Estimating Energy Consumption of Residential Buildings at Scale with Drive-by Image Capture

WOC Ward X Li Y Sun M Dai H Arbabi D Densley Tingley M Mayfield
Department of Civil and Structural Engineering, The University of Sheffield, UK

Abstract

Data driven approaches to addressing climate change are increasingly becoming a necessary solution to deal with the scope and scale of interventions required to reach net zero. In the UK, housing contributes to over 30% of the national energy consumption, and a massive rollout of retrofit is needed to meet government targets for net zero by 2050. This paper introduces a modular framework for quantifying building features using drive-by image capture and utilising them to estimate energy consumption. The framework is demonstrated on a case study of houses in a UK neighbourhood, showing that it can perform comparatively with gold standard datasets. The paper reflects on the modularity of the proposed framework, potential extensions and applications, and highlights the need for robust data collection in the pursuit of efficient, large-scale interventions.

1 Introduction

Buildings and their operation contribute to nearly 17% of global carbon emissions (UNEP, 2020). Of these emissions, 61% can be mitigated, according to the IPPC (2022), with the largest share of mitigation potential coming from the retrofit of existing buildings in developed countries such as the UK. The report also highlights that the next decade is critical for building technical capacity to ensure this potential is realised (IPPC, 2022). Technical solutions to the mass implementation of retrofit require robust, large scale data and modelling.

Large scale modelling of residential buildings with a resolution of information at the individual building level requires high volumes of data. Capturing and processing high quality data that can be used in decision making both reliably and efficiently, in high volumes, will require a substantial degree of automation. However, access to reliable sources of built environment data can be a challenge. Building stock models, for example, have been developed for use in modelling energy usage and occupant behaviour at an individual building level, however such methods have relied on a set of predefined archetypes (Shorrock et al., 2005). Such archetypes, that describe e.g. age cohorts, can miss particular nuances in different construction types, or building performance. In Great Britain (GB), datasets such as those provided by Ordnance Survey (Ordnance Survey, 2022b) and Verisk (Verisk, 2022) provide attributes for individual properties on a national scale, including building footprints, building heights and usage. Previous work has looked at this data, along with aerial point cloud data, to produce city-level stock models (Steadman et al., 2020),

however while the aerial data can provide large scale topographic information, there are limited resources to provide information on facades and other street-level features that are essential in aiding understanding of the urban environment, something that is essential for enacting retrofit measures.

Energy consumption per unit floor area is a metric used to assess the energy efficiency of a building. In the UK, this is most commonly reported in the form of energy performance certificates (EPC). The generation of EPCs requires a manual survey of the property, which is then used to input information into the so-called Standard Assessment Procedure (SAP) to estimate energy consumption (BRE, 2012). These assessments take approximately 45 minutes per building to complete and are conducted as required by law: most commonly when a property is sold, or every ten years in the case of rental properties (ReallyMoving.com, 2022). However, reports of issues in EPC reporting are widespread (Hardy and Glew, 2019). For example, the Retrofit Playbook, a guide to retrofit for policy makers in the UK published by the UK Green Buildings Council, describes EPCs as "not fit for purpose", and highlight this as a barrier to enacting home retrofit in the UK (UKGBC, 2021). To help overcome these barriers, a framework is presented in this paper that uses drive-by captured image data to generate energy models at a high scale.

The framework is designed as a set of modular components, defined in terms of their input and output, with the aim to simulate energy consumption for individual properties at scale. The framework is a data-driven approach that utilises computer vision techniques, including machine learning and 3-D reconstruction, to measure and assess buildings using street-level images and use this to build models to simulate energy consumption. Each component of the framework is discussed, and the approach applied to a case study of residential buildings in a neighbourhood in the UK. The capabilities of the proposed framework are compared against available data, and the results critically appraised in terms of performance against this existing data. The framework also utilises a modular approach that could be used in conjunction with other methods. Furthermore, the paper addresses the potential scope for extension and generalisation of the framework with additional modalities of data, such as thermal imaging, and the application to other quantification problems in built environment research.

2 Framework Design

This section presents a detailed overview of the proposed framework for capturing and localising street-level images and using them to estimate the physical properties of houses for simulation of energy consumption. Figure 1 provides an illustration of the framework as a pipeline of modular components mapping from data collection and localisation to energy modelling. Each framework component highlighted in Figure 1 is described in the following sections.

2.1 Data Collection and Localisation

A scalable platform for estimating building properties and energy consumption requires a scalable solution to data generation and processing. Data analysis and decision making that can be

performed at neighbourhood- or city-scale requires large quantities of high quality, localised data.

2.1.1 Capture

Drive-by data capture is not uncommon, and has been used for applications such as mapping (Anguelov et al., 2010) or in the development of self-driving vehicle technology (Gwak et al., 2019). To effectively perform feature and geometry extraction, image data needs to be available in high temporal and spatial resolution. In this research, a bespoke mobile sensing vehicle is used to prototype the proposed framework (Arbabi et al., 2021).

Image data is captured by driving the sensing vehicle along residential streets. The mobile sensing vehicle uses a multi-camera rig[1] to capture spherical image data using five radial cameras

