
Proceedings Paper:
Dai, M., Meyers, G.M. orcid.org/0000-0003-4157-3991, Densley Tingley, D.O. orcid.org/0000-0002-2477-7629 et al. (1 more author) (2019) Initial investigations into using an ensemble of deep neural networks for building façade image semantic segmentation. In: Erbertseder, T., Chrysoulakis, N., Zhang, Y. and Baier, F., (eds.) Proceedings of SPIE. Remote Sensing Technologies and Applications in Urban Environments IV, 09-12 Sep 2019, Strasbourg, France. Society of Photo-optical Instrumentation Engineers . ISBN 9781510630178

https://doi.org/10.1117/12.2532828

Initial investigations into using an ensemble of deep neural networks for building façade image semantic segmentation

Menglin Dai, Gregory Meyers, Danielle Densley Tingley, Martin Mayfield

Menglin Dai, Gregory Meyers, Danielle Densley Tingley, Martin Mayfield, "Initial investigations into using an ensemble of deep neural networks for building façade image semantic segmentation," Proc. SPIE 11157, Remote Sensing Technologies and Applications in Urban Environments IV, 1115708 (2 October 2019); doi: 10.1117/12.2532828

SPIE.

Event: SPIE Remote Sensing, 2019, Strasbourg, France

Initial investigations into using an ensemble of deep neural networks for building façade image semantic segmentation

Menglin Dai*[a], Gregory Meyers[a], Danielle Densley Tingley[a], Martin Mayfield[a]
[a]Department of Civil & Structural Engineering, The University of Sheffield, Sheffield S1 3JD, UK
Tel: +44 (0) 114 2225728

ABSTRACT

Due to now outdated construction technology, houses which have not been retrofitted since construction typically fail to meet modern energy performance levels. However, identifying at a city scale which houses could benefit the most from retrofit solutions is currently a labour intensive process. In this paper, a system that uses a vehicle mounted camera to capture pictures of residential buildings and then performs semantic segmentation to differentiate components of captured buildings is presented. An ensemble of U-Net semantic segmentation models are trained to identify walls, roofs, chimneys, windows and doors from building façade images and differentiate between window and door instances which are partially visible or obscured. Results show that the ensemble of U-Net models achieved high accuracy in identifying walls, roofs and chimneys, moderate accuracy in identifying windows and low accuracy in identifying doors and instances of windows and doors which were partially visible or obscured. When U-Net models were retrained to identify doors or windows, irrespective of partially visible and obscured instances, a significant rise in door and window identification accuracy was observed. It is believed that a larger training dataset would produce significantly improved results across all classes. The results presented here prove the operational feasibility in the first part of a process to combine this model with high-resolution thermography and GPS for automating building retrofitting evaluations.

Keywords: deep learning; image segmentation; building retrofit; environmental modelling; U-Net

1 INTRODUCTION

Energy imprints on human lives everywhere and drives human society[1]. The high-speed development of the global economy and overpopulation has led to many problems including the energy crisis. The ever-increasing demand for limited non-renewable energy resources exceeding their supply will challenge the global economy and even threaten human's survival. Therefore, the necessity of saving energy for sustainability purposes becomes a significant concern under the current circumstances. The UK government has targeted a reduction of 80% carbon emissions by 2050[2].

Globally, a significant share of total energy end-use is consumed by buildings through their lifetime[3]. While about 35% of buildings in the EU are over 50 years old[4], the building renovation rate of the EU building stock is only 0.4-1.2% (varied for different countries) per year[4]. This has meant that approximately 75% of buildings in the EU can currently be classed as energy inefficient. This is a significant concern when buildings are believed to consume 40% of the energy demand and are responsible for 36% carbon emission in the EU at present[4]. Compared to replacing the current outdated energy-inefficient buildings with renovated energy-efficient buildings, retrofitting is usually a more cost-effective and feasible approach. In 2010, the UK government committed to reduce carbon emissions by enhancing the energy efficiency of seven million British homes by 2020[3].

The process of building retrofit evaluation is usually comprised of two stages: a building survey and subsequent decision making. Some efforts have been made to assess the sustainability of building retrofitting[5–7]. However, the building survey stage which usually aims to detect insulation absence, thermal leakage, defective installation and other similar issues is currently a labour intensive process. Prioritizing retrofit at city level becomes an important challenge. Therefore, an approach to automate the building energy efficiency survey is essential for retrofit plan making at city scale.

*menglin.dai@sheffield.ac.uk; phone +44 (0) 744 6885153

Remote Sensing Technologies and Applications in Urban Environments IV, edited by Thilo Erbertseder, Nektarios Chrysoulakis, Ying Zhang, Proc. of SPIE Vol. 11157, 1115708 · © 2019 SPIE
CCC code: 0277-786X/19/$21 · doi: 10.1117/12.2532828

Image segmentation which is a vital technology in the computer vision area has been applied to many areas (e.g. urban scene understanding[8]). The technology is aimed to divide an image into individual pixel groups which contain all pixels of the image and pixels in each group sharing similar properties[9]. Therefore, to automate the building thermal condition evaluation process, this technology is a critical step here.

