
Article:
Dai, M., Ward, W.O.C., Meyers, G. et al. (2 more authors) (2021) Residential building facade segmentation in the urban environment. Building and Environment, 199. 107921. ISSN 0360-1323

https://doi.org/10.1016/j.buildenv.2021.107921

Residential building facade segmentation in the urban environment

Menglin Dai*[1], Wil O.C. Ward[1], Gregory Meyers[2], Danielle Densley Tingley[1], and Martin Mayfield[1]

[1] *Department of Civil & Structural Engineering, The University of Sheffield*
[2] *Department of Aeronautical and Automotive Engineering, Loughborough University*

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. An advanced vehicle-mounted data capturing system has been built to collect urban environmental multi-spectral data. The data contains substantial information that is essential in identifying building retrofit needs. Although the data capturing system is able to collect data in a highly-efficient manner, the data analysis is still a big data challenge to apply the system into delivering building retrofit plans. In this paper, a street-view building facade image segmentation model is designed as the foundation of the holistic data analysis framework. The model is developed on the deep learning-based semantic segmentation technology and uses an ensemble learning strategy. The object detection technology is fused into the model as an magnifier to improve the model performance on small objects and boundary predictions. The model has achieved state-of-the-art levels of accuracy on a built street-view building facade image dataset.

Keywords:
building retrofit; building facade; deep learning; environmental modelling

1 Introduction

Under the circumstances of increasingly severe global climate change, reducing greenhouse gas (GHG) emissions is an inevitable worldwide challenge. To reach the Paris Agreement target [1], the UK government has committed to reducing the UK's net GHG emissions by 100% of their 1990 levels by 2050 [2]. Across all sectors contributing towards GHG emissions in 2019, residential buildings are responsible for 15% of the total GHG emissions [3] and consume 29% of the total energy [4] in the UK. In the meantime, the GHG reductions from the residential sector and the efforts of adapting the current housing stock for the climate change risks are stalled [3, 5]. In the situation, the Climate Change Committee in the UK has requested the housing retrofit to be an infrastructure priority [5].

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 conducting to collect required data to assess the building energy condition [6]. Such process collects the key building data such as building geometry, thermal characteristics e.g. 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. Such

*Corresponding author, Email: menglin.dai@sheffield.ac.uk, Address: Department of Civil and Structural Engineering, Sir Frederick Mappin Building, Mappin Street (via Broad Lane), Sheffield, S1 3JD

a highly labour-intensive and time-consuming process makes conducting building retrofit at large scale extremely difficult. 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 [7]. If we are to accelerate and scale the decarbonisation of the building stock, it is vital that we develop methods to automate the gathering of spatial information. 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 at large scale.

Vehicle-mounted sensors provide an efficient way to collect urban environmental data at scale. This type of data is widely used in many ways such as assisting land-use recognition [8] and automatically recognising building typology [9]. A famous example is the Google Street View [10] project which captures visual image data in the urban environment. More examples emerged in the past five years with additional point cloud and thermal image data available to support the autonomous driving research [11, 12, 13]. 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 [14]. The platform is vehicle-mounted and designed to collect visual images, thermal images, hyperspectral images and point clouds with high automation level. The platform is named at MARVEL (Multispectral Advanced Research VEhicLe) An demonstration image is shown in Figure 1. The collected data contains substantial building information which can be used to identify the building retrofit needs, e.g. point cloud data contains building geometry information, thermal image and hyperspectral image data can be used to identify the thermal faults and facade component material types.

Integrating the captured data and extracting building information with least manual intervention is essential to improving the efficacy and efficiency of large-scale building retrofit. The platform mentioned above is designed to capture data simultaneously, and the motion of the vehicle is recorded over its operation with a GNSS(Global Navigation Satellite System)/IMU(Inertial measurement unit) localisation/acceleration unit. The design ensures that the multi-spectral data can be integrated after collection. Recognising the building facades and their components in a visual image with high accuracy level is the foundation of acquiring other higher-level information. The pixel-wise building recognition removes the redundant information of the urban environment. By aligning the visual image data and other types of data, all unnecessary information can be removed for further data analysing.

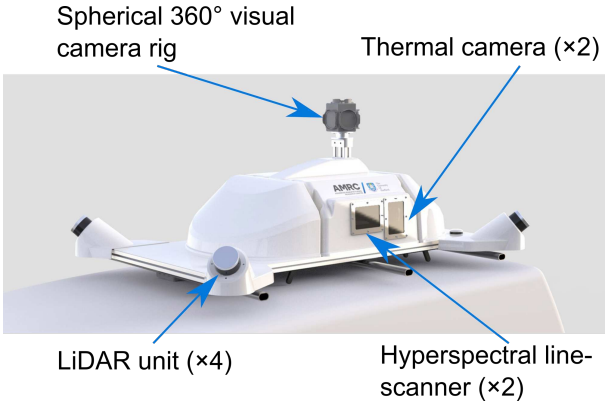

Figure 1: The developed multi-spectral data collection platform, MARVEL [14]. The visual camera rig is installed on the top of the platform; four LiDAR units are installed on each corner of the platform; thermal cameras and the hyperspectral line-scanners are installed on both sides of the platform.

The pixel-wise building recognition problem is not rare in the computer vision area. A classic topic referred to facade segmentation can be traced back to the 1970s [15]. The state-of-the-art approaches are dominantly based on deep learning-technology with extra thinking in using the structural features of buildings to improve the performance on predicting smaller objects e.g. windows and doors [16, 17, 18, 19]. This includes encoding the structural properties of buildings into models [17, 19] and loss functions [18]. Another approach is to use bounding box to refine the outlines of building components [16, 18]. This approach shows strong capability of using the bounding box and two-stage strategy in facade segmentation. These works show the powerful capability of deep learning models in this area and the necessity of taking extra care in refining the performance on predicting smaller facade components.

However, through research of literature under the typical facade segmentation topic in the past decade [16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26], we have identified a considerable difference between our task and the facade segmentation task. A building facade segmentation algorithm usually only concentrates on the building facade itself and without considering the interactions between the facade and the urban environment. The images used in this research area are usually with front views, and rectification & partition pre-processing. These interventions maintain the buildings' structural features, e.g. symmetry and straight outline. However, as we are using an automated image capturing system and our goal is to apply the method further to city-scale building analysis, the images we use are street-view images with substantially more diverse scenes, lighting conditions and viewing angles than state-of-art generic facade datasets [20, 27, 28, 29].

On the contrary, our task can be described as identifying the position of buildings in an urban scene image and classifying a building's main facade components. Efforts using urban scene images to locate buildings are also seen in the area of urban scene segmentation research. Urban scene segmentation is also a computer vision task that aims to assign a category label to each of the pixels within an image, for example a road, tree, or building. The technology is of significant interest in the area of autonomous vehicle research [30]. Well-cited public urban scene parsing datasets include ApolloScape [12], Cityscapes [31] and Mapillary Vistas [32]. The category 'building' appears in all of these datasets as it is an important component in the urban environment. However, although the building objects in the urban scene segmentation datasets are imaged with irregular viewpoints, such objects are only labelled as a whole rather than segmented into sub-components, e.g. window, door, wall etc. Therefore, the methods in the area did not focus on refining the model performance on recognising smaller building components particularly for windows and doors which are significant for energy modelling.

Multi-scale problems are a universal challenge of designing an urban scene segmentation model. The 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 [33], feature pyramid [34], and dilated convolution [35]. These methods show the obvious capability of deep learning-based models to solve the problem. However, these approaches are designed to be a universal solution of urban scene segmentation, and accordingly lack refinement for a certain scenario.