[1]Teledyne FLIR Ladybug5+

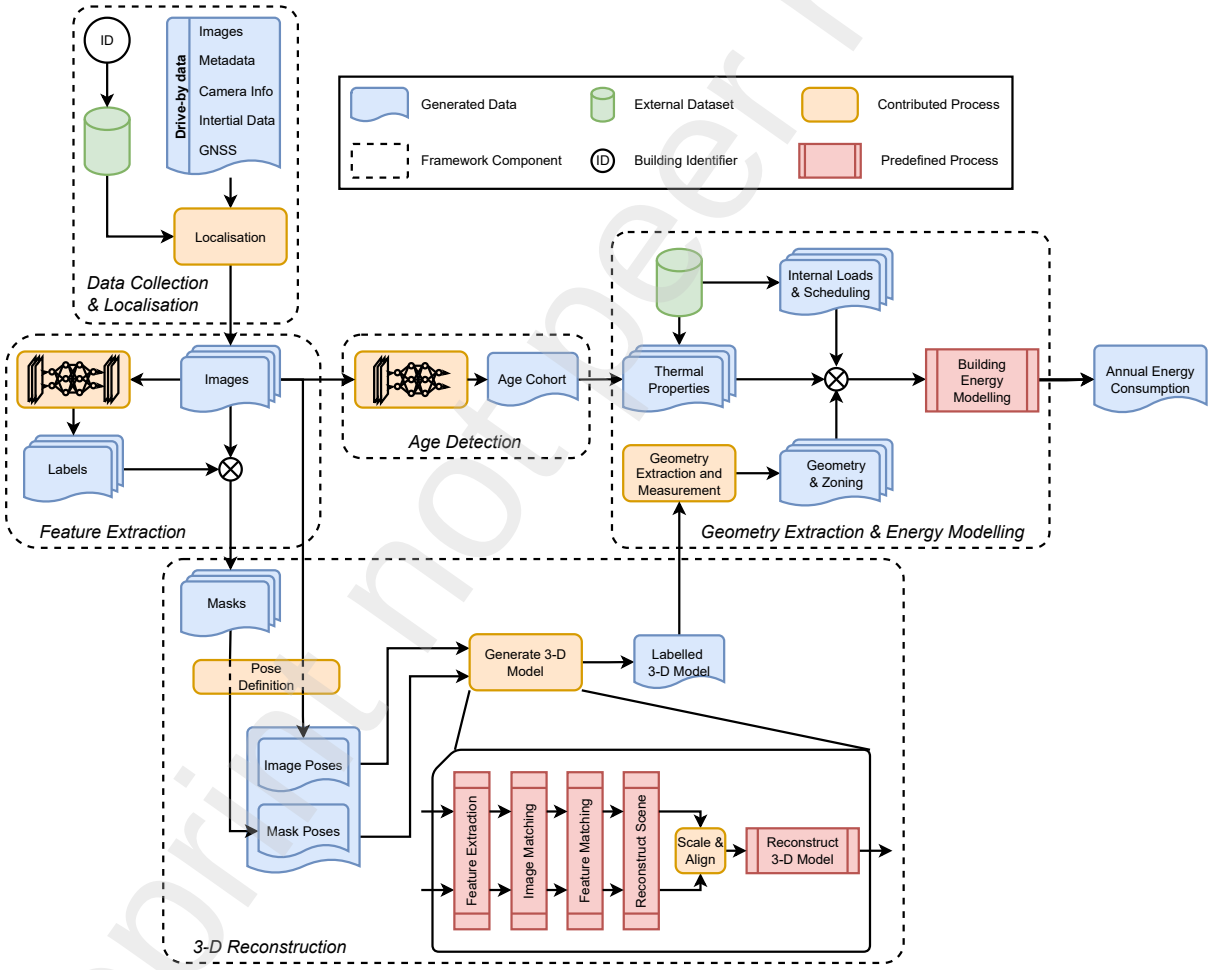


Figure 1: Overview of proposed energy consumption prediction framework, outlining individual components of the framework. Components of the framework are data collection and localisation (Section 2.1); age detection (Section 2.2); feature extraction (Section 2.3) and 3-D reconstruction (Section 2.4); and geometry extraction and energy modelling (Section 2.5). Aspects of the framework highlight where data is generated by the process, through either a process contributed by the authors or through predefined processes from publicly available software

and one upwards facing camera, each with a resolution of 5 megapixels (MP), capturing with a frequency of up to 30 frames per second (FPS). In practise, a trade-off is made between pixel resolution and capture frequency due to limitations of bandwidth in saving the images: uncompressed images comprise a huge amount of data so cannot be captured at very high frequencies. In this paper, higher resolution images are prioritised, with fewer high resolution images having been generally found to produce better quality results in the 3-D reconstruction component of the energy prediction framework; to this end, six 2048×2464 pixel images are captured at a rate of 10 FPS. Driving through a neighbourhood at approximately 4.5 m/s ($\approx$ 16 km/h) around 12 images are captured per metre driven. At a distance of 10 m from the van, each pixel corresponds to approximately 2.5 cm^2 of, e.g., building facade.

A contemporary analogue to the image data produced is Google Street View (Anguelov et al., 2010), which has been used in both urban data projects (Li et al., 2017) and wider socioeconomic research (Gebru et al., 2017; Nguyen et al., 2020). Google Street View data is made available through a paid-for API. However, images are only available at a maximum pixel resolution of 0.4 MP and are restricted in the available spatial resolution, with the API returning only the nearest image to a given location, which can limit details that can be extracted for a given property (Google, 2022).

2.1.2 Localisation

With the high volume of image data captured using a mobile sensing vehicle, a clear indexing scheme is required. Turning the images into a geospatial dataset requires robust localisation of the captured data, allowing images to be associated with a spatial pose which, in this paper, can be used to associate views with given houses. Onboard the sensing vehicle used, a georeferencing system[2] comprising an inertial measurement unit (IMU) and global navigation satellite system (GNSS) are used to monitor and map the location of the vehicle with an accuracy of up to 0.1 m. The IMU also provides the orientation of the vehicle, up to an accuracy of 0.1°. Localisation of the vehicle can be performed at a frequency of up to 100 Hz, equivalent to every hundredth of a second.

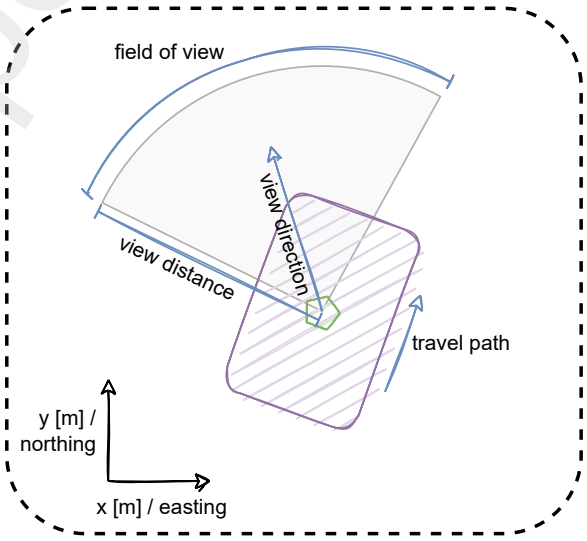


Figure 2: Sketch of a localised "view", associated with an image, generated relative to the position and travel path of the sensing vehicle

The onboard imaging system and IMU/GNSS share a time synchronisation system, which allows for reliable identification of the IMU/GNSS position for each image frame. With high frequency georeferencing of the van location, the position of each camera can be identified with linear

[2] OxTS Survey+

interpolation of the vehicle's position and orientation at capture time. At a driving speed of 4.5 m/s, there is an approximate accuracy assumption, for each camera frame's position, of 0.25 m. The output position is recorded in World Geodetic System (WGS 84), i.e. longitude and latitude. For use in the UK, the coordinates are reprojected into the Ordnance Survey National Grid reference system (OSGB 1936), which gives the van's position within the UK in metres, allowing for direct measurement of 3-D models generated later in the framework.

Localising each frame is essential for extracting views of a given house. Individual "views" are constructed to represent the perspective of a given image: the orientation and position of each camera relative to IMU/GNSS unit, combined with the post-processed measurement of the vehicle's location and orientation will give the centre point of the camera, in OSGB 1936, and the view direction of the camera. From this, a view is designed by creating a circular sector from the absolute position of the camera, with some predefined view distance and field-of-view. A sketch of this view is shown in Figure 2, highlighting one of the five radial views generated from image data. The upward facing camera is disregarded in further processing.

As input to the framework, some building identifier is required to indicate the property of focus. Two such identifiers commonly used for UK buildings are the unique property reference number (UPRN) and Ordnance Survey topographic identifier (TOID). There are widely available resources to link between these two references, as well as with other identifiers such as address (Ordnance Survey, 2022a,c). With these identifiers, existing geospatial information of the property such as its footprint can be extracted (Ordnance Survey, 2022b). Such information can be used to associate localised images from drive-by capture with individual properties, by finding intersections between the generated view and geospatial building information.

2.1.3 Generating perpendicular views

Due to the setup of the cameras, the images captured in the drive-by do not show perpendicular "face-on" views of properties. For the age detection component of the framework, these views are desirable as they provide a clear focal point for learning models. Such perspectives, however, can be generated by reconstructing all images in a single frame as a panoramic image showing the radial view of the sensing vehicle. Slicing this panorama can be used to create artificial views to show new perspectives, including views perpendicular to the vehicle. However, these views are not suitable for the reconstruction component, as they augment the images and remove contextual information about the camera.

2.2 Age Detection

To estimate thermal properties of a building, statistical estimates of thermal transmittance, in the form of u-values, are used due to the lack of comprehensive knowledge of thermal flux in exterior features, such as walls, windows and roofs. The TABULA project provides estimates of u-values based on age-based statistical archetypes for countries in the European Union, including for Great Britain in the UK (BRE, 2014; Loga et al., 2016). Predicting the age-cohort will allow the u-values from TABULA to be used to predict energy consumption.

2.2.1 Model

To classify the image data, a deep convolutional neural network (DCNN)-based model is used to estimate age cohort. Similar to the model presented in Zeppelzauer et al. (2018), the age detection model used in this paper relies on a patch-based classification and fusion approach, whereby the image is divided into subregions and each region is classified before an average pooling of the predicted age for each patch produces a single estimate.

