Measuring the Cityscape: A Pipeline from Street-Level Capture to Urban Quantification

Measuring the Cityscape: A Pipeline from Street-Level Capture to Urban Quantification

W Ward[1]*, M Dai[1], H Arbabi[1], Y Sun[1], D Tingley[1] and M Mayfield[1]

[1]Department of Civil and Structural Engineering, The University of Sheffield

*w.ward@sheffield.ac.uk

Abstract. Any solution to achieving climate targets must be performed at scale. Data driven methods allow expert modelling to be emulated over a large scope. In the UK, there are nearly 30 million residential properties, contributing to over 30% of the national energy consumption. As part of the UK Government's requirement to meet net-zero emissions by 2050, retrofitting residential buildings forms a significant part of the national strategy. This work addresses the problem of identifying, characterising and quantifying urban features at scale. A pipeline incorporating photogrammetry, automatic labelling using machine learning, and 3-D geometry has been developed to automatically reconstruct and extract dimensional and spatial features of a building from street-level mobile sensing.

Keywords: building stock, 3-D modelling, street-level capture, computer vision

1. Introduction

Global solutions to achieving climate change targets will require large-scale action. Decision-making at a large scale needs high volumes of information. Capturing and processing such information requires a significant degree of automation, and the generation of high-quality data that can be used to inform decisions reliably and efficiently.

According to the UN environment programme, in 2020, residential properties contributed to 17% global emissions [1]. In the UK, there are nearly 30 million residential properties, contributing to over 30% of the national energy consumption [2]. As part of the UK Government's requirement to meet net-zero emissions by 2050, retrofitting residential buildings forms a significant part of the national strategy [3]. We calculate that, in the UK, on average two houses per minute must be retrofitted between now and 2050 to meet net-zero targets. The sheer scope of this undertaking means that efficient, large-scale solutions are needed: solutions that require high quantities of robust information [4].

Access to reliable sources of data can be a challenge for large scale decision making. For example, stock models for residential buildings have been developed for use in modelling energy usage and occupant behaviour at an individual building level, however such methods rely on a set of predefined archetypes [5]. Data sets from Ordnance Survey [6] and Verisk [7] provide attributes for individual properties on a national scale, including building footprints, height attributes and usage. Some research has looked at incorporating these sources of information, in addition to aerial LiDAR information, to generate 3-D stock models [8]. However, while aerial information can provide large scale topographic information, details of the facade that can aid understanding is lacking.

Reliable and scalable methods for extracting structural features from residential buildings can contribute to a wealth of applications, for a wide range of stakeholders. Building 3-D geometric models with representative dimensions can be used to expedite and automate energy modelling, potentially augmenting or improving the Standard Assessment Procedure (SAP) for energy performance certification [9]. Automating the generation of geometric models might be used as part of a larger framework automating the extraction of features for building scalable interventions for retrofit or manufacturing.

This work looks at the development of a scalable pipeline for the generation of three-dimensional representations of residential buildings with accurate geometric information. Using street-level drive-by data capture, the pipeline identifies buildings; individual components on those buildings' facades; and builds a 3-D geometric model with localised features that can be used to extract measurements such as facade dimensions, and window-to-wall ratios.

The rest of this paper outlines related work, and outlines the proposed methodology for measuring 3-D geometry of properties from drive-by capture. The pipeline is evaluated and compared with available data, and scalability, limitations and future directions for the pipeline are discussed.

2. Related Work

Data-driven solutions to categorising and quantifying the built environment are numerous and long-standing [10]. Much of the research has been around understanding material stock [4] and predicting energy performance [11] at large scales. To this end, building automated energy models from building data has been researched in [12] and [13]. The former looks at developing retrofit scenarios at city-scale using the building data [12]. In [13], the authors investigate building 3-D data models to simulate energy usage.

The proposed methodology relies on the projecting automatically identified features in drive-by images. Identifying properties using existing sources such as Google Street View [14] has been applied to building understanding of the urban environment [6, 7]. Feature detection and mapping from Google Street View images has been used to build estimate building heights and facade understanding [17]. The main limitation with Google Street View data, however, is the spatial and temporal resolution at which it is available means that it is not possible to use for reconstructing high quality 3-D geometries.

Detection of facade features using machine learning has become a popular topic in the last few years [8–11]. Tailor-made facade segmentation solutions such as in [19] and [20] 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. Due to lack of code availability, and specific requirements for the format of images, neither solution were used in this work. Other features that have previously been identified from street-level images using computer vision and machine learning techniques include building age [21] and heating energy demand [22].