To summarise, reducing building GHG emissions is an urgent topic. Large-scale building retrofit plays an essential role in reducing building GHG emissions. The building energy modelling is essential in providing supportive information for making building retrofit plans at scale. The building data availability is still a barrier, although the energy modelling analytical process can be automated in the state-of-the-art. The developed multi-spectral collection platform enjoys high potentiality to provide key building information including geometry, material and thermal characteristics for the energy modelling process with a high automation level.

The data collected by the multi-spectral collection platform contains a significant amount of environmental noise irrelevant to buildings. To extract building information such as the

building geometry and thermal characteristics mentioned above, from the data collected, a facade segmentation model with accuracy priority is essential. The state-of-the-art related works in facade segmentation and urban scene segmentation area 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. The 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 provide many contributions to the multi-scale challenge. However, these works lack the component-level consideration in segmenting buildings.

The overall contribution of this work is the development of a new scalable approach to the automation of residential building facade component recognition. The solutions outlined in this paper are summarised thus:

1. a labelled dataset, focused on housing stock in a UK city, has been created, aimed at representing complex urban scenes at high resolution, with features that are not represented in existing datasets;
2. we have developed a novel ensemble segmentation model tailored to handle class imbalance, and to segment facade images with inter-category size discrepancies, such as between walls and windows;
3. we have expanded on the model to counter intra-category size discrepancies, e.g. due to perspective, by incorporating a novel magnifier strategy.

Section 2 describes the data collection process and the produced dataset, and the development of the facade image semantic segmentation models, FacMagNet. Section 3 demonstrates the performance of both models on the created dataset, showing their capabilities on images with different sized features, and handling miscellaneous urban furniture, such as on-street flora. Both models are benchmarked against state-of-the-art methods for semantic segmentation, that have previously been applied to both building facade segmentation problems, and more general segmentation problems. Finally, in Section 4, we discuss the importance of our model and the impact on building retrofit frameworks of using the proposed model as a foundation for identifying building retrofit needs.

2 Methodology

This section introduces the process of building the dataset and the study area. Two different semantic segmentation models derived from the U-Net architecture are designed based upon the characteristics of our dataset, and integrated in an ensemble way. The first model is designed to segment the relatively smaller sized features: windows, doors and chimneys. The second is designed for the features that take up a larger number of pixels: the wall and roof of a building.

2.1 Dataset and Study Area

Street view images were collected by the visual camera rig equipped on a multi-spectral data-capture system, as shown in Figure 1, The data-capture system is mounted on a vehicle and driven along residential streets. The visual camera rig comprises six separate Sony IMX264 CMOS sensors with 2048×2448 pixels resolution. 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.

A cube mapping technique is applied on the captured spherical-view images to map the data to six environmental mapping images, each with 2048×2048 [36]. The six images form a cube

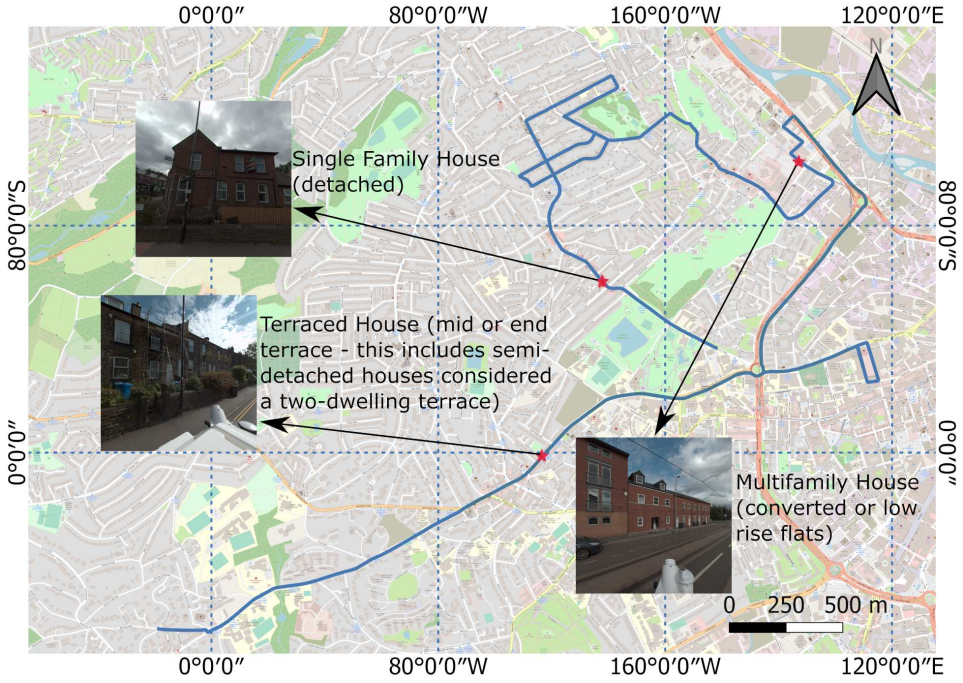

Figure 2: The data collection route is marked in blue. The route is selected in a typical suburb of the North of England residing in the outside of the Sheffield city centre. The route contains a wide range of residential building typologies defined in the TABULA database [37, 38]. Examples of the three main residential building typologies with corresponding descriptions are marked by the red stars.

covering the entire FOV with a front, right, back, left, top and bottom view. The top and bottom views, which predominantly show the sky, and road and sensors, respectively, were not used in training or prediction.

The images used for building the dataset were captured in the city of Sheffield, UK. The buildings in our dataset are visually matched with the British residential building typology database [37, 38]. The database classifies the building typologies based on building types, e.g. detached, terraced, and sub-classifies based on building age bands. The building age in the area ranges from 19th to the 21th 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 2 with examples of each building type highlighted, along with its corresponding location.

The dataset is built with 997 urban scene building images. The dataset is 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 [39, 40].

We categorise the facade image data into five classes: windows, doors, chimneys, roofs, and walls. 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 labelled regardless of whether they occur in the foreground or background. Examples of labelled facade images with these categories are shown in Figure 3.

The choice of categorisation is selected with consideration of the usage of positional information of their respective features, while maintaining a clear semantic taxonomy for which we have a

large number of training data. For example, segmenting walls in visible light images will allow the inference of building properties such as the total area of the external building facade, and the height of the building. With localisation of the windows on a facade, it is possible to infer the number of storeys, total window area and potentially room layout. Properties such as this are useful information in applications such as building energy modelling [41], and material stock analysis [42]. When considering multi-spectral sensing, as in the use of MARVEL, features that are easily identified in visible light images can also inform analysis of data from other streams: for example, understanding the thermal properties of different components with thermography capture can be simplified by having pre-localised features. Fault detection in glazing using thermal imaging could be easily automated given the locations of windows.

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 1.

Occlusion is an inevitable feature in the urban data captured. We employ two strategies, 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 the 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.

The collected images show extremely high complexity and variability in multiple ways, e.g. unexpected obstacles: billboards, antenna, trees, vehicles, etc., ambient light and in particular

Table 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 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

object-size variance.

2.2 Categorical Semantic Segmentation Models

Supervised deep learning-based semantic segmentation models are usually developed based on a fully convolutional neural network (FCN) [43]. These semantic segmentation models are a combination of a series of mathematical operations such as discrete convolution and pooling, and the overall aim is to minimise a designed loss function. The first FCN was designed initially as an image classification model, based on common models in the field. An image classification model is one that aims to assign a class label to an image based on its features. Convolutional neural network (CNN)-based classification models typically map an image to a feature vector by passing the outputs of the convolution layers through a fully connected neural network to create a vector output. However, the FCN replaces the fully connected layers with convolution layers that act as deconvolution operators. The deconvolution operations restore the output feature maps to the original input resolution, resulting in a class label corresponding to each pixel.