The age detection model can be considered as three distinct elements: the feature selection, which extracts patches from the input image and stacks them into a single tensor; the backbone, a DCNN, that identifies and emphasises specific features within the patches; and a prediction layer, that distinguishes the features into classes, pooling classified patches to create a single unified prediction for the age cohort of the given image. Figure 3(a) shows an illustration of the model from input to prediction. The hyperparameters of the model are the number and size of patches, and the choice of backbone model. In the implementation of the framework, ResNet-18 is used as a backbone, a common DCNN used in classification problems, with widely available implementations (He et al., 2016; PyTorch, 2022).

Table 1: Alignment of categories of age cohort used in Verisk UKBuildings, used to train the age detection model; TABULA archetypes used to estimate u-values, and aggregation reported in energy performance certificates (EPC)

UKBuildings	**TABULA**	**EPC**
Historic	pre-1919	before 1990
Interwar	1919-1944	1900-1929
Postwar	1945-1964	1930-1949
Sixties Seventies	1965-1980	1950-1966
Modern	1981-1987	1967-1975
	1987-1990	1976-1982
	1991-2003	1983-1990
	2004-2009	1991-1995
		1996-2002
		2003-2006

2.2.2 Training and Validation

The age detection model was trained using a sample of 2463 images of houses in South Yorkshire, UK, captured perpendicular to the mobile sensing vehicle, using the panorama slicing algorithm described in Section 2.1.3. Building age cohorts were obtained using Verisk UKBuildings (Verisk, 2022), a geospatial dataset with a small number of attributes for residential properties in GB. The aggregation of ages into cohorts is shown in Table 1, showing how the cohorts align with the TABULA age categories, as well as the aggregation used in EPCs. While there is no perfect alignment, the reliable availability of building ages in UKBuildings dictated its use as a label set for training and validation.

The houses were randomly sampled from the capture data such that the number of houses in each age cohort was approximately equal: 487 houses were 'Historic'; 'Interwar' and 'Postwar' each comprised 496 houses; and 492 each of 'Sixties Seventies' and 'Modern'. Houses for training and validation were sampled from captured data in the South Yorkshire region, including neighbourhoods in Sheffield and Barnsley, but excluding Doncaster as this is used in the case study later.

The dataset was randomly subdivided into training, validation and testing sets, at a ratio of 80:10:10%, respectively. Training was performed initially for 50 epochs with early stopping using validation loss at 32 epochs to prevent overfitting. Each epoch involved an evaluation of the model, performed with a batch of $N = 8$ images before updating the parameters. Each image was subdivided into $P = 32$ random 32×32 patches extracted from the middle 50% of the image, and stacked and reshaped into an $(NP) \times 64 \times 64$ batch tensor before propagating through the backbone and classification layers. The output predictions are reshaped and averaged to produce a prediction for each class. The loss used cross entropy, and the Adam optimiser was used with a learning rate of 0.003. The weights of the backbone were initialised with pretrained weights for the classification of the ImageNet dataset to provide a well generalised starting point, but were not fixed (Kornblith et al., 2019; PyTorch, 2022). All other weights in the model were initialised randomly.

After 32 epochs of training, the validation accuracy was 71.2%, and the testing accuracy was 68.9%. The performance of the trained model is higher than that of the model proposed in Zeppelzauer et al. (2018), but this could be accounted for by difference in the number of classes, number of training, relative homogeneity of neighbourhoods in the training set and the different approach to patching the images.

2.3 Feature Extraction

2.3.1 Labelling Images

Understanding images at a pixel level provides a means to identify features of a building facade. In the framework, projecting labelled features into 3-D allows for the measurement and localisation of individual components of the building. The action of labelling, or segmenting images, is to assign each pixel to a set of semantic categories that inform the scene, essentially highlighting what is in an image and where it is located. In the proposed framework, the features of interest

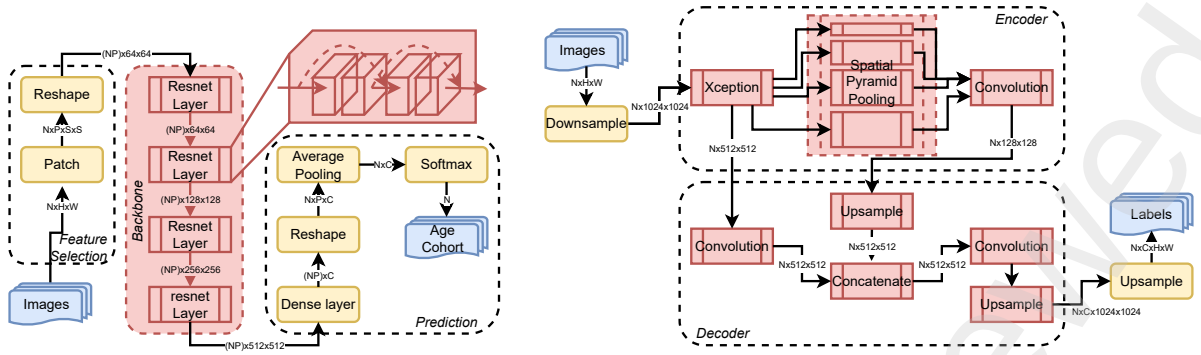


Figure 3: Illustrations of the two machine learning models used in the implementation of the proposed framework. (a) Age detection, using image patching to select features and pass through a backbone network, e.g. ResNet-18, and predict age cohorts based on average patch classification; (b) Feature extraction using semantic segmentation of images using the DeepLabv3+ architecture that encodes image features to different levels using a deep convolutional network, e.g. Xception, and spatial pyramid pooling, before decoding the features to pixel classes using convolution and concatenation of mixed-level features. Both illustrations indicate the dimensions of data throughout, where N is the number of $H \times W$ images processed, P is the number of $S \times S$ patches extracted, and C is the number of categories in the respective classification

are properties of the building facade and roof, namely windows, doors and chimneys, along with classification of the wall and roof. Semantic segmentation of building facades will return pixel-level labels of each category and treat anything else as "background".

Manual segmentation of images is possible, but to do so on the scale required in the framework would be prohibitively time-consuming. To this end, machine learning approaches are utilised, as with the age detection component of the framework. Semantic segmentation of building facades is a well studied topic, with dedicated models designed around extracting building features (Dai et al., 2021; Ma et al., 2022). In this paper, we use DeepLabv3+, an out-of-the-box DCNN-based encoder-decoder model that is used for a wide range of semantic segmentation problems (Chen et al., 2018). The decision to use DeepLabv3+ was predominantly due to easily accessible implementations (PyTorch, 2022). The relative simplicity of the segmentation problem, in that facade features are typically simple rectangular shapes, lends to the idea that a generalised semantic segmentation model will perform well.

The DeepLabv3+ model is built around an encoder-decoder architecture. The encoder part generalises so-called "high-level" and "low-level" abstract features using a DCNN backbone, most commonly Xception (Chollet, 2017), and a series of algorithms called spatial pyramid pooling that learn to generate a low-dimensional representation of the models. These features are fed into the decoder part of the model, which learns a transformation to map these features to pixel level for classification. Training the model on a set of manually labelled building facades adapts the model to take in street-level images of houses and return a pixel map of semantic labels. Figure 3(b) shows an illustration of the basic architecture of the semantic segmentation model.

The model was trained using a set of 6000 directly captured images from the mobile sensing vehicle, which were manually annotated to highlight windows, walls, roofs, doors and chimneys. A small

proportion of the images used contained no labelled features, to provide better generalisation of the trained model. The image was split 80:10:10% between training, validation and testing, and was trained for 100 epochs. No early stopping was employed, but epochs were capped at 100 due to limits in computational resources. The test accuracy of the model, i.e. the average percentage of pixels correctly classified, was 93.6%. A more discerning metric, the mean intersection over the union (IOU) across all labels, was recorded as 78.9%. The IOU quantifies the degree of overlap between predicted regions and true segmented regions, and is widely used in classification problems (Dai et al., 2021). The results for the trained model are in line with state-of-the-art semantic segmentation work, e.g. (Dai et al., 2021; Ma et al., 2022). Figure 4 shows the results of the trained model alongside the ground truth for an example image in the test set.

