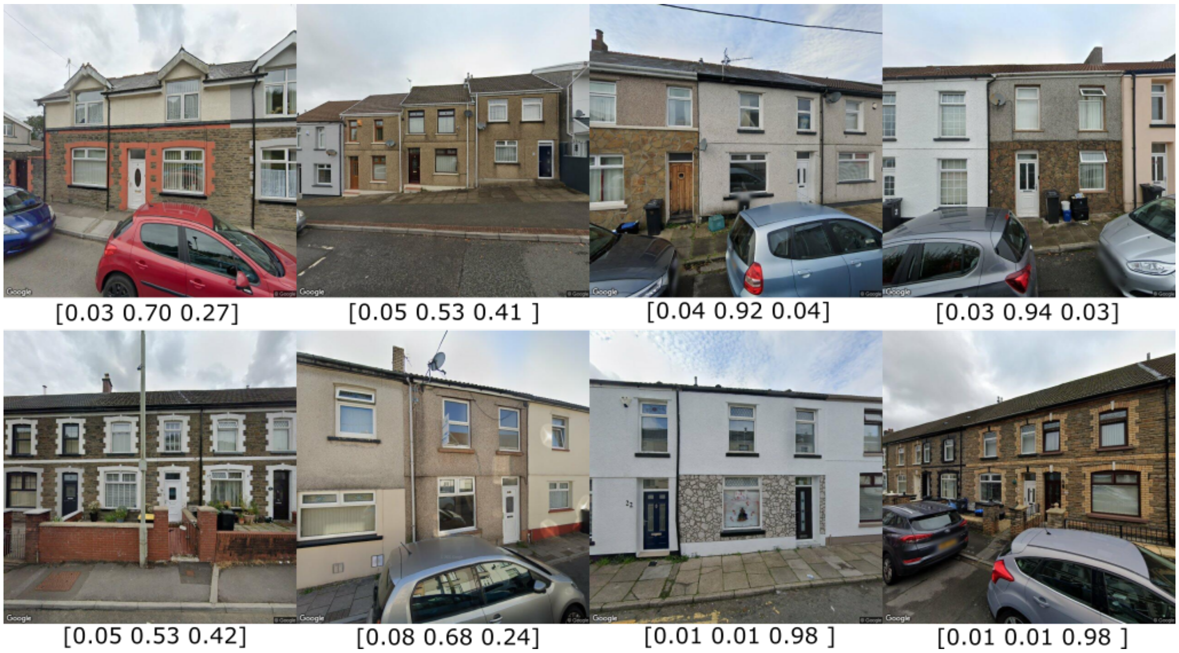


Figure 6: *The visualisation of model predictions on buildings which are tagged with 'solid brick' but have stone-made appearances. The three values below each image are probabilities of cavity, nature stone and solid brick.*

Conclusion

In summary, a house outer wall construction type dataset containing 9802 facade images has been built in this appendix. To the best of the candidate's knowledge, the dataset is the first that aims to use street view images for wall type predictions. The wall type ground truth data is provided by British EPC records and manually filtered by the candidate first. The benchmark test shows that deep learning models can differentiate cavity and stone walls with high confidence. However, they will meet problems when dealing with solid brick wall recognition. Potential reasons include the fact that there are no sufficient features on painted facades that can be used to identify wall types and EPC survey mistakes.