The spatial resolution of the feature maps, i.e. the outputs of each convolution layer, decreases throughout the feature extraction process. This allows the learned feature maps to be more invariant to small translations of the inputs. Consequently, the ratio of the input image resolution to the output feature map resolution, called downsampling rate, becomes a significant concern as redundant spatial resolution reductions will lead to target objects vanishing and insufficient resolution reduction may result in 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 [43].

The U-Net model is another semantic segmentation model, based on the FCN, that was developed initially for medical images [44]. The 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 [45, 46]. 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 [34], the symmetric architecture is able to retain small object information better. 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 our problem.

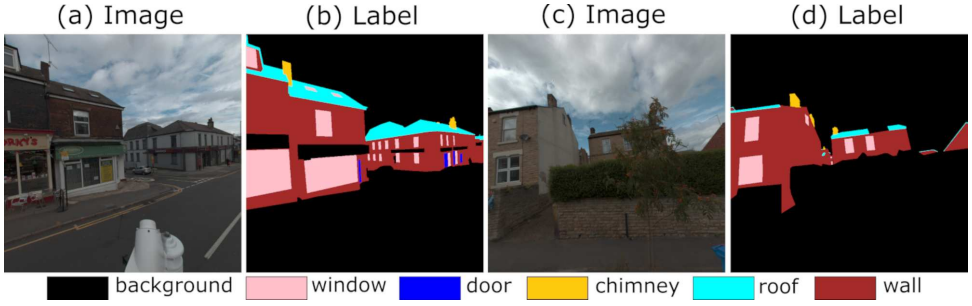

Figure 3: Annotation samples, all target objects appearing in an image are labelled regardless of their sizes, occlusions are avoided. Objects occluded by sparse obstacles are labelled depending on the obstacle density.

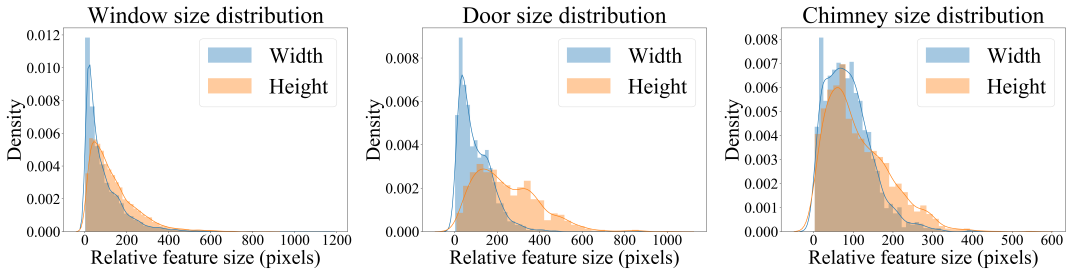

Figure 4: Relative feature size statistics of window, door and chimney under the raw data resolution; the plot shows the width and height distributions, the distributions show high varieties in sizes.

Another benefit of using the 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 [47]. Diffuse boundaries require low-level high-resolution edge information to refine the segmentation boundaries. In our facade image set, boundary ambiguity has been identified in all classes. Additionally, we have a degree of semantic information in the structure of a building, e.g. chimneys are typically located on roofs. This type of information is often found in medical images on which U-net has been found effective, such as human brain with defined interior structure [48]. The high-level semantic information can support the detection of the target objects.

Employing the original U-Net architecture directly to our facade segmentation dataset is ill-considered. The original U-Net takes inputs with resolution 572×572 and has a downsampling rate of 16. Our data has a much greater resolution, and has properties such as high size discrepancy, and class imbalance. As such, a new model is developed.

In our dataset, most of the wall objects can occupy the major area of an image and the roof objects are mainly slender shape 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 facts including viewing perspective. Through measuring the minimum bounding rectangles (MBR) size of the three smaller-size categories, the size distributions are plotted in Figure 4. These plots show that the objects of the three categories are distributed highly widely and unevenly. An effective method of solving the high size discrepancy problem is building several neural networks aiming for different scales [49]. Therefore, we have decided to use the ensemble learning strategy to build different models for different classes in this paper.

We have determined three different downsampling rates. For the three smaller-size categories, the downsampling rate is chosen to be 32, i.e. $log_2 32 = 5$ layers, which means the model will reduce the the feature map size to 32 times smaller than the input resolution. The decision is in terms of the largest objects in the three categories occupy a significantly larger proportion than the target objects in medical image 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 all significantly larger than the smaller-size categories and thus require deeper models to extract semantic information.

Figure 5 shows an example of the model structure with downsampling rate 32. The black arrows indicate the skip connection which is the operation to concatenate the feature maps in encoder network to the their symmetric ones in decoder network. As the encoder network increases the translate invariance of the model, it loses detailed edge information and location information. The skip connection are included to combine low-level features with high-level features, which helps the

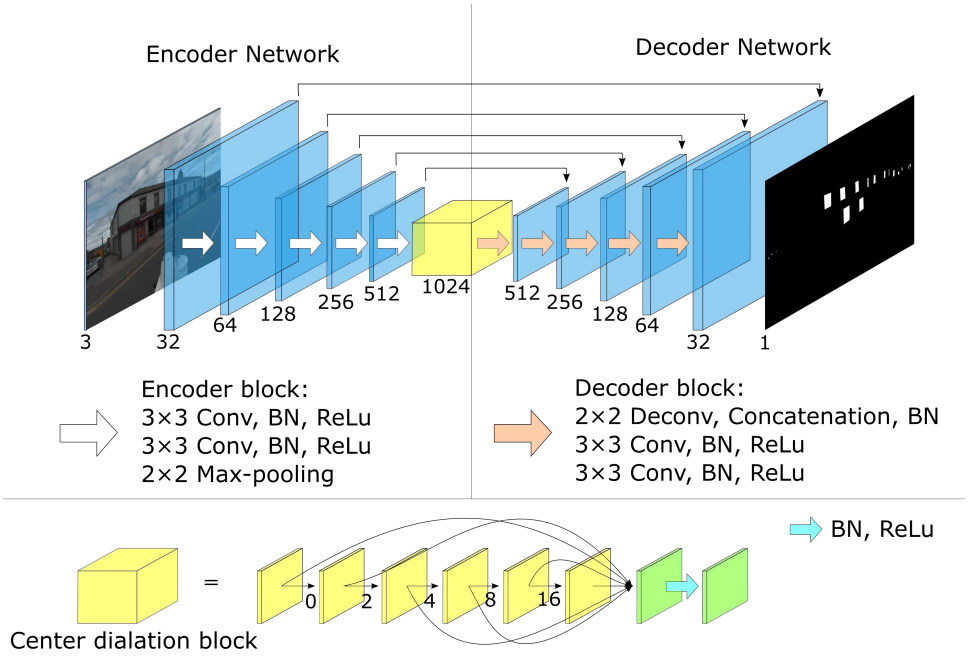

Figure 5: The semantic segmentation model for window, door and chimney categories, the 'Conv' stands for the convolution operation and 'BN' is the abbreviation of batch normalisation which is a common way to prevent over-fitting. The numbers in the encoder-decoder network represent the channel number and numbers in the dilation block are the dilation rates. The feature maps from the dilation convolution layers are added in the end with subsequent batch normalisation and activation layer in the centre dilation block.

model maintain information from different scales. In the three smaller-size 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 [50, 51], 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 5.

The detection of the roof and walls require higher downsampling rates, and the convolution layers in the model are replaced with residual blocks[52] to deal with 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 apart of the model structure, that determines how effective our 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 [53, 54].

In this work, we use a joint loss function 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 compared 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. We have used symmetric model structure and ensemble learning strategy 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.

2.3 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 [55]. 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, we have proposed a new method: 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 adopted for only the three category models where we observe small objects and class imbalance 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 [56]. However, as the design purpose is entirely 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 as same as the semantic segmentation technique. The technique is 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 [16, 18]. As in the front-view rectified facade images, objects such as windows and doors are commonly in rectangular shape. In our dataset, it is not possible to use the technique in a shape refinement manner, however we identified potential for integrating the technique in our task 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 normally is not parallel to the axes, the coordinates of the minimum rectangle which covers the MBR and parallel to the axes are calculated as the bounding box information.