2.3.2 Masking Images

The trained segmentation model is used to automatically create label maps for facades to be used for projection and measurement in the framework. An additional benefit of these label maps is that they can be used to mask the original images to remove background features, which is beneficial during 3-D reconstruction, as the final model will only contain features belonging to a building, without additional objects like cars or other urban furniture such as trees and lampposts. Reconstructing only the building in the images reduces the amount of post-processing required to extract geometries from the 3-D model. An example of a masked image is shown in Figure 4.

2.4 3-D Reconstruction

Once a set of views of a building has been labelled and masked, the 3-D reconstruction component of the framework is performed. Using the known localised views, as described in Section 2.1.2, along with intrinsic properties of the cameras, such as focal distance and field-of-view, poses can

(a) (b) (c)

Figure 4: Demonstration of labelling and masking of a building facade in a drive-by captured image. (a) The raw image; (b) A label mask resulting from evaluation of the trained semantic segmentation model; (c) The masked image obtained by removing “background” features

be reconstructed to build 3-D models with real-world coordinates. Mapping image and label data onto these 3-D models also allows for specific facade features to be measured in real-world coordinate space for use in energy modelling.

2.4.1 Defining Poses

Generating poses from the data is required to codify the spatial information about each image. In the framework, for a given building, there is a geospatially located polygon representing its global position within GB in metres. Also localised are the position and 'view' of associated images that contain the given building, identified during the localisation step. To generate intermediary data used for 3-D reconstruction, positions are centred relative to the centroid of the polygon, by simply translating the global position of the images to be positioned relative to the polygon, allowing for easier measurement of the output reconstruction.

Additionally, intrinsic camera properties such as the focal length and lens distortion, as well as the orientation in 3-D space of the camera, are attributed to each image pose. This process is repeated for both the original images and the masked images to create two sets of reconstruction data.

2.4.2 Generating 3-D Building Model

Once a set of poses is generated from the labelled and masked images, the 3-D reconstruction component of the framework can be used to build geometric models of buildings. Using a combination of structure-from-motion and multi-view stereoscopy, the multiple perspectives of the building can be used to localise features and create a surface model in 3-D space (Griwodz et al., 2021). The first step in the generation of a 3-D model is the detection and extraction of the so-called "features" in each image. These features are identified using the scale-invariant feature transform (SIFT), a widely used algorithm that detects abstract descriptive properties in an image, based on various properties such as sudden changes in colour or shape. The invariance of the features allows them to be paired together regardless of any perceptive transformation or distortion they are affected by, e.g. rotation, translation or shearing (Lowe, 2004). SIFT features have been used across computer vision applications, including object recognition in video tracking and image stitching, as well as 3-D reconstruction (Peng et al., 2009). The ability of SIFT to provide a generalised representation of features in images allows for the pairing and matching of objects to create correspondences in the building facade and wider urban furniture in the localised images.

The list of features extracted from each image is used to pair images based on their relative poses. Typically, in a structure-from-motion pipeline, this process relies on finding common SIFT descriptors between images and assigning pairings based on matches. However, the localisation gives known poses which allows for simpler assumptions to be made on the association between images. The assumption made here is that each feature can only have one corresponding match, which reduces the computational requirements that impact processing time, but limits the effectiveness, especially on repetitive structures. Despite this, the process remains fairly robust. To improve matching, this process is repeated twice, once for the original images and once for the

Table 2: Information and parameters required for energy modelling

Geometry and Zoning	Coordinates of footprint
	Building height
	Window coordinate locations
	Number of storeys
Thermal Properties	Thermal transmittance of wall, floor, roof and window
	Solar heat gain coefficient of windows
	Outdoor air filtration
Metereological conditions	Weather data
Internal loads and scheduling	Intensities of occupancy, lighting and equipment
	Schedules of occupancy, lighting and equipment
	Building service systems

masked images. The original images contain many more features within an image that can be used to infer context in the scene, including objects such as trees and cars that increase the total number of matches to improve the quality of image and feature referencing. However, since the intended output of the reconstruction is just the building, applying the process to the masked images generates a secondary representation of the scene. An alignment of the mask-based scene with the full feature reconstruction, in conjunction with the predefined poses acts as a corrective transformation that minimises potential errors and inaccuracies in, e.g. the location of the van, and the lesser contextual information in the masked images.

With the aligned and matched features, a reconstruction of the 3-D model can be performed by creating a surface mesh by connecting features. The generated mesh is automatically post-processed to remove artefacts and reduce regions with a large number of nodes. Following this, a texture map is created, which essentially projects the images onto the reconstructed 3-D model. Applying the texturing process with both the masked images and the label maps provides two representations of the house: one with photographic detail, and the other with a semantic label localised in 3-D space. In the former case, this can be used for visualisation, and checking the quality of a reconstruction, while the projected labels allow for the extraction and measurement of geometry of the building.

2.5 Geometry Extraction and Energy Modelling

To scale up the (partial) automation of building energy models requires condensing and formatting of building properties needed to simulate consumption over a defined period of time. In this section, the individual considerations of the building model are described and processes to extract them from drive-by imaging, or otherwise, are detailed.

2.5.1 Geometry Extraction and Measurement

Building a process to automate the extraction of geometry used to build energy models requires some preparation of the 3-D models generated in the pipeline.

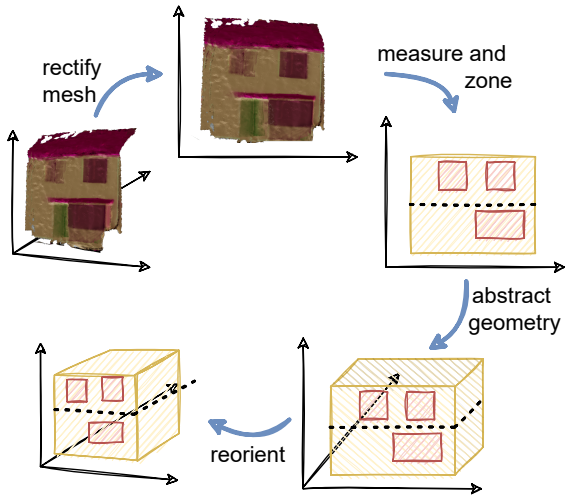


Figure 5: Illustration of pipeline for building geometric representation of building from a labelled reconstructed mesh, showing the process of rectifying a mesh to orient with the yz-plane; measuring the facade with bounding boxes and zoning based on storeys; abstracting the facade to a 3-D geometry; and transforming the building to its original orientation.

Due to the automated nature of the framework, the mesh generated during 3-D reconstruction may contain artefacts or low quality regions. A preprocessing step to remove low-quality features is performed by identifying natural clusters in the 3-D model: the DBSCAN algorithm identifies distinct spatial regions based on areas with dense detail (Ester et al., 1996); by virtue of the semantic segmentation and masking, largest distinct cluster is considered to represent the building, and the rest is discarded.

The orientation of the mesh is transformed to best align with the unit axes, such that the front facade aligns with the yz-plane. Bounding boxes are fitted to the features on the facade, such as the wall and each disconnected window, based on the projected labels. These bounding boxes are used to generate the measured geometry used in the building energy models. The geometry is separated into zones based on the number of storeys, which is inferred from the orientation of windows. The generated geometric representation is retransformed to the mesh's original orientation. Figure 5 illustrates the geometry extraction and measurement component of the framework.

2.5.2 Building Energy Modelling

The culmination of the building information extracted from drive-by images is to construct a model from which energy consumption, and therefore performance, can be effectively simulated. In this work, EnergyPlus, an industry-standard whole building energy simulation program, is used to estimate energy consumption given the inputs from the framework. These inputs are defined in an intermediary data format, generated from the information described in Table 2.