Identifying building heights at scale using data from aerial remote sensing is also an active area of research. Using aerial photography, machine learning methods have been used to build 3-D models of buildings from a top-down perspective by predicting depth from single images [14, 15]. Aerial LiDAR has also been used to generate large scale datasets of building height attributes, including OS MasterMap [6] and the Verisk UKBuildings dataset [7]. The generation of these datasets generally relies on automated processes, and as such there can be significant uncertainty where this data has not been verified.

3. A Pipeline for Urban Quantification

The pipeline for developing a scalable means of quantifying building facades by drive-by capture is outlined in Figure 1. The two major components can be summarised as feature localisation and identification; and 3-D projection of those features. The projected features can be used to provide

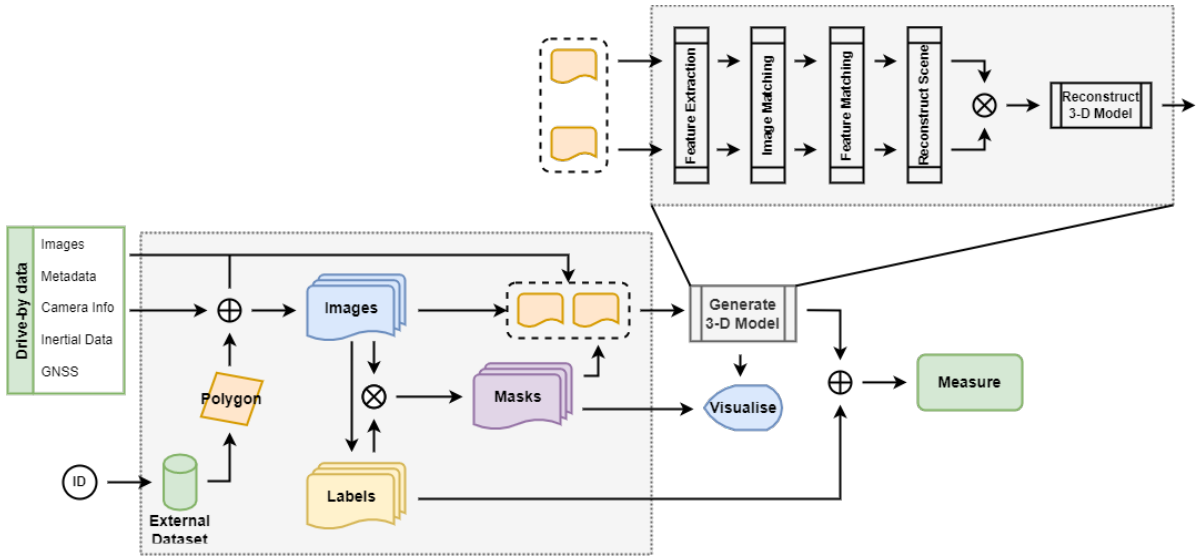

Figure 1. Overview of the pipeline for extracting quantitative information on individual residential buildings. The pipeline takes data from the drive-by capture, along with a building identifier to localise, extract and process the raw images of the building. The processed information is then used to generate a virtual 3-D representation that can then be visualised and measured.

measurements of a building. The remainder of this section outlines the key stages in the pipeline from capture, through to measuring the 3-D models.

3.1. Data Capture

One of the main requirements for building a process for data analysis and decision making is that it can be scaled to neighbourhood or city-level. Drive-by capture of data is not uncommon and is used for applications such as mapping [14] or developing self-driving vehicle technology [25]. In the context of measuring building facades, the collection of image data with both high temporal and high spatial resolution is essential. To this end, a bespoke mobile sensing vehicle is used to prototype the proposed pipeline [4].

Image data is captured using a FLIR Ladybug5+, a multi-sensor camera rig that captures spherical 360° image data, at a resolution of 30 MP, and frequency of up to 30 FPS. In practice, there is a trade-off between pixel resolution and capture frequency due to limitations of bandwidth in saving the image files: uncompressed image capture contains the greatest amount of spatial information and so cannot be captured at very high frequencies. For the projection and measurement of the processed data, precedence is placed on spatial accuracy, so images are captured at full, uncompressed resolution, resulting in six images per frame, at a rate of 10 frames per second. Given one of the sensors on the Ladybug5+ faces up, we disregard the images from this sensor. At a driving speed of 16 kph, this means we capture approximately 10 images per metre driven: two per sensor, each with a resolution of 2048 x 2464 pixels. At a distance of 10 m from the sensor, each pixel corresponds to approximately 2.5 cm2.