An object detection model is trained to locate the bounding boxs of the three smaller-size 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

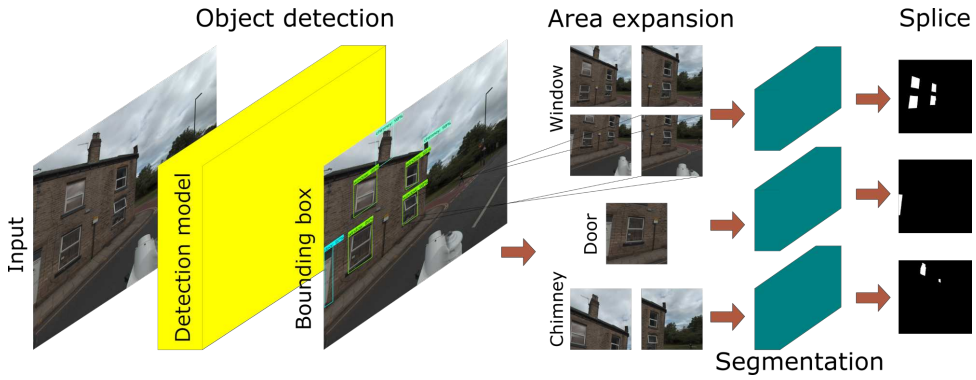

Figure 6: Model workflow with object detection; the input image passes into the detection model to generate bounding boxes first; the interested areas are then expanded and extracted; the extracted patches are magnified to a unified size then fed into the corresponding semantic segmentation models; the output score maps are resized to the original size and spliced together in the end.

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 6.

To learn and predict bounding boxes, the Faster R-CNN model is used [57]. 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 9 different anchor boxes for each point of in the feature maps and determines if each of the anchor boxes contains a target, along with their coordinates [57]. 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 [57], we adopt the Inception ResNet-V2 [58]. The anchor box ratios are fixed, as in the original paper, at 1:1, 1:2, and 2:1 [57]. The NMS threshold is fixed at 0.7.

3 Experiments

3.1 Training Strategy and Evaluation

Experiments are conducted in each category to explore the best combination among various choices discussed in Section 2. A base U-Net model is built for the purpose of comparison. For the roof and wall category, a deeper U-Net model with larger down-sampling rate, as well as a residual connection version of the deeper U-Net model. 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 is tested. Finally, the combined model is compared with existing state-of-the-art semantic segmentation models including Deeplab-v3plus [35], PSPNet [34], and SegNet [33]. Deeplab-v3plus is one of the top performing methods in the PASCAL VOC semantic segmentation challenge [59]. PSPNet and SegNet were introduced in the literature review was two models used for urban scene segmentation task.

To maintain detail of 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 feeding into the smaller category models. A data augmentation technique is used during model training: geometric transformations and colour adjustments were applied to the base dataset to produce a larger training set. Horizontal mirroring, vertical & horizontal translations and small rotations were applied randomly to 50% of the data. The hue of the images were adjusted randomly by up to 10%.

The adaptive moment estimator (Adam) optimiser with a learning rate reduction strategy was used [60]. 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 [61].

Categorical models performances are evaluated both qualitatively and quantitatively. Qualitative evaluation is based on visual inspection, and the quantitative include use of a confusion matrix and comparative evaluation metrics on component models. We use accuracy, precision, recall, true negative rate (TNR), intersection of union (IOU) and the F1 score to indicate the quality of models. Each of these metrics relies true positive, true negative, false positives 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 widely used to measure the overall model performance by considering the impact of the both precision and recall values. The ensemble model will be evaluated by 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 where implemented with the TensorFlow library [62]. All models were trained a workstation with Windows 10, 16GB RAM, an Intel Xeon E5-1620 v4 CPU and an NVIDIA Quadro P5000 GPU.

3.2 Categorical Model Evaluation

Metrics for the small component segmentation models are given in Table 2. Both our proposed model architectures outperform the base U-Net structure. Looking at the F1 score and IOU metrics, we find 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 a same image, this phenomenon indicates that the magnifier integration model tends to predict more positive pixels. Besides, the results show 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 to be used to compare model performances in this task.

The qualitative analysis demonstrates the same overall results. Examples of segmentations are shown in for the detection of doors and windows in Figure 7 and Figure 8 respectively. The magnifier integration model generally shows better performance in handling boundaries and small objects.

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 8 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 8(b), and these kind of surfaces widely exist on building facades, the tendency lowers its overall quantitative performance.

The evaluation metrics for the roof and wall categories with different downsampling rates are shown in Table 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 9.

3.3 Ensemble Model Evaluation

Based on the findings in the evaluation of each categorical model, the model ensemble, FacMagNet, uses our magnifier integration model for the three smaller-sized categories, because of its

Table 2: Smaller-size categories segmentation performance; The 'U-Net 32' is the base U-Net with downsampling rate 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

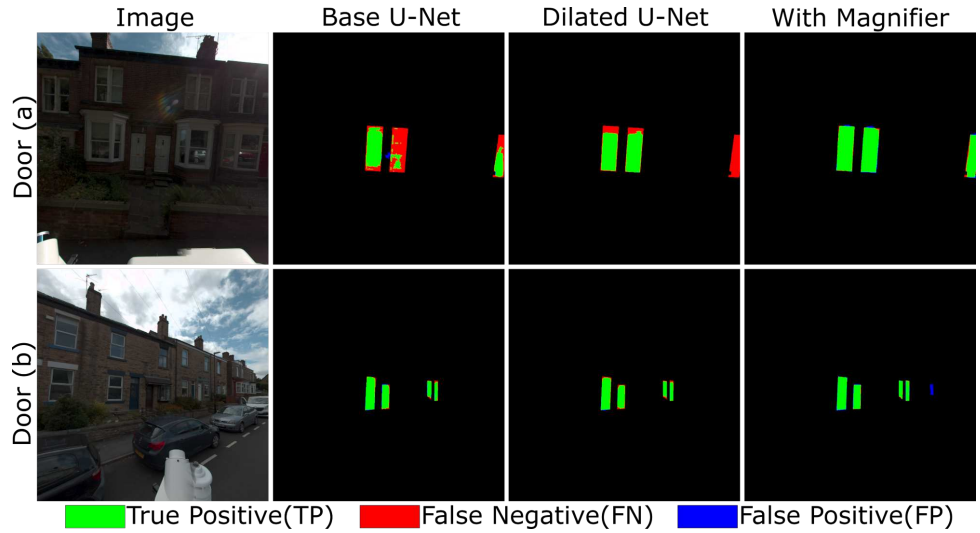

Figure 7: Door qualitative examples; (a) clearly shows the performance improvements and (b) shows the object detection integration model predicting the object without annotation.