This approach allows for the modelling of each building independently, providing energy consumption information at a high level of granularity. The core aspects of the building that are identified as important for the simulation are the geometry and zoning; the thermal properties; and the internal loads and schedules. In the case of the former two, data obtained from drive-by captured data is used.

Geometry and Zoning Creating the physical representation of a building is, in essence, the extrusion of its footprint using the calculated geometry. Thermal zones can be inferred from the number of storeys, obtained by counting windows. The zones, representing a storey, are assumed to be of equal height, for simplicity in the generalisation of zones. The windows are represented as coordinated quadrilaterals on the facade, extracted from the minimum bounding rectangle for each window on the 3-D reconstructed mesh. For non-visible faces of the building, the window-to-wall ratio is instead encoded: equal to that of the measured facade on the opposite face; and a low number, e.g. 10%, for side faces. Symmetry of the window-to-wall ratio and a low non-zero number for side faces were used in the absence of directly observed data, as reasonably considered assumptions of the average construction of a residential building. A visual representation of the process is shown in Figure 5.

Thermal Properties To infer the thermal properties of the building, a set of age-based typologies for GB, developed by BRE, were used to infer the u-values of different features. The TABULA age-cohorts, which were used in the age detection component of the proposed framework, contain statistical assumptions of u-values for different properties, including walls and windows. The estimates given by the age detection model were used to generate estimates for the thermal transmittance and solar heat gain coefficient to be input in the energy model.

Internal Loads and Scheduling Due to the lack of observable information on the internal properties of each building, a uniform assumption was made for all simulations. To keep these assumptions as close to those used in EPCs as possible, the heating schedules were sourced from SAP 2012 guidelines: 9 hours on week days and 16 hours on weekends (BRE, 2012); schedules were also modelled to approximately represent reported diurnal patterns in energy usage (Few et al., 2022). Lighting and electrical equipment, and occupancy scheduling, were referenced from literature and industry guidelines (Baden-Powell et al., 2011; BRE, 2012).

2.5.3 Simulating Energy Consumption

In addition to the systems defined, associated weather data was obtained from International Weather for Energy Calculations (IWE, 2001). Based on the extracted geometry and generated model, energy consumption based on space heating, lighting and equipment is simulated using EnergyPlus giving an estimation of annual usage in kWh/m^2.

3 Related Work

Data driven solutions to categorising and quantifying the built environment, particularly energy consumption, are numerous and long-standing (Foucquier et al., 2013). Much of the research has focussed on understanding material stocks and predicting energy performance at large scales (Arbabi et al., 2021; Wei et al., 2018). To this end, building automated energy models from building data has been researched, for example in Chen et al. (2017) and Wate and Coors (2015). In Chen et al. (2017), the authors develop scenarios for retrofit at city-scale, using building data. 3-D data models have also been used to simulate energy usage (Wate and Coors, 2015).

Detection of facade features using machine learning and computer vision has become a popular topic in the last few years (Chen et al., 2017). Tailor-made facade segmentation solutions, such as in Liu et al. (2017) and Dai et al. (2021) report high accuracy but are limited in that they are applied predominantly to rectified images, i.e. those that have had lens distortion features removed, similar to the format used in the age detection component of the framework. Due to a lack of code availability, and specific requirements for the format of images, neither solution was used in this work. Other features that have been identified from street-level images include building age (Zeppelzauer et al., 2018), and heating energy demand (Despotovic et al., 2019).

Another component of the proposed framework relies on the projection of features to 3-D for the extraction of geometry. Identifying properties of buildings using existing sources such as Google Street View (Anguelov et al., 2010) has been applied to improving understanding of the urban environment (Hara et al., 2013; Campbell et al., 2019). Feature detection and mapping from Google Street View has been used to estimate building heights and improve facade understanding (Yuan and Cheriyadat, 2016). One of the main limitations with Google Street View data, however, is the spatial and temporal resolution at which it is available, meaning it can be difficult to reconstruct high quality 3-D geometries.

Aerial remote sensing is has also been used in urban quantification: datasets such as UKBuildings utilise LiDAR to estimate building heights (Verisk, 2022); Bayomi et al. (2021) use thermography to calibrate building envelopes; and remote sensing has been used to develop material stock models (Lanau et al., 2019). Stock models of buildings, such as in London, have been developed to build representations of cities, utilising data sources including aerial LiDAR (Steadman et al., 2020).

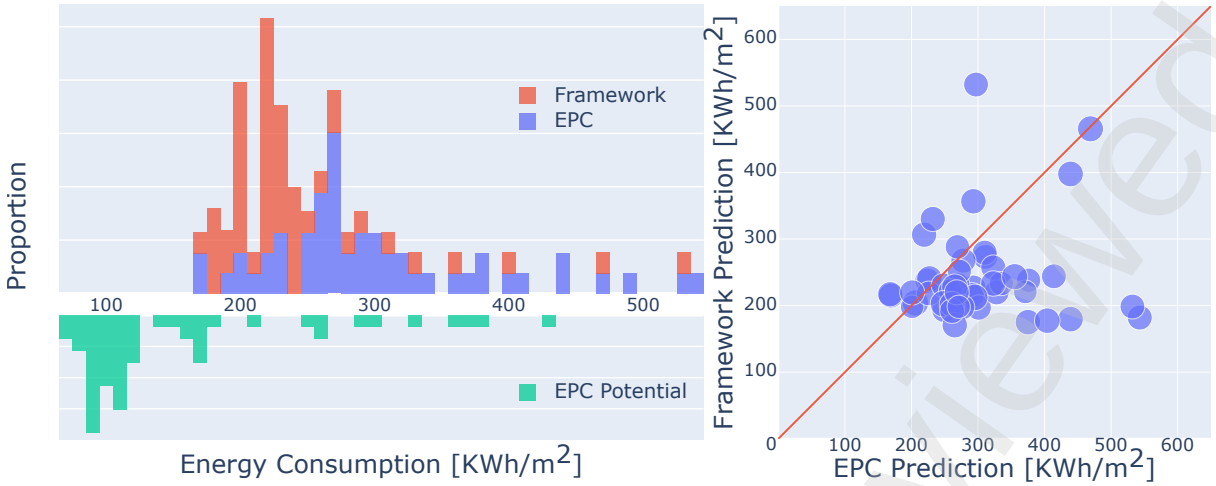
4 Case Study: Doncaster, UK

The framework as outlined was applied to houses in a residential neighbourhood in Doncaster, UK. Data from 53 council-owned social houses were selected for the case study, based on available information including addresses. Publicly available energy performance certificates (EPC) were obtained for each of the properties and filtered for the most up-to-date version. The framework was applied to each house and the estimated features of the building, including height and energy consumption, are compared against existing data sets.

4.1 Existing Data

The three main datasets used to compare and validate the outputs of the framework are OS MasterMap, Verisk UKBuildings and EPC reports. In the case of the former two datasets, properties such as the building footprint, building height and building age are available, with varying degrees of quality and confidence. In EPC reports, predictions of the energy consumption, alongside features used in the inputs of the SAP modelling are provided, including dwelling features such as heating systems, room space, and assumed insulation.

While polygons from OS MasterMap were used to localise the data, all other available properties are kept separate. In practice, one might combine these sources of available data to gain greater



(a) (b)

Figure 6: Comparison of predicted energy consumption by the MARVel-based framework and EPCs. (a) the distribution of predicted energy consumption from SAP-based EPCs and from the proposed framework. The lower plot shows the distribution of "potential" energy consumption based on recommendations provided in the EPC reports. (b) the energy consumption predictions from each source for each property in the case study.

representation of a building, but this is beyond the scope of the paper.