Matching images to specific locations requires each frame to have an accurate geospatial reference. The mobile sensing vehicle also houses an OxTS Survey+ inertial navigation system: a joint inertial measurement unit and global navigation satellite system (IMU/GNSS). The IMU/GNSS system provides positional information with an accuracy of up to 0.1 m, and provides information of the orientation of the vehicle, including its heading, with an accuracy of up to 0.1°. Vehicle localisation can be performed at a frequency of 100 Hz.

Frames from the Ladybug5+ and measurements from the IMU/GNSS are synchronised with an integrated time synchronisation server running on the sensing vehicle. We can thus reliably identify the IMU/GNSS position for each image frame. Given the high frequency of IMU/GNSS data, the position

of the vehicle is mapped to each frame using linear interpolation in time. At 16 kph, this means we have an approximate accuracy for each frame's position of 0.25 m.

The output position information is captured in World Geodetic System (WGS 84), which is then reprojected to the Ordnance Survey National Grid reference system (OSNG). The units of positions are now in metres, which easily allows for direct comparison when measuring projected information later in the pipeline.

3.2. Localising Views

The geolocation of each frame is essential for extracting views of a given residential property. The pipeline requires, as input, a building identifier that can be used to associate a given property. This identifier can be used to extract location information, such as a polygon from OS MasterMap [6]. Using this georeference, it is possible to extract views of the property from drive-by data.

Given a polygon, the individual image frames can be selected by disregarding all those outside a circular region with some pre-defined radius about the polygon, e.g. 20 m. Given the orientation of the cameras relative to the van's heading is known, and the heading is measured using the onboard IMU/GNSS, each image for each frame has some geospatial identity, in terms of its location and orientation. Images from frames within the predefined region about the property that are considered facing the property, i.e. the polygon exists within the field-of-view (FOV) of the camera at a given point, are selected.

With the subset of images, and known geospatial properties of the cameras, the reconstruction process can be initialised with known camera poses: intrinsic properties of the camera lens, such as FOV; the "centre" of the pose indicating its position in space; and a rotation matrix describing the orientation of the camera in 3-D space. This information will help identify features between images and reference them for scene reconstruction, as well as perform the reconstruction at real-world scale.

3.3. Feature Identification

To accurately measure a building using drive-by imaging and projection, we need to understand the input images at a pixel level. The action of labelling, or segmenting, images, is to assign pixels in the image to a set of semantic categories that inform the scene. Semantic segmentation essentially indicates what is in the image, and where in the image it is located. In the context of the proposed pipeline, the features of interest are those that are present on the building facade and roof: windows, doors, and chimneys; as well as classification of the wall and roof. Segmentation of building facades will return pixel-level labels of these categories and treat anything else as "background".

While it is possible to manually label images, this is a time-consuming activity. Manual segmentation would be a bottleneck in the scaling of the pipeline, requiring human intervention for every image. The

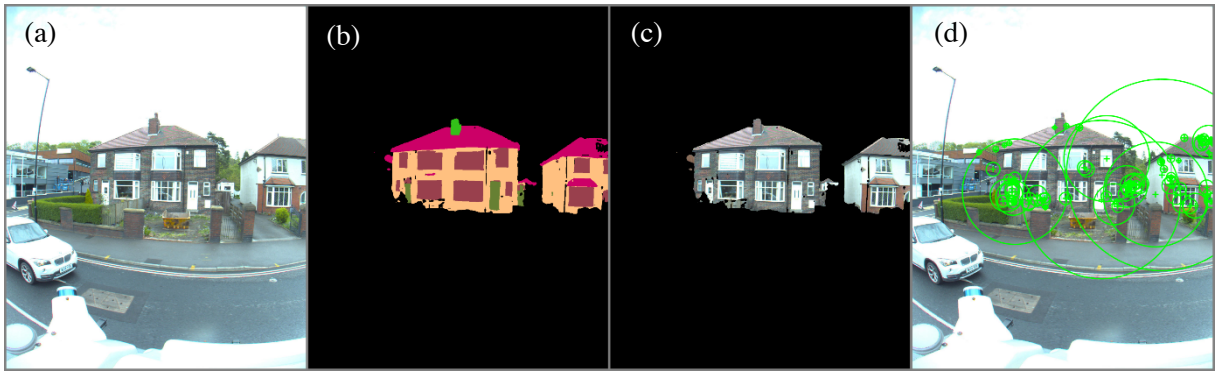

Figure 2. (a) An example image showing the view of a residential building, captured by the mobile sensing vehicle; (b) facade label information automatically generated by the trained machine learning model; (c) the view of the building masked by the label information; (d) a selection of SIFT features detected in the masked image, overlaid on the original view.