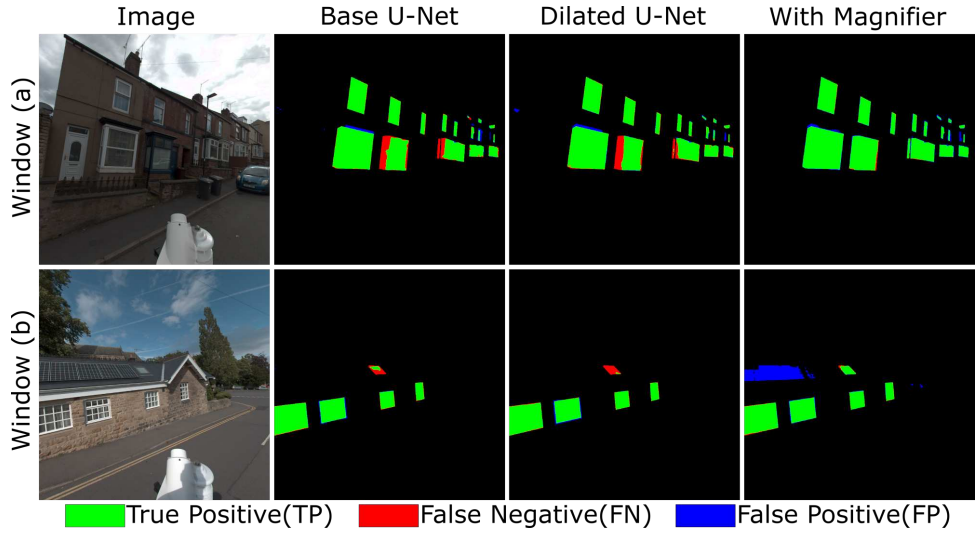

Figure 8: Window qualitative examples; (a) shows the performance improvements and (b) shows the object integration model improving the performance but also recognise the solar panel as a window

advantages in boundary and small-object predictions. The base U-Net for the roof category and our proposed residual U-Net 128 for the wall are selected due to sharing highest IOU and F1 values in each category.

Figure 10 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 due to wrongly classifying pixels belonging to objects as background. Walls are the category to which the model will incorrectly assign pixels second-most often.

Our proposed FacMagNet is compared with using other models which are widely adopted in the semantic segmentation area, the categorical IOU values of each model are shown in Table 4. The mean IOU (mIOU) is calculated by computing the IOU average of all classes excluding the background.

From Table 4, it is clear that our 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 11 shows that our FacMagNet visually achieves high performance

Table 3: Roof & Wall segmentation performance; The 'U-Net' '64' and '128' means using the U-Net structure with downsampling rate 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**

segmentations, even when dealing with high-distortions, small-objects and obstacles. We can also see from Figure 11 that our approach can easily handle segmentation, even when the components are partially occluded by objects such as trees and fences.

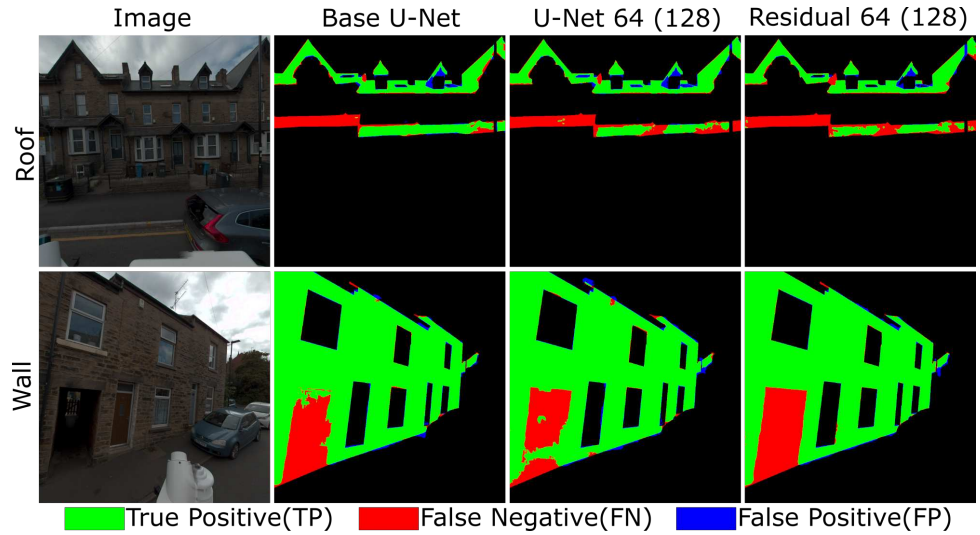

Figure 9: Roof & Wall qualitative examples, the number in bracket indicates the down-sampling ratio of the wall category model; the roof category is more suitable for the base model and the residual model is more friendly to the wall category.

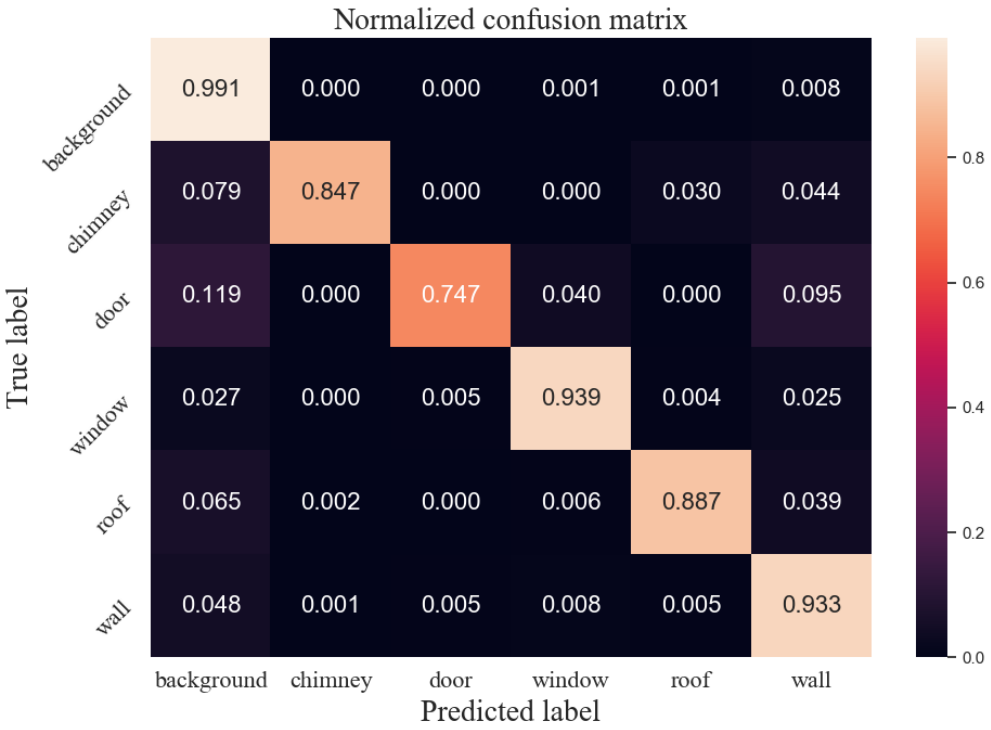

Figure 10: Normalised ensemble model confusion matrix, the confusion matrix is normalised by dividing the sum of the ground-truth pixels in each categories, the diagonal show the percentage of the correctly predicted pixels over the sum of corresponding ground-truth pixels.

Table 4: Categorical IOU, the first row is the metrics of using our developed FacMagNet, the second row is using the base U-Net with ensemble strategy. The third, fourth, fifth and sixth rows are using Deeplab-v3plus, U-Net, PSPNet, and SegNet models, correspondingly, as multi-class classifiers.

	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
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

4 Discussion and Future work

4.1 The FacMagNet model

In this paper, a building facade semantic segmentation model is developed to recognise residential building facades in component level. The model has been carefully designed to detect the features of residential buildings from street-level imaging. A key characteristic of our data is that it contains a substantial number of small objects, and there is a high size discrepancy both between different classes and within the same class. Our proposed model employs a symmetric structure, dilated convolution and an ensemble learning strategy, as well as a magnifier to handle class imbalance. The results presented in Section 3 demonstrate the efficacy of our model on facade segmentation against contemporary and state-of-the-art models.