In this paper, a methodology is presented which captures residential building images through the use of a vehicle-mounted camera and subsequently performs semantic segmentation to segment building façades in these captured images. The dataset is used to train an ensemble of U-Net semantic segmentation models to perform the building façade semantic segmentation.

The paper is organized as follows, the following section presents the related work of façade segmentation, as well as classic and state-of-art deep learning models, are reviewed. Section 3 provides details of the data capture system and description of the semantic segmentation model in depth. Details of the dataset and experimental results are given in section 4. Finally, conclusions are presented in section 5.

2 BACKGROUND

Pixel-wise façade segmentation is a significant challenge in many aspects, for example, the complex shape and miscellaneous exteriors of building facade objects (e.g. windows, doors) increase the difficulty to find a general solution for segmenting different facades. In some early façade segmentation works, efforts have been made by using multiple images for window localization and 3D reconstruction[10,11]. However, in a paper published in 2018, the authors argued that, currently, single-image façade segmentation was attracted more focus compared to multiple-image-based façade segmentation[12]. Also, two types of single-image-based façade segmentation methods were identified in the same paper, which are Grammar-based and classification-based methods[12].

Grammar-based methods parse façade images by generating initial labelling of façade object and then using shape grammars. Sets of rules, handcrafted or learnable, need to be defined and which will be integrated into a parse-tree to segment façades in a nodes-split manner[13]. The shape parse-tree grammar usually needs to be designed manually and often requires the involvement of experts[14]. The methods are able to exploit the space information of the façade objects (e.g. hierarchy and distribution) and architectural features of the dataset[12]. However, the grammar-based prior approaches were usually constrained to the dataset regularity which means they would not perform very well in datasets with fewer regularities[14]. And the size of these datasets with strong architectural inconsistency limits grammars to improve the segmentation performance by learning the dataset[14]. The grammar-based façade segmentation method has achieved satisfactory results in some building façade image benchmark datasets. An approach has achieved 90.8% pixel accuracy on the ECP[15] dataset but it is highly time-consuming[13,16]; in another grammar-based approach, the authors have validated their work on four state-of-art façade image datasets and obtained good results[14]. However, in general, the accuracy of grammar-based methods is usually below 85%, and efficiency is sacrificed for a higher accuracy algorithm[13,14].

Classification-based methods perform pixel-wise classification and usually are combined with an optimization method. An attempt has been made by using dynamic programming (DP) approach[17]. This approach encodes hard architectural constraints in the DP and optimizes more complex objects such as roof, sky, and chimney[17]. Then the approach is improved by exploiting the symmetry characteristics of the façade[18]. It contributes to deal with the problem of objects with occlusions and increase algorithm accuracy. The state-of-art efforts in façade segmentation area have also been made by using the Structured Random Forest (SRF) algorithm[12,13]. An iterative optimization approach is applied to improve the segmentation results from the SRF in the earlier approach[13]. In the following work, authors use the Regional Proposal Network (RPN) to detect the existence of interesting objects and generate features in the form of rectangular intensity boxes[12]. The features are employed as channels in the following SRF, and the results are optimized by a rectangular fitting optimizer[12]. This work achieved state-of-art accuracy at the time. Instead of employing the Convolutional Neural Network (CNN) as part of a façade segmentation pipeline[12], a trial has also been made on directly using CNN to segment façade images[19].

CNN technology has the advantages of without the need for involvement from an expert in certain areas to design handcraft features for image segmentation. With the advances of CNN[20] and Graphics Processing Units (GPUs) in recent years, many CNN models have been developed for image semantic segmentation purpose. Many of the image semantic segmentation models are based on a Fully Convolutional Network (FCN). This model replaces the fully connected layers initially for image-level classification purpose with convolution and pooling layers to realize pixel-wise classification[21].

This model also concatenates fine low-level features with coarse high-level features to combine information at different levels[21]. This approach has been widely adopted and tested on natural images[27–29]. Based on the structure of a FCN, U-Net was developed for the purpose of medical image segmentation[22]. The model uses a U-shape architecture with skip structures as well as data augmentation techniques[22]. The model proves the effectiveness of multiscale information exploitation and data augmentation in dataset size limitation conditions.

3 METHODOLOGY

3.1 Dataset

Images are captured from a Ladybug5+ camera rig mounted on a vehicle as it is being driving around an urban environment. A Ladybug5+ visual camera rig (30 Hz, resolution per sensor: 2048 x 2448 pixels, FOV: 90% of full sphere, IP65) is composed of six separate Sony IMX264 CMOS sensors with one on the top pointing upwards and the other five positioned along the sides forming a pentagon. After the raw images are captured via the vehicle-mounted Ladybug5+, they are labelled to create mask images before feeding into the ensemble of U-Net models. In this dataset, the raw images were segmented into 8 groups: wall, roof, chimney, door, alt_door, window, alt_window, background. In some of the state-of-art building façade datasets, e.g. eTRIMS[23], non-building objects like sky and vegetation have also been labelled. However, in this work, since evaluating the thermal performance of buildings is our future target, objects which are not related to buildings are not labelled.