alternative is to utilise machine learning methods to automatically perform semantic segmentation of the building facades. Such an approach has been addressed in previous research, such as [19]. Building on their findings, a deep convolutional encoder-decoder model was trained. The architecture of the model is based on Deeplabv3+ [26] with an Xception decoder [27]. This model essentially builds a classifier by finding a low dimensional representation of images that accentuates features that can be easily classified, the encoder; and a pathway for these low dimension features to be mapped back to classified features back to the original image, the decoder. Training the model on a set of manually labelled building facades adapts this encoder-decoder architecture to take in street-level images and return a pixel map of labels.

The semantic segmentation model training setup used a set of 6000 manually labelled images, split 80:10:10% between training, validation and testing. The starting model parameters were initialised randomly, and the model was trained for 100 iterations. The test accuracy, i.e., the average percentage of pixels correctly classified, was 93.6%. The mean intersection over the union (IOU) across all labels, which indicates the degree of overlap between predicted and true segmented regions, was 78.9%. These metrics are in line with state-of-the-art (SOTA) segmentation methods, e.g., [19], [20], for both general purpose datasets and dedicated facade segmentation research.

The trained segmentation model is used to automatically create label maps for facades in the pipeline. These labels can be projected to 3-D to aid measurement, as discussed in the next section. An additional benefit of these label maps is that they can be used to mask the original images and remove background features. This will be beneficial during the 3-D reconstruction, as the final model will contain only features belonging to the building, without additional objects like cars, or other urban furniture such as trees and lampposts. Figure 2(a-c) shows an example of a labelled and masked image using the process outlined in this section.

3.4. Projecting the Scene

Once the set of views has been labelled and masked, the reconstruction aspect of the pipeline is used to generate 3-D models. Details of the image file paths, intrinsic camera properties, such as field-of-view, and pose information are collated into intermediary files used to initialise the 3-D model generation process.

The reconstruction process uses a combination structure-from-motion techniques, and multi-view stereoscopy [28]. From multiple perspectives of a single object, in this case a residential building, it is possible to localise features within 3-D space, and from this build a 3-D surface model of the object.

Reconstructing the scene, i.e., registering camera poses in 3-D space and matching features between images, is performed through a feature detection and matching pipeline. The first step is to identify "features" in the images: these are calculated using the scale-invariant feature transform (SIFT), an algorithm that detects descriptive properties in an image that can be paired together regardless of any perceptive transformation or distortion they are affected by, e.g., rotation, translation or shearing [29]. SIFT features are widely used for object recognition in applications such as video tracking and image stitching, as well as 3-D reconstruction [30]. Figure 2(d) illustrates the SIFT features detected from a drive-by captured image.

With the list of features extracted from each image, the next step is to pair images based on their relative poses. Typically, this process relies on finding common SIFT descriptors between images and assigning pairings based on matches. However, since we have initialised the process with known poses, it is straightforward as all images are chosen with views of the same objects. Each image is paired with all other images, and between these pairings, the SIFT features are compared. Features are matched between images where they contain the same descriptive information, independent of any distortion or transformation [29]. The algorithm for feature matching simplifies pair matching by assuming a feature can only have one corresponding match. Such an assumption has reduced capacity on structures that contain repetitive properties: a common occurrence in pictures of buildings due to brick patterns, for example. Despite this limitation, the process is fairly robust. To improve pairwise matching, we duplicate the feature detection and matching process on both the original images and the masked images.