One drawback of our model, due to the ensemble nature of the model, is the high time and computational resource requirements, both for the training and more importantly for prediction. As we integrate six individual models with differing architecture, our model requires more

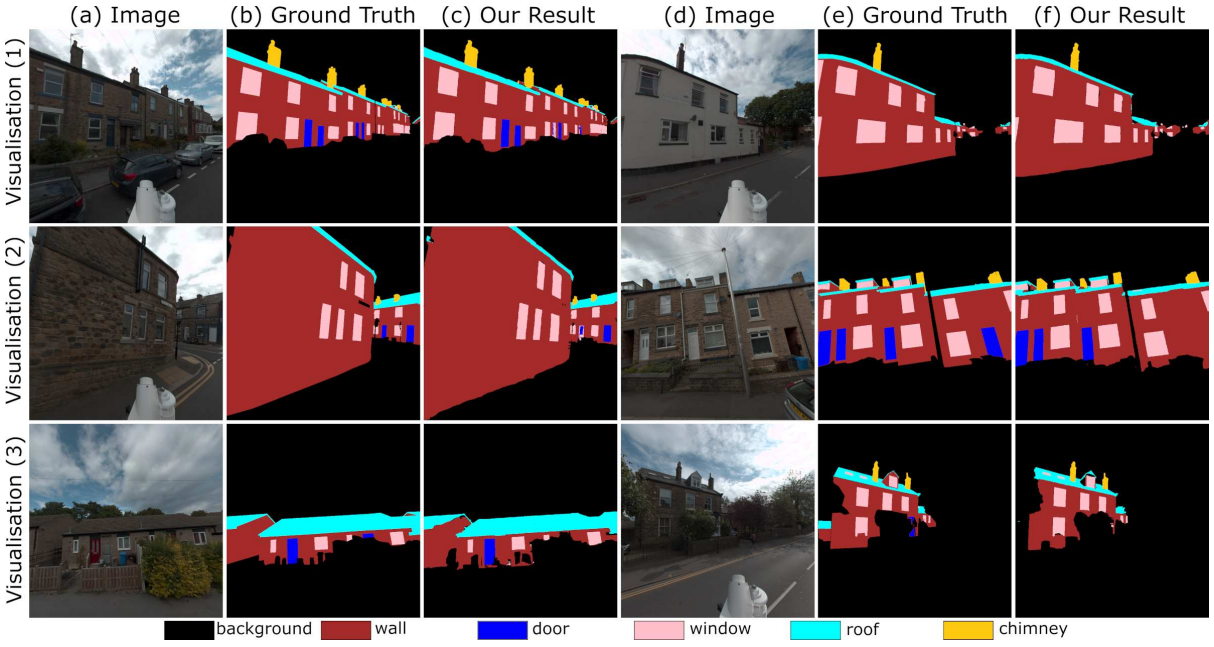

Figure 11: Qualitative examples of the FacMagNet model; the visual results show our model has achieved high accuracy in large object, and is friendly in handling small-object and occlusion problems.

resources compared with using an end-to-end model, such as the Deeplab-v3plus model [35]. 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 our proposed model.

4.2 Facade segmentation as a foundation for scalable retrofit

As discussed in the Section 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, 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.

MARVEL realises a highly efficient approach 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. For example, from a visible-light image, we can segment the wall and extract directly properties such as total area, height and number of stories. Incorporating hyperspectral information, it may be possible to characterise 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.

As we have created an accurate process for segmenting features on a building facade, we can reliably make inferences on the properties such as those outlined in Table 1. Performing this in a scalable manner is the first step towards automating or semi-automating the development of retrofit solutions for residential buildings. The model has been shown to give high quality segmentations of all components, 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 MARVEL 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 also be used to evaluate the building natural illumination and ventilation conditions. For example, a building with low glazing ratio might need the introduction of synthetic ventilation, such as the THEX (Total Heat EXchanger) [63] during retrofit.

The thermal images and hyperspectral images are to determine the thermal performance and material types of buildings, respectively. The use of integration of point-cloud and infrared thermography data to detect the thermal leakage was previously explored in [64]. With the integration of the facade segmentation approach, the certain building components of thermal leakages can be determined. This will contribute to a more precise retrofitting plan in terms of material replacement or strengthening. The spectral information can be used to identify building materials [65]. Incorporating the hyperspectral data with the integrated 3D model could realise accounting building stock. Compared to the traditional methods, this method has less

constraints since it does not require historic 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.

The facade segmentation information also can be directly applied to contribute towards building energy modelling. [66] automatically measure the view factor of street canyons with using the Google street view data and deep learning-based semantic segmentation model. The model choice in this paper is not component-level oriented, thus lacking attentions on small object problem.

4.3 Future work

The holistic target of future work is to realise an automatic, scalable, comprehensive, building analysis system towards efficient scalable retrofit solutions. The system will largely contribute towards renovating the existing building analysis approaches with fewer data constraints and higher efficiency. We identify some key areas towards which our proposed model can contribute.

Given the nature of the data capture, with both visible-light photography and LiDAR capture, a 3-D representation of the buildings can be constructed with real-world dimensions. Multiple perspectives, as the capture moves past a building, can be utilised to use multi-view stereo photogrammetry techniques [67, 68]. Alignment of projected 3-D models with LiDAR can serve for validation. Because the semantic segmentation model proposed accurately localises features at a pixel-level, these labels can be projected onto the photogrammetry models. A labelled 3-D model can be used to extract dimensional properties of features that can be used to infer a great deal of information about a property, as discussed in Table 1.

Multi-spectral capture, as we are able to perform with MARVEL, allows for co-registered information on buildings. Thermal profiles of a building can indicate a number of features, such as insulation quality, as well potential faults or thermal leaks. Localising objects within the building front allows for better understanding of a thermal image: for example, the difference between the thermal bridge of a window and wall could cause incorrect inferences without the context that they are two different objects, with different material properties. The information we can obtain from segmenting visible-light images can help automatically extract relevant context to fine tune analysis.

Isolating objects on a building facade is also essential when characterising materials. With hyperspectral data, a distribution of properties representing material properties can be identified. Knowledge of material, and even glazing, can impact the effectiveness of building energy models [7, 41]. The information characterised about a building can also be used to build stock models, the creation of which would greatly benefit from scalable solutions [42]. Beyond retrofit, knowledge of building properties on a scalable level can be useful in a number of urban applications. For example, radio signals can be negatively impacted by both building height and materials, which can affect indoor signal [69]. Understanding the nature of urban environments with accurate information can aid in modelling network coverage, which may inform distribution needs.

4.4 Conclusions

In this paper, we have collected and built a novel street-view building facade images dataset focusing on the UK residential housing stock. A data labelling framework was identified, considering potential needs in assisting scalable building retrofit. We have also presented a novel ensemble model for the semantic segmentation of building facade components, termed at FacMagNet. The model is purpose build for the task, utilising contemporary deep learning architectures and utilising an ensemble learning strategy to best categorise each object. We have demonstrated that it can effectively and accurately label images at a level that exceeds other state-of-the-art models for the given task.

Along with the development of the model and evaluation on urban street-level data, we have identified clear motivation 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.

Acknowledgements

We thank Maud Lanau, Richard Johnson, Xinyi Li and Oktay Cetinkaya for their insights into applications of localised features to retrofit and network modelling.

The multispectral advanced research vehicle (MARVEL) was supported under EPSRC UK Collaboratorium for Research in Infrastructure & Cities: Urban Observatories (Strand B) [EP/P016782/1]; MD and WOCW were funded by the EPSRC Active Building Centre [EP/S016627/1] and Active Building Centre Research Programme [EP/V012053/1]; GM was supported under EPSRC UKCRIC-CORONA: City Observatory Research Platform for Innovation and Analytics [EP/R013411/1].