4.2 Validation

The predicted annual energy consumption in EPC reports is provided in kWh/m^2, which is used to generate ratings. Using the energy modelling component of the framework, annual energy consumption, in kWh/m^2, can be simulated. In these models, simulation takes less than a minute on a mid-tier laptop.

Figure 6 shows the predicted energy consumption by each method. In Figure 6(a), the distribution of energy consumption predictions over the sample houses is shown, highlighting the similarities in the overall values predicted across the houses estimated. The bottom plot indicates potential energy consumption provided in the EPC based on retrofit recommendations, to give a visual reference for the potential range of energy consumption values. In Figure 6(b), the energy consumption by each method for each building is shown, for direct comparison of estimation. From Figure 6(a), there is demonstrable agreement in the estimations for the majority of houses, with the framework-based estimation tending towards slightly lower estimations of energy consumption, on average. This is corroborated in the scatter plot, with most properties predicted as equal or slightly lower energy consumption with the framework. In some cases, however, there is a larger difference between the predictions, either with the EPC reporting higher values for energy consumption, or vice versa. In such cases, these can largely be attributed to the assumptions made both in the generation of the energy model using the drive-by data, which considers all internal systems and scheduling to be uniform across all houses; and in the assumptions made by the EPC provider. For example, for one property, the EPC-based energy consumption has been predicted significantly higher than the framework-based approach. Looking at the report features, the EPC highlights issues with poor efficiency from windows,

walls and the water heating system, the latter of which forms the most significant aspect of the recommendation to reduce energy consumption. On the other hand, the framework-based approach has assumed an average internal condition due to the lack of other data made available to it, and assumed thermal properties based on statistical archetypes of the building, characterised by its predicted age.

To validate the geometry extracted from the 3-D reconstructed models, the difference in calculated height with data available from a 3rd party dataset for each house is shown in Figure 7. The relative height-to-the-eaves, reported as "RelH2" in OS MasterMap Building Height Attribute (OS, 2022), is used as a benchmark, as the building heights are not reported in the EPC data. Figure 7 shows a general agreement between the two estimates of facade height, with a mean difference of 0.245 m.

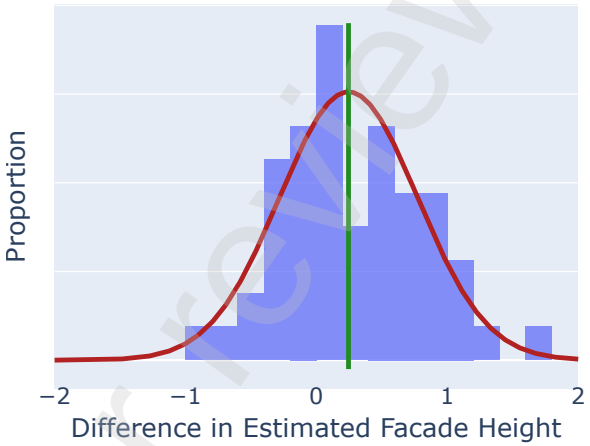


Figure 7: Histogram showing the proportion of differences between the facade predicted in the framework and the value reported in OS MasterMap for the buildings in the case study. When the difference is greater than zero, this indicates the framework height is greater than the OS value. The mean difference is 0.245 m highlighted by the vertical line. A Gaussian curve fitted to the mean and standard deviation is overlaid.

The case study conducted has demonstrated the feasibility of street-level drive by capture, in that it has the capability to provide estimates that are largely in agreement with those reported by EPCs, without the need for entry to the property for direct inspection. Factoring in the degree of assumptions also made in EPCs on aspects such as insulation thickness, the drive-by approach has viability, even with the approximations of interior state required.

5 Limitations, Modularity and Extensibility

The framework proposed and outlined in this paper is designed to perform large-scale generation of digital representations of buildings, with the end-goal of the framework to measure energy consumption. There are a number of limitations with each component of the framework, as in most data-driven approaches, including the requirements for large scale deployment and the limited availability of high quality ground truth data for validation.

These are discussed in turn in the following subsections. However, the modularity of the framework allows for clear adaptation and extension to both overcome some of the identified limitations and facilitates application to wider research problems in the urban environment.

5.1 Framework Design

Data Collection and Localisation A source of uncertainty is the localisation of the sensing vehicle, which can be located to within 0.1 m, but in practice may be less accurate. Furthermore, the difference in projection of location used between different sources of data, and the conversion between, is also a potential source of error or inaccuracy. While the GNSS/IMU unit in the van measures the global position in longitude and latitude, with units in degrees, building identifiers, such as OS TOIDs, as well as the measurement of structures, are made in metres. The conversion from longitude and latitude to easting and northing used by Ordnance Survey has an inaccuracy up to 1 m, and this conversion is only available in third-party GIS softwares.

The availability of other sources of data is limited, even for nationwide datasets. For example, UKBuildings only has building age for 71% of the houses in the neighbourhood from which the case study buildings were taken, and there is no clear way to validate the accuracy of these values.

It may also be possible to expand on the information provided from drive-by data collection by including additional data modalities, such as thermal and LiDAR. Greater insight and reliability of the structure of the buildings that these modalities might add will lead to more confident estimations of the energy consumption, as well as providing additional inferences, such as insulation thickness or fault detection by assessing thermal properties and more accurate geometric structure using the LiDAR point clouds.

Independent of uncertainty, another limitation is the sheer volume of data obtained by drive-by capture. In contrast to parameter based datasets, where each house is represented as a set of variables, typically text- or number-based, the drive-by process of capturing images and geolocation information creates a huge amount of data that needs to be processed and stored. While this paper outlines the framework as a proof-of-concept, the quantity of computational storage and other resources required need to be considered before deploying this type of image-based solution at a city, regional or national scale. In terms of raw image data, 1000 images takes up approximately 1 GB space. At a capture rate of 7.5 FPS, the total storage for a 75-minute drive is around 150 GB.

Feature Extraction Other sources of uncertainty in the framework include the trained machine learning models used to identify age and component features. In the latter case, most feature pixels are classified with an accuracy well over 95%; an exception to this are roofs, which are classified with accuracy 81% – the source of uncertainty here may be in the varied inclusion of eaves and gutters in the training data. For the age detection part of the framework, the model is only accurate approximately 70% of the time. While this is in line with state-of-the-art models for detection, it will likely have an impact on the results.

Despite the relatively high levels of accuracy in the models used in this work, they are not necessarily the best performing of all available models. As mentioned previously, bespoke facade detection methods might result in higher quality label data, and as research progresses, the state-of-the-art will improve. As the framework is designed to be modular in terms of defined

inputs and outputs, replacing components with new methods should be simple. However, with the addition of multiple modalities, such as thermal data, joint representation learning could further extend the capabilities of feature extraction: in the identification and localisation of facade features; or providing additional insight into the thermal properties of the building (Theodosiou et al., 2021).

Geometry and Zoning When creating the geometric representation of the building relies on a number of simplified assumptions. Thermal zoning is assumed by storey and, due to the lack of consistent reconstruction of roofs, all buildings are assumed to have a flat roof even if the contrary is true. As shown in the results, neither of these assumptions have drastic effects on estimated energy consumption, compared to the values reported in EPCs. Incorporating additional knowledge, for example structural archetypes, to better generalise the assumptions may yield more representative estimations of energy consumption.

The 3-D reconstruction component of the framework is based on multiview stereoscopy, but using different approaches to photogrammetry may yield more accurate measurements. LiDAR, for example, measures the 3-D scene directly, and aerial point clouds have been used for urban quantification (Steadman et al., 2020). More contemporary methods in 3-D scene representation include neural radiance fields (NeRF), which utilise generative machine learning models to generalise views and poses (Mildenhall et al., 2020). NeRFs have, for example, been used to represent cities at different spatial scales (Xiangli et al., 2022).