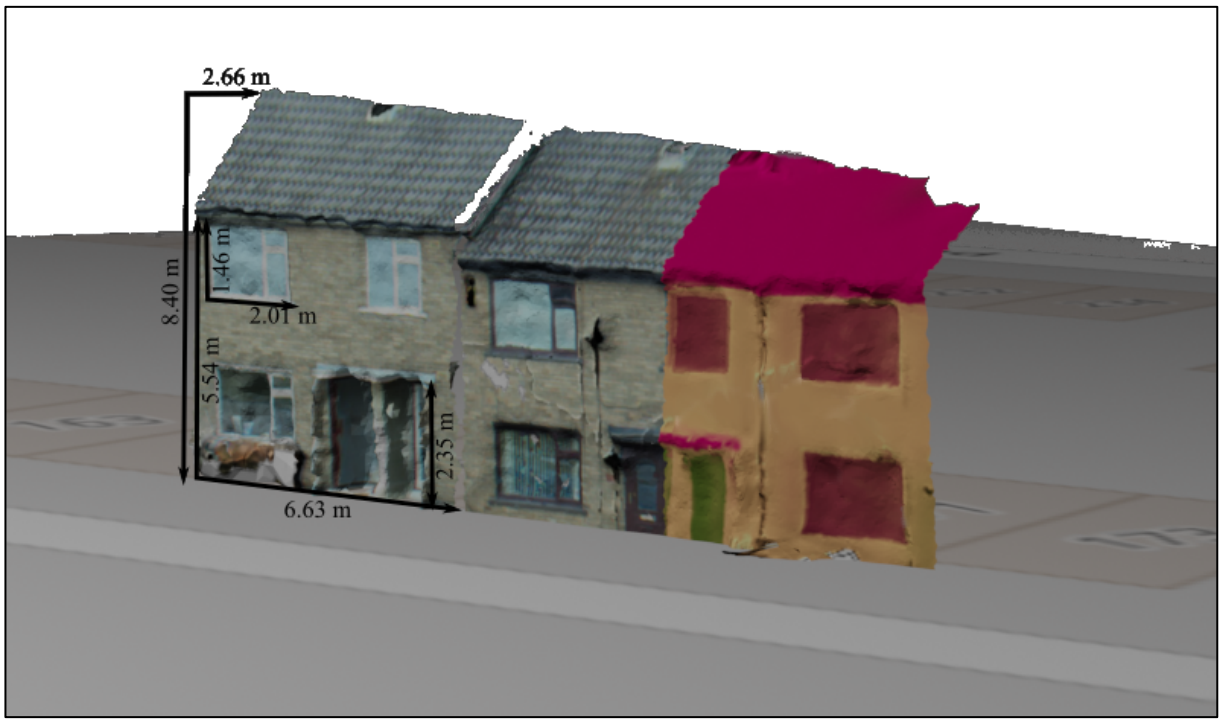

Figure 3. Examples of three reconstructed houses. Measurements overlaid on the left-most property show the dimensions of the facade, total structure, and a window and door. The central property shows the reconstructed facade with no annotation. The right-most property shows the results of labels projected onto the surface, which are used to measure the buildings.

In the latter case, we use masks to limit the amount of background that is reconstructed. However, the removal of extra information about the scene can limit the effectiveness of scene reconstruction. To mitigate this, applying feature matching on the original image will allow other urban furniture to be used to give spatial context to the scene and improve estimation of the 3-D projection. With the combined information of known poses, and reconstructed scenes from both original and masked images, the corrected camera poses and 3-D features can be used to build a surface map.

To build a 3-D surface map of the building, which will be output as a mesh of points, edges and faces, depth maps of each input are generated by comparing matching points with the properties of the cameras, such as focal distance and sensor sizing[31]. These depth maps, along with the reconstructed scene, can be used to create a mesh by triangulating nearby points [32]. The resulting 3-D model is constructed relative to the original camera poses, and thus its geometry corresponds to real-world sizing.

Images and label maps can be projected to the 3-D model using a process called texturing [33]. In the case of the former, this serves largely for visualisation, but in the case of the latter, the labels projected onto the 3-D surface can be used for feature localisation and measurement of the facade. Figure 3 shows an example of residential buildings reconstructed from drive-by capture.

3.5. Measuring the Cityscape

From drive-by capture to projection, the proposed pipeline creates a 3-D representation of residential buildings in real-world coordinates. Features of the facade are mapped into 3-D space. Measuring such features on the projection can give estimates of the structure, to be used for creating an understanding of the geometric properties of the building. Such properties include building height, facade height, window-to-wall ratio, and roof pitch; these can be used to generate building energy models, or to inform stock models.

4. Evaluation

The proposed pipeline was applied to a set of buildings using images obtained from drive-by sensing. Polygons were obtained using topographic identifiers (TOID) in OS MasterMap [6]. These polygons were used to isolate views of the corresponding properties and execute the reconstruction and localisation process. The reconstruction process is implemented using the Python bindings for AliceVision Meshroom [28]. Label information from the image segmentation is projected onto the resulting 3-D surface maps.