References

[1] United Nations, "The Paris Agreement," 2016. [Online]. Available: https://unfccc.int/process-and-meetings/the-paris-agreement/the-paris-agreement

[2] Great Britain, *The Climate Change Act 2008 (2050 Target Amendment) Order 2019*, ser. Statutory Instruments Series. Stationery Office, 2019. [Online]. Available: https://www.legislation.gov.uk/ukdsi/2019/9780111187654/article/2

[3] Department for Business, Energy & Industrial Strategy, "Final UK greenhouse gas emissions national statistics: 1990 to 2019," 2021. [Online]. Available: https://www.gov.uk/government/statistics/final-uk-greenhouse-gas-emissions-national-statistics-1990-to-2019

[4] ——, "UK energy in brief 2020," 2020. [Online]. Available: https://www.gov.uk/government/statistics/uk-energy-in-brief-2020

[5] Commitee on Climate Change, "UK housing: Fit for the future?" 2019. [Online]. Available: https://www.theccc.org.uk/wp-content/uploads/2019/02/UK-housing-Fit-for-the-future-CCC-2019.pdf

[6] Z. Ma, P. Cooper, D. Daly, and L. Ledo, "Existing building retrofits: Methodology and state-of-the-art," *Energy and Buildings*, vol. 55, pp. 889 – 902, 2012, cool Roofs, Cool Pavements, Cool Cities, and Cool World. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0378778812004227

[7] T. Hong, Y. Chen, X. Luo, N. Luo, and S. H. Lee, "Ten questions on urban building energy modeling," *Building and Environment*, vol. 168, p. 106508, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0360132319307206

[8] J. Kang, M. Körner, Y. Wang, H. Taubenböck, and X. X. Zhu, "Building instance classification using street view images," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 145, pp. 44–59, 2018, deep Learning RS Data. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0924271618300352

[9] D. Gonzalez, D. Rueda-Plata, A. B. Acevedo, J. C. Duque, R. Ramos-Pollán, A. Betancourt, and S. García, "Automatic detection of building typology using deep learning methods on street level images," *Building and Environment*, vol. 177, p. 106805, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0360132320301633

[10] D. Anguelov, C. Dulong, D. Filip, C. Frueh, S. Lafon, R. Lyon, A. Ogale, L. Vincent, and J. Weaver, "Google street view: Capturing the world at street level," *Computer*, vol. 43, no. 6, pp. 32–38, 2010.

[11] W. Maddern, G. Pascoe, C. Linegar, and P. Newman, "1 year, 1000 km: The oxford robotcar dataset," *The International Journal of Robotics Research*, vol. 36, no. 1, pp. 3–15, 2017. [Online]. Available: https://doi.org/10.1177/0278364916679498

[12] X. Huang, X. Cheng, Q. Geng, B. Cao, D. Zhou, P. Wang, Y. Lin, and R. Yang, "The apolloscape dataset for autonomous driving," *CoRR*, vol. abs/1803.06184, 2018. [Online]. Available: http://arxiv.org/abs/1803.06184

[13] Y. Choi, N. Kim, S. Hwang, K. Park, J. S. Yoon, K. An, and I. S. Kweon, "Kaist multi-spectral day/night data set for autonomous and assisted driving," *IEEE Transactions on Intelligent Transportation Systems*, vol. 19, no. 3, pp. 934–948, 2018.

[14] G. Meyers, C. Zhu, M. Mayfield, D. D. Tingley, J. Willmott, and D. Coca, "Designing a vehicle mounted high resolution multi-spectral 3d scanner: Concept design," in *Proceedings of the 2nd Workshop on Data Acquisition To Analysis*, ser. DATA'19. New

York, NY, USA: Association for Computing Machinery, 2019, p. 16–21. [Online]. Available: https://doi.org/10.1145/3359427.3361921

[15] Y.-i. Ohta, T. Kanade, and T. Sakai, "An Analysis System for Scenes Containing objects with Substructures - The Robotics Institute Carnegie Mellon University," in *Proceedings of the Fourth International Joint Conference on Pattern Recognitions*, Jan 1978, pp. 752–754.

[16] K. Rahmani and H. Mayer, "High quality facade segmentation based on structured random forest, region proposal network and rectangular fitting," in *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 4, no. 2. Copernicus GmbH, may 2018, pp. 223–230.

[17] J. Femiani, W. R. Para, N. J. Mitra, and P. Wonka, "Facade segmentation in the wild," *CoRR*, vol. abs/1805.08634, 2018. [Online]. Available: http://arxiv.org/abs/1805.08634

[18] H. Liu, Y. Xu, J. Zhang, J. Zhu, Y. Li, and C. S. Hoi, "DeepFacade: A Deep Learning Approach to Facade Parsing with Symmetric Loss," *IEEE Transactions on Multimedia*, pp. 1–1, feb 2020.

[19] W. Ma, W. Ma, S. Xu, and H. Zha, "Pyramid alknet for semantic parsing of building facade image," *IEEE Geoscience and Remote Sensing Letters*, pp. 1–5, 2020.

[20] H. Riemenschneider, U. Krispel, W. Thaller, M. Donoser, S. Havemann, D. Fellner, and H. Bischof, "Irregular lattices for complex shape grammar facade parsing," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2012, pp. 1640–1647.

[21] O. Teboul, I. Kokkinos, L. Simon, P. Koutsourakis, and N. Paragios, "Parsing facades with shape grammars and reinforcement learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 7, pp. 1744–1756, 2013.

[22] V. Jampani, R. Gadde, and P. V. Gehler, "Efficient facade segmentation using auto-context," in *Proceedings - 2015 IEEE Winter Conference on Applications of Computer Vision, WACV 2015*. Institute of Electrical and Electronics Engineers Inc., feb 2015, pp. 1038–1045.

[23] R. Gadde, R. Marlet, and N. Paragios, "Learning Grammars for Architecture-Specific Facade Parsing," *International Journal of Computer Vision*, vol. 117, no. 3, 2015. [Online]. Available: https://hal.inria.fr/hal-01069379v2

[24] M. Mathias, A. Martinović, and L. Van Gool, "ATLAS: A Three-Layered Approach to Facade Parsing," *International Journal of Computer Vision*, vol. 118, no. 1, pp. 22–48, may 2016.

[25] R. Gadde, V. Jampani, R. Marlet, and P. V. Gehler, "Efficient 2D and 3D Facade Segmentation Using Auto-Context," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 5, pp. 1273–1280, may 2018. [Online]. Available: http://arxiv.org/abs/1606.06437

[26] A. Cohen, M. R. Oswald, Y. Liu, and M. Pollefeys, "Symmetry-Aware façade parsing with occlusions," in *Proceedings - 2017 International Conference on 3D Vision, 3DV 2017*. Institute of Electrical and Electronics Engineers Inc., may 2018, pp. 393–401.

[27] F. Korč and W. Förstner, "eTRIMS Image Database for Interpreting Images of Man-Made Scenes," University of Bonn, Tech. Rep., 2009. [Online]. Available: http://www.ipb.uni-bonn.de/projects/etrimsdb/

[28] O. Teboul, "Ecole Centrale Paris Facades Database," 2011. [Online]. Available: http://vision.mas.ecp.fr/Personnel/teboul/data.php

[29] R. Tyleček and R. Šára, "Spatial Pattern Templates for Recognition of Objects with Regular Structure," in *Proc. of German Conference on Pattern Recognition (GCPR)*, 2013, pp. 364–374.

[30] E. Yurtsever, J. Lambert, A. Carballo, and K. Takeda, "A survey of autonomous driving: Common practices and emerging technologies," *CoRR*, vol. abs/1906.05113, 2019. [Online]. Available: http://arxiv.org/abs/1906.05113

[31] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," *CoRR*, vol. abs/1604.01685, 2016. [Online]. Available: http://arxiv.org/abs/1604.01685

[32] G. Neuhold, T. Ollmann, S. R. Bulò, and P. Kontschieder, "The mapillary vistas dataset for semantic understanding of street scenes," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 5000–5009.