Thermal Properties Detection of materials from visual images alone is difficult due to issues such as paint and texture, so the framework uses an age detection model and statistical values for the thermal transmittance for use in the energy modelling component. However, combined information from data sources and drive-by data might be able to infer u-values with greater resolution. For example, whether a facade wall is solid brick, or a cavity wall, filled or unfilled, can be used to estimate u-values with greater degree of accuracy using industry documentation, such as reduced data SAP (rdSAP) (BRE, 2012). How to infer such properties with drive-by data is an open question that requires further research.

Internal Loads and Scheduling As highlighted in the evaluation of the case study, it is difficult to quantify the internal loads and scheduling used for energy modelling assumptions from drive-by capture. There are many studies which cover modelling occupant behaviour that can be used to provide "best guesses" (Carlucci et al., 2020; Yang et al., 2021). Focusing on this is beyond the scope of the paper, but introducing seasonality in the internal loads and scheduling, based on trends identified from qualitative research, or from statistical models based on smart meter data could introduce a more realistic model of occupant behaviour (Few et al., 2022).

5.2 Validation and Verification

On validation of the approach, there is limited data, especially on the interior of the property, that is consistently reliable and available. EPCs, UKBuildings and OS Mastermap all use similar

data driven approaches, or rely on flawed assessment as in the case of EPCs, and due to this there is no robust information for verifying data (Hardy and Glew, 2019).

Validation of the geometry measured with the framework is limited due to the availability of consistent and reliable data. In the case study, building height calculated in the framework was compared with OS data. However, for every house in the case study, the data is considered unverified by Ordnance Survey. In fact, this is the case for almost all houses in the area local to the case study. For example, of over $382,000$ houses analysed in South Yorkshire, UK, 99.65% were reported as unverified in the OS MasterMap Building Height Attribute dataset (OS, 2022).

5.3 Scalability

Consumer advice sites estimate that conducting an EPC survey can take between 45 and 60 minutes per house (ReallyMoving.com, 2022). With drive-by capture, up to 30 houses can be imaged per minute, based on the assumptions made in this paper. While the current implementation of the framework was built to model and simulate energy consumption for a single house at a time, it is possible to expand the scale of the framework by incorporating parallel computing, to reconstruct and simulate energy for multiple houses simultaneously. Similarly, additional efficiencies might be found in preprocessing of the image data. Captured data contains multiple houses and could be batch processed, including with labelling and masking.

5.4 Beyond Energy Consumption

While the focus in this paper has been on using drive-by capture data to model energy consumption, there are many other uses for the data and the processes discussed. Understanding material inventories and dimensions can help build up a picture of material stock (Lanau et al., 2019). An extension of the proposed framework, with statistical assumptions based on factors such as age and build factor, would be to build a database to help quantify material stock for a specific region (Arbabi et al., 2021).

Similar to understanding material intensity in a region, knowledge of the dimensions and retrofit needs of a building can yield solutions to facilitate efficient manufacturing for retrofit interventions, e.g. for mass-produced panelised systems (Orlowski, 2020). Quantifying housing stocks at a neighbourhood, town or city level would allow for efficient manufacturing, resulting in an economy of scale benefit (Andronie et al., 2021).

6 Conclusions

This paper has outlined a multi-aspect, modular framework spanning from the capture and localisation of drive-by image data to reliable, scalable energy consumption prediction of individual residential buildings. The resulting predictions produce similar estimates for energy consumption as to EPCs. Each component of the framework is discussed and critically appraised both in terms of its individual performance and its contribution to the wider output of the framework. Limitations and extensions to the framework are discussed and more general aspects of the process are explored.

The current implementation of the feature extraction and measurement aspect of the pipeline is designed to measure geometry and simulate energy consumption for a single property. This design allows for parallel execution when scaling up to neighbourhood- or city-level. While reliant on the same input data set of drive-by capture, each reconstruction is independent, computational requirements notwithstanding. As opposed to reconstructing and measuring whole streets, the highly parallelisable approach in our methodology is much better designed for future scaling.

In practice, a hybrid method is likely needed, one that takes into account data from multiple sources and scales, with a view to confidence in data, its quality and any uncertainty in the process. The modular aspect of the framework presented in this framework offers some facilitation for improvements as methods improve, and for extensions into new avenues.

Authorship Contribution Statement

Wil O C Ward: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Writing – Original Draft, Funding **acquisition. Xinyi Li:** Conceptualization, Methodology, Software. **Yushu Sun:** Data Curation, Software. **Menglin Dai:** Software. **Hadi Arbabi:** Conceptualization, Writing – Review & Editing, Project administration, Funding acquisition. **Danielle Densley Tingley:** Writing – Review & Editing, Supervision, Funding acquisition. **Martin Mayfield:** Supervision, Funding acquisition.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported by EPSRC Active Building Centre [EP/V012053/1]. WOCW, XL and HA were additionally supported by Towards Turing 2.0 under EPSRC [EP/W037211/1] and The Alan Turing Institute. Neither EPSRC nor The Alan Turing Institute had any involvement in study design; execution; or in the writing of this article.

References

International weather for energy calculations. *American Society of Heating, Refrigerating and Air-Conditioning Engineers, ASHRAE*, 2001.

M. Andronie, G. Lăzăroiu, M. Iatagan, I. Hurloiu, and I. Dijmărescu. Sustainable cyber-physical production systems in big data-driven smart urban economy: a systematic literature review. *Sustainability*, 13(2):751, 2021.

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*, 43:32–38, 6 2010. doi: 10.1109/MC.2010.170.

H. Arbabi, M. Lanau, X. Li, G. Meyers, M. Dai, M. Mayfield, and D. D. Tingley. A scalable data collection, characterization, and accounting framework for urban material stocks. *Journal of Industrial Ecology*, 2021. doi: 10.1111/JIEC.13198.

C. Baden-Powell, J. Hetreed, and A. Ross. *Architect's pocket book*. Oxford: Architectural P., Oxford, 4th ed. edition, 2011. ISBN 9780080969596.

N. Bayomi, S. Nagpal, T. Rakha, and J. E. Fernandez. Building envelope modeling calibration using aerial thermography. *Energy and Buildings*, 233:110648, 2021.

BRE. The Government's standard assessment procedure for energy rating of dwellings, 2012. URL `https://www.gov.uk/guidance/standard-assessment-procedure`.

BRE. Building typology brochure England, 2014.

A. Campbell, A. Both, and Q. C. Sun. Detecting and mapping traffic signs from google street view images using deep learning and gis. *Computers, Environment and Urban Systems*, 77: 101350, 2019.

S. Carlucci, M. De Simone, S. K. Firth, M. B. Kjærgaard, R. Markovic, M. S. Rahaman, M. K. Annaqeeb, S. Biandrate, A. Das, J. W. Dziedzic, et al. Modeling occupant behavior in buildings. *Building and Environment*, 174:106768, 2020.

L. C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 11211 LNCS:833–851, 9 2018. doi: 10.1007/978-3-030-01234-2_49. URL `https://link.springer.com/chapter/10.1007/978-3-030-01234-2_49`.

Y. Chen, T. Hong, and M. A. Piette. Automatic generation and simulation of urban building energy models based on city datasets for city-scale building retrofit analysis. *Applied Energy*, 205:323–335, 11 2017. doi: 10.1016/j.apenergy.2017.07.128. URL `https://linkinghub.elsevier.com/retrieve/pii/S0306261917310024`.

F. Chollet. Xception: Deep learning with depthwise separable convolutions. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017-January:1800–1807, 7 2017. doi: 10.1109/CVPR.2017.195.

M. Dai, W. O. Ward, G. Meyers, D. D. Tingley, and M. Mayfield. Residential building facade segmentation in the urban environment. *Building and Environment*, 199:107921, 7 2021. doi: 10.1016/J.BUILDENV.2021.107921.