In Figure 3, three examples of reconstructions are shown along a street in Sheffield, UK. The original images are applied as textures in the first two properties, and the right-most property has pixel labels projected on. Extracted dimensions are also overlaid on one of the properties, showing the dimensions of the facade, as well as extracted geometry of windows and doors. In addition, the total height of the building and depth of the reconstruction are shown. Comparing these results with the properties from the Building Height attribute for the corresponding TOID [34], the total height of the building is measured at 8.40 m, with 8.8 m (reported only to 1 d.p.) in OS MasterMap. The height-to-the-eaves is reported as 4.9 m, versus the 5.54 m measured from the 3-D reconstruction. The discrepancies in values here may be attributed to limitations in the respective measurement methods and the uncertainties that they introduce. In the case of extracting total building height, it is likely that the roof is not fully reconstructed given limitations of the view from the street – a combination of camera FOV and slope of the roof mean the apex is not captured, especially at close range. According to the TOID polygons, the total depth of the building is 8.33 m by 6.78 m, meaning that width-ways the reconstruction is close, but we have only captured approximately two-thirds of the roof from eave to apex. However, it is possible to infer the roof pitch as 42.9°.

5. Discussion

The pipeline outlined in this paper is designed to build spatially accurate models using visual and spatial information that can be measured. Whereas with many existing datasets that infer properties of buildings using aerial surveys, the proposed pipelines are generated from street-level surveys. This gives an alternative perspective to inform quantitative methods and decision making. This section discusses the potential scalability of the proposed pipeline and outlines where street-level modelling can mitigate limitations in aerial surveying and highlight new sources of uncertainty. Possible future research expanding and utilising the pipeline is discussed and, finally, conclusions are drawn summarising the work.

5.1. Scalability

The current implementation of the pipeline is designed to get geometry on a single property. The design allows for parallel execution when seeking to build 3-D geometry for individual properties. While reliant on the same input data set of drive-by capture, each reconstruction is independent, computational requirements notwithstanding.

The computational time for the reconstruction process grows quadratically with the number of views, e.g., for 5 times as many images, the computational time will grow 25-fold. This was part of the motivation for building a single-house reconstruction pipeline, versus generating a 3-d model of a whole street or neighbourhood and analysing this, similar to aerial LiDAR-based approaches. In Figure 4, the relative computation time for three scenarios is illustrated: the full

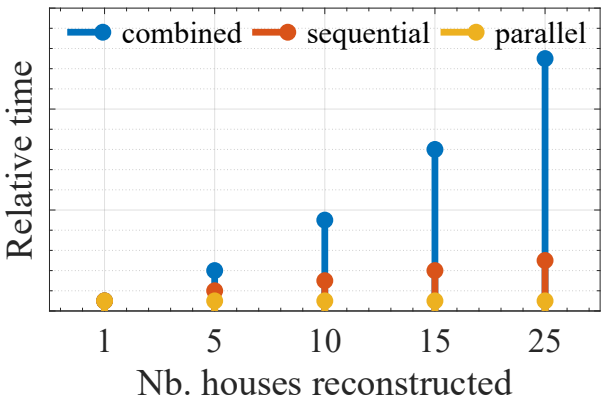

Figure 4. Relative time to reconstruct houses based on a combined reconstruction of all houses; using the proposed pipeline sequentially; and using the proposed pipeline in parallel.

reconstruction is shown to have high relative runtime; and the outlined single-house reconstruction executed both in sequence and in parallel. The figure demonstrates the scalability of the current approach, and the further benefits that might be gained by parallelisation.

Additional speed-ups can be found through preparation of the input data during the reconstruction of multiple households. In the drive-by dataset, a given image may contain views of multiple houses: this can be seen in Figure 2. Pre-labelling and masking the dataset would avoid some duplication of computation in the preparation of data. Similar gains may be made by pre-calculating SIFT features on the images.

5.2. Uncertainty and Limitations

As discussed throughout the paper, there are some limitations to the proposed methodology. The first quantifiable source of uncertainty is in the localisation of the sensing vehicle, which can only reliably be located to within 0.1 m, but may be less accurate, in practise. This inaccuracy is mitigated somewhat by reducing constraints in the scene reconstruction process, where only the overall scale and orientation are maintained. Similarly, the segmentation of facade features will always contain some error, despite having SOTA accuracy: for example, occluding objects such as trees will reduce the quality of segmentation. Notably, the metrics for the trained models are given as average over all labels. However, metrics for individual classes indicate that most features are classified with accuracy well over 95%. The exceptions are the roof features, which are classified by the model with pixel accuracy of 81%.