[33] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *CoRR*, vol. abs/1511.00561, 2015. [Online]. Available: http://arxiv.org/abs/1511.00561

[34] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," *CoRR*, vol. abs/1612.01105, 2016. [Online]. Available: http://arxiv.org/abs/1612.01105

[35] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[36] M. Lambers, "Survey of cube mapping methods in interactive computer graphics," *The Visual Computer*, pp. 1–9, 2019.

[37] T. Loga, B. Stein, and N. Diefenbach, "TABULA building typologies in 20 european countries—making energy-related features of residential building stocks comparable," *Energy and Buildings*, vol. 132, pp. 4–12, 2016, towards an energy efficient European housing stock: monitoring, mapping and modelling retrofitting processes. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0378778816305837

[38] ——, "TABULA database webtool," accessed: 2021-04-16. [Online]. Available: https://webtool.building-typology.eu/#bm

[39] V. Syrris, O. Pesek, and P. Soille, "Satimnet: Structured and harmonised training data for enhanced satellite imagery classification," *Remote Sensing*, vol. 12, no. 20, 2020. [Online]. Available: https://www.mdpi.com/2072-4292/12/20/3358

[40] R. Robinson, O. Oktay, W. Bai, V. V. Valindria, M. M. Sanghvi, N. Aung, J. M. Paiva, F. Zemrak, K. Fung, E. Lukaschuk, A. M. Lee, V. Carapella, Y. J. Kim, B. Kainz, S. K. Piechnik, S. Neubauer, S. E. Petersen, C. Page, D. Rueckert, and B. Glocker, "Real-time prediction of segmentation quality," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*, A. F. Frangi, J. A. Schnabel, C. Davatzikos, C. Alberola-López, and G. Fichtinger, Eds. Cham: Springer International Publishing, 2018, pp. 578–585.

[41] J. Henderson and J. Hart, "Bredem 2012–a technical description of the bre domestic energy model," *Building Research Establishment, UK*, 2012. [Online]. Available: https://www.bre.co.uk/page.jsp?id=3176#:~:text=The%20BRE%20Domestic%20Energy%20Model,work%2C%20such%20as%20stock%20modelling.

[42] M. Lanau, G. Liu, U. Kral, D. Wiedenhofer, E. Keijzer, C. Yu, and C. Ehler, "Taking Stock of Built Environment Stock Studies: Progress and Prospects," *Environmental Science and Technology*, vol. 53, no. 15, pp. 8499–8515, 2019.

[43] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," *CoRR*, vol. abs/1411.4038, 2014. [Online]. Available: http://arxiv.org/abs/1411.4038

[44] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 9351. Springer Verlag, may 2015, pp. 234–241.

[45] J. McGlinchy, B. Johnson, B. Muller, M. Joseph, and J. Diaz, "Application of unet fully convolutional neural network to impervious surface segmentation in urban environment from high resolution satellite imagery," in *IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium*, 2019, pp. 3915–3918.

[46] Z. Chu, T. Tian, R. Feng, and L. Wang, "Sea-land segmentation with res-unet and fully connected crf," in *IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium*, 2019, pp. 3840–3843.

[47] M. Havaei, A. Davy, D. Warde-Farley, A. Biard, A. Courville, Y. Bengio, C. Pal, P.-M. Jodoin, and H. Larochelle, "Brain tumor segmentation with deep neural networks," *Medical Image Analysis*, vol. 35, pp. 18–31, 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1361841516300330

[48] A. Kermi, I. Mahmoudi, and M. T. Khadir, "Deep convolutional neural networks using u-net for automatic brain tumor segmentation in multimodal mri volumes," in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, A. Crimi, S. Bakas, H. Kuijf, F. Keyvan, M. Reyes, and T. van Walsum, Eds. Cham: Springer International Publishing, 2019, pp. 37–48.

[49] P. Hu and D. Ramanan, "Finding tiny faces," *CoRR*, vol. abs/1612.04402, 2016. [Online]. Available: http://arxiv.org/abs/1612.04402

[50] L. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *CoRR*, vol. abs/1606.00915, 2016. [Online]. Available: http://arxiv.org/abs/1606.00915

[51] L. Zhou, C. Zhang, and M. Wu, "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 (CVPR) Workshops*, June 2018.

[52] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *CoRR*, vol. abs/1512.03385, 2015. [Online]. Available: http://arxiv.org/abs/1512.03385

[53] F. Milletari, N. Navab, and S. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," *CoRR*, vol. abs/1606.04797, 2016. [Online]. Available: http://arxiv.org/abs/1606.04797

[54] X. Li, X. Sun, Y. Meng, J. Liang, F. Wu, and J. Li, "Dice loss for data-imbalanced NLP tasks," *CoRR*, vol. abs/1911.02855, 2019. [Online]. Available: http://arxiv.org/abs/1911.02855

[55] A. V. Etten, "You only look twice: Rapid multi-scale object detection in satellite imagery," *CoRR*, vol. abs/1805.09512, 2018. [Online]. Available: http://arxiv.org/abs/1805.09512

[56] K. He, G. Gkioxari, P. Dollár, and R. B. Girshick, "Mask R-CNN," *CoRR*, vol. abs/1703.06870, 2017. [Online]. Available: http://arxiv.org/abs/1703.06870

[57] S. Ren, K. He, R. B. Girshick, and J. Sun, "Faster R-CNN: towards real-time object detection with region proposal networks," *CoRR*, vol. abs/1506.01497, 2015. [Online]. Available: http://arxiv.org/abs/1506.01497

[58] J. Huang, V. Rathod, C. Sun, M. Zhu, A. Korattikara, A. Fathi, I. Fischer, Z. Wojna, Y. Song, S. Guadarrama, and K. Murphy, "Speed/accuracy trade-offs for modern convolutional object detectors," *CoRR*, vol. abs/1611.10012, 2016. [Online]. Available: http://arxiv.org/abs/1611.10012

[59] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results," http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html, 2012.

[60] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[61] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," *CoRR*, vol. abs/1502.01852, 2015. [Online]. Available: http://arxiv.org/abs/1502.01852

[62] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng, "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015, software available from tensorflow.org. [Online]. Available: https://www.tensorflow.org/

[63] A. Fukami and K. Okamoto, "Total heat exchanger," Jul. 17 1984, uS Patent 4,460,388.

[64] L. Hoegner and U. Stilla, "Building facade object detection from terrestrial thermal infrared image sequences combining different views," *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. II-3/W4, pp. 55–62, 2015. [Online]. Available: https://www.isprs-ann-photogramm-remote-sens-spatial-inf-sci.net/II-3-W4/55/2015/

[65] R. Ilehag, A. Schenk, Y. Huang, and S. Hinz, "Klum: An urban vnir and swir spectral library consisting of building materials," *Remote Sensing*, vol. 11, no. 18, 2019. [Online]. Available: https://www.mdpi.com/2072-4292/11/18/2149

[66] F.-Y. Gong, Z.-C. Zeng, F. Zhang, X. Li, E. Ng, and L. K. Norford, "Mapping sky, tree, and building view factors of street canyons in a high-density urban environment," *Building and Environment*, vol. 134, pp. 155–167, 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0360132318301148

[67] Z. Ma and S. Liu, "A review of 3d reconstruction techniques in civil engineering and their applications," *Advanced Engineering Informatics*, vol. 37, pp. 163 – 174, 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1474034617304275

[68] O. Özyesil, V. Voroninski, R. Basri, and A. Singer, "A survey on structure from motion," *CoRR*, vol. abs/1701.08493, 2017. [Online]. Available: http://arxiv.org/abs/1701.08493

[69] R. Rudd, K. Craig, M. Ganley, and R. Hartless, "Building materials and propagation," *Final Report, Ofcom*, vol. 2604, 2014. [Online]. Available: https://www.ofcom.org.uk/research-and-data/technology/general/building-materials