M. Despotovic, D. Koch, S. Leiber, M. Doeller, M. Sakeena, and M. Zeppelzauer. Prediction and analysis of heating energy demand for detached houses by computer vision. *Energy and Buildings*, 193:29–35, 2019.

M. Ester, H.-P. Kriegel, J. Sander, X. Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd*, volume 96, pages 226–231, 1996.

J. Few, M. Pullinger, E. McKenna, S. Elam, E. Webborn, and T. Oreszczyn. Smart energy research lab: Energy use in GB domestic buildings 2021. *Smart Energy Research Lab (SERL) Statistical Reports*, 1, 2022.

A. Foucquier, S. Robert, F. Suard, L. Stéphan, and A. Jay. State of the art in building modelling and energy performances prediction: A review. *Renewable and Sustainable Energy Reviews*, 23:272–288, 2013.

T. Gebru, J. Krause, Y. Wang, D. Chen, J. Deng, E. L. Aiden, and L. Fei-Fei. Using deep learning and Google Street View to estimate the demographic makeup of neighborhoods across the united states. *Proceedings of the National Academy of Sciences*, 114(50):13108–13113, 2017.

Google. Street View Static API, 2022. URL `https://developers.google.com/maps/documentation/streetview`. [Online; accessed: 11-July-2022].

C. Griwodz, S. Gasparini, L. Calvet, P. Gurdjos, F. Castan, B. Maujean, G. D. Lillo, and Y. Lanthony. Alicevision meshroom: An open-source 3d reconstruction pipeline. ACM Press, 2021. doi: 10.1145/3458305.3478443.

J. Gwak, J. Jung, R. D. Oh, M. Park, M. A. K. Rakhimov, and J. Ahn. A review of intelligent self-driving vehicle software research. *KSII Transactions on Internet and Information Systems (TIIS)*, 13:5299–5320, 11 2019. doi: 10.3837/TIIS.2019.11.002.

K. Hara, V. Le, and J. Froehlich. Combining crowdsourcing and google street view to identify street-level accessibility problems. pages 631–640, 2013.

A. Hardy and D. Glew. An analysis of errors in the energy performance certificate database. *Energy policy*, 129:1168–1178, 2019.

K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

IPPC. Climate change 2022: Mitigation of climate change, 2022.

S. Kornblith, J. Shlens, and Q. V. Le. Do better imagenet models transfer better? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2661–2671, 2019.

M. Lanau, G. Liu, U. Kral, D. Wiedenhofer, E. Keijzer, C. Yu, and C. Ehlert. Taking stock of built environment stock studies: Progress and prospects. *Environmental Science & Technology*, 53(15):8499–8515, 2019. doi: 10.1021/acs.est.8b06652.

X. Li, C. Zhang, and W. Li. Building block level urban land-use information retrieval based on Google Street View images. *GIScience & Remote Sensing*, 54(6):819–835, 2017.

H. Liu, J. Zhang, J. Zhu, and S. C. Hoi. Deepfacade: A deep learning approach to facade

parsing. *IJCAI International Joint Conference on Artificial Intelligence*, 0:2301–2307, 2017. doi: 10.24963/IJCAI.2017/320.

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*, 132:4–12, 2016.

D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision 2004 60:2*, 60:91–110, 11 2004. doi: 10.1023/B:VISI.0000029664.99615.94. URL https://link.springer.com/article/10.1023/B:VISI.0000029664.99615.94.

W. Ma, S. Xu, W. Ma, and H. Zha. Multiview feature aggregation for facade parsing. *IEEE Geoscience and Remote Sensing Letters*, 19, 2022. doi: 10.1109/LGRS.2020.3035721.

B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng. NeRF: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020.

Q. C. Nguyen, Y. Huang, A. Kumar, H. Duan, J. M. Keralis, P. Dwivedi, H.-W. Meng, K. D. Brunisholz, J. Jay, M. Javanmardi, et al. Using 164 million Google Street View images to derive built environment predictors of COVID-19 cases. *International journal of environmental research and public health*, 17(17):6359, 2020.

Ordnance Survey. AddressBase, 2022a. URL https://www.ordnancesurvey.co.uk/business-government/products/addressbase. [Online; accessed: 12-July-2022].

Ordnance Survey. MasterMap topography layer, 2022b. URL https://www.ordnancesurvey.co.uk/business-government/products/mastermap-topography. [Online; accessed: 12-July-2022].

Ordnance Survey. Open Linked Identifiers, 2022c. URL https://www.ordnancesurvey.co.uk/business-government/products/open-linked-identifiers. [Online; accessed: 12-July-2022].

K. Orlowski. Automated manufacturing for timber-based panelised wall systems. *Automation in construction*, 109:102988, 2020.

OS. Mastermap building height attribute, 2022. URL https://www.ordnancesurvey.co.uk/business-government/products/mastermap-building.

K. Peng, X. Chen, D. Zhou, and Y. Liu. 3d reconstruction based on sift and harris feature points. *2009 IEEE International Conference on Robotics and Biomimetics, ROBIO 2009*, pages 960–964, 2009. doi: 10.1109/ROBIO.2009.5420735.

PyTorch. Models and pre-trained weights, 2022. URL https://pytorch.org/vision/stable/models.html. [Online; accessed: 13-July-2022].

ReallyMoving.com. How to get an EPC, 2022. URL https://www.reallymoving.com/energy-performance-certificates/guides/how-to-get-an-energy-performance-certificate-epc. [Online; accessed: 23-August-2022].

L. D. Shorrock, J. Henderson, and J. I. Utley. Reducing carbon emissions from the uk housing stock. 2005. URL www.bre.co.uk.

P. Steadman, S. Evans, R. Liddiard, D. Godoy-Shimizu, P. Ruyssevelt, and D. Humphrey. Building stock energy modelling in the uk: the 3dstock method and the london building stock model. *Buildings and Cities*, 1:100–119, 5 2020. doi: 10.5334/BC.52. URL http://journal-buildingscities.org/articles/10.5334/bc.52/.

T. Theodosiou, K. Tsikaloudaki, K. Kontoleon, and C. Giarma. Assessing the accuracy of predictive thermal bridge heat flow methodologies. *Renewable and Sustainable Energy Reviews*, 136:110437, 2021.

UKGBC. The Retrofit Playbook. Technical report, 2021.

UNEP. 2020 global status report for buildings and construction: Towards a zero-emission, efficient and resilient buildings and construction sector, 2020.

Verisk. UKBuildings, 2022. URL https://digimap.edina.ac.uk/verisk. [Online; accessed: 12-July-2022].

P. Wate and V. Coors. 3d data models for urban energy simulation. *Energy Procedia*, 78: 3372–3377, 2015.

Y. Wei, X. Zhang, Y. Shi, L. Xia, S. Pan, J. Wu, M. Han, and X. Zhao. A review of data-driven approaches for prediction and classification of building energy consumption. *Renewable and Sustainable Energy Reviews*, 82:1027–1047, 2018.

Y. Xiangli, L. Xu, X. Pan, N. Zhao, A. Rao, C. Theobalt, B. Dai, and D. Lin. Bungeenerf: Progressive neural radiance field for extreme multi-scale scene rendering. In *The European Conference on Computer Vision (ECCV)*, 2022.

B. Yang, F. Haghighat, B. C. Fung, and K. Panchabikesan. Season-based occupancy prediction in residential buildings using machine learning models. *e-Prime-Advances in Electrical Engineering, Electronics and Energy*, 1:100003, 2021.

J. Yuan and A. M. Cheriyadat. Combining maps and street level images for building height and facade estimation. pages 1–8, 2016.

M. Zeppelzauer, M. Despotovic, M. Sakeena, D. Koch, and M. Döller. Automatic prediction of building age from photographs. In *Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval*, pages 126–134, 2018.