Robust validation of the work is also currently a limitation. Building measurements for height were compared against OS MasterMap building attributes. As mentioned, there are limitations with reliably extracting the total building height, but the facade height, i.e., height to the eaves, can be compared. However, the availability of verified data in Sheffield is limited: of over 382,000 TOID polygons available in OS MasterMap, the 99.65% of the building height attributes have a confidence of "99", meaning they are unverified [34]. The remaining polygons comprise buildings "for which [OS] have not been able to calculate some or all" attributes. Full validation of the findings in this work, and of the 3rd party datasets, such as OS MasterMap and Verisk UKBuildings, will require manual survey.

Independent of uncertainty, another limitation is the sheer volume of data obtained by drive-by capture. In terms of raw image data, 1000 images is approximately ~1 GB, which given each frame comprises six images, means the Ladybug5+ can capture 1 GB of data every six seconds. Even at a reduced framerate of 7.5 FPS, used in the prototype, the total storage required for a 75-minute drive is 150 GB. Any solution applying the pipeline at scale would need to consider the storage requirements, as well as the computational requirements.

5.3. Future Work

For large-scale reconstruction, it may be beneficial to pre-process much of the data independently of individual buildings. As discussed in section 5.1. , localising and labelling views in preparation would minimise repeat computations where views contain multiple properties. The preparation of such a dataset may also have benefits for other research that relies on high resolution, localised views of properties, such as [15] and [16].

Augmenting additional sensing information, for example thermography, is one possible avenue of research. Projecting localised temperature data, in a similar fashion to the feature labels that have been projected in the proposed pipeline, would give 3-D thermal information that could be used for understanding material properties of the facade, or for fault detection, e.g., in double-glazing windows.

Given the respective limitations of both the street-level sensing used to generate the 3-D models, and aerial sensing used to calculate building heights in OS MasterMap and Verisk UKBuildings, there may be some merit in combining both sources of data to get a greater perspective of properties: combining facade information that is difficult to obtain from aerial capture with the top-down information not captured by the mobile sensing vehicle.

The scope of this paper addresses the pipeline that localises and reconstructs 3-D geometry, and measurement of the resulting projected features is done manually. However, with spatially localised

features it may be possible to automatically extract geometry for use in scalable modelling: a drive-by capture that is then automatically processed, reconstructed and measured would provide an invaluable data source for stock modelling, for example, providing facade information that is not available from datasets generated from aerial remote sensing. Automatically generated data may also be used to build up large-scale digital twins.

5.4. Conclusions

In this paper, we have outlined a scalable process for building 3-D reconstructions of residential buildings. We outline the data capture and modelling pipeline, and demonstrated the efficacy and scalability of extracting accurate building features. We discussed the usability of large-scale quantification to energy and stock modelling, and have described how these can be used to build urban digital twins to model retrofit interventions. Future work will involve manually validating the results given the unavailability of verified data at scale.

The outputs for this work may be incorporated into building stock models, or used as inputs for energy modelling – both key requirements in the modelling of retrofit solutions. The raw data may also be combined with other sensing data, including thermography and LiDAR, both from street-level and aerial sources.

Acknowledgements

W Ward and Y Sun were supported by EPSRC Grant EP/V012053/1. W Ward was also supported by Towards Turing 2.0 under the EPSRC Grant EP/W037211/1 and The Alan Turing Institute.

References

[1] United Nations Environment Programme, "2020 Global Status Report for Buildings and Construction: Towards a Zero-Emission, Efficient and Resilient Buildings and Construction Sector." United Nations Environment Programme Nairobi, Kenya, 2020.
[2] E. & I. S. Department for Business, "Energy Consumption in the UK (ECUK)," UK.
[3] HM Government, "Net Zero Strategy: Build Back Greener," 2021.
[4] H. Arbabi *et al.*, "A scalable data collection, characterization, and accounting framework for urban material stocks," *Journal of Industrial Ecology*, 2021.
[5] L. D. Shorrock, J. Henderson, and J. I. Utley, "Reducing carbon emissions from the UK housing stock," 2005, Accessed: Feb. 15, 2022. [Online]. Available: www.bre.co.uk
[6] OS, "MasterMap topography layer," *Ordnance Survey*. https://www.ordnancesurvey.co.uk/business-government/products/mastermap-topography (accessed Feb. 15, 2022).
[7] Geomni, "UKBuildings." https://digimap.edina.ac.uk/geomni (accessed Feb. 15, 2022).
[8] 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*, vol. 1, no. 1, pp. 100–119, May 2020.
[9] "Standard Assessment Procedure - GOV.UK." https://www.gov.uk/guidance/standard-assessment-procedure (accessed May 30, 2022).
[10] 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*, vol. 23, pp. 272–288, 2013.
[11] Y. Wei *et al.*, "A review of data-driven approaches for prediction and classification of building energy consumption," *Renewable and Sustainable Energy Reviews*, vol. 82, pp. 1027–1047, 2018.
[12] 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*, vol. 205, pp. 323–335, Nov. 2017.

[13] P. Wate and V. Coors, "3D data models for urban energy simulation," *Energy Procedia*, vol. 78, pp. 3372–3377, 2015.
[14] D. Anguelov *et al.*, "Google street view: Capturing the world at street level," *Computer (Long Beach Calif)*, vol. 43, no. 6, pp. 32–38, Jun. 2010.
[15] K. Hara, V. Le, and J. Froehlich, "Combining crowdsourcing and google street view to identify street-level accessibility problems," in *Proceedings of the SIGCHI conference on human factors in computing systems*, 2013, pp. 631–640.
[16] 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*, vol. 77, p. 101350, 2019.
[17] J. Yuan and A. M. Cheriyadat, "Combining maps and street level images for building height and facade estimation," in *Proceedings of the 2nd ACM SIGSPATIAL Workshop on Smart Cities and Urban Analytics*, 2016, pp. 1–8.
[18] H. Liu, J. Zhang, J. Zhu, and S. C. H. Hoi, "Deepfacade: A deep learning approach to facade parsing," *IJCAI International Joint Conference on Artificial Intelligence*, vol. 0, pp. 2301–2307, 2017.
[19] M. Dai, W. O. C. Ward, G. Meyers, D. Densley Tingley, and M. Mayfield, "Residential building facade segmentation in the urban environment," *Building and Environment*, vol. 199, p. 107921, Jul. 2021.
[20] W. Ma, S. Xu, W. Ma, and H. Zha, "Multiview Feature Aggregation for Facade Parsing," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, 2022.
[21] 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*, 2018, pp. 126–134.
[22] 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*, vol. 193, pp. 29–35, 2019.
[23] C.-J. Liu, V. A. Krylov, P. Kane, G. Kavanagh, and R. Dahyot, "IM2ELEVATION: Building Height Estimation from Single-View Aerial Imagery," *Remote Sensing*, vol. 12, no. 17, p. 2719, 2020.
[24] Y. Xie *et al.*, "Multi-Scene Building Height Estimation Method Based on Shadow in High Resolution Imagery," *Remote Sensing*, vol. 13, no. 15, p. 2862, 2021.
[25] 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)*, vol. 13, no. 11, pp. 5299–5320, Nov. 2019.
[26] 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)*, vol. 11211 LNCS, pp. 833–851, Sep. 2018.
[27] F. Chollet, "Xception: Deep Learning with Depthwise Separable Convolutions," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 2017-January, pp. 1800–1807, Jul. 2017.
[28] C. Griwodz *et al.*, "AliceVision Meshroom: An open-source 3D reconstruction pipeline," 2021.
[29] D. G. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," *International Journal of Computer Vision 2004 60:2*, vol. 60, no. 2, pp. 91–110, Nov. 2004.
[30] 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*, pp. 960–964, 2009.
[31] C. Strecha, R. Fransens, and L. van Gool, "Combined depth and outlier estimation in multi-view stereo," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2, pp. 2394–2401, 2006.

[32] M. Jancosek and T. Pajdla, "Exploiting Visibility Information in Surface Reconstruction to Preserve Weakly Supported Surfaces," *International Scholarly Research Notices*, vol. 2014, pp. 1–20, Aug. 2014.
[33] C. Allène, J. P. Pons, and R. Keriven, "Seamless Image-Based Texture Atlases using Multi-band Blending," *Proceedings - International Conference on Pattern Recognition*, 2008.
[34] OS, "MasterMap Building Height Attribute," *Ordnance Survey*. https://www.ordnancesurvey.co.uk/business-government/products/mastermap-building (accessed Feb. 15, 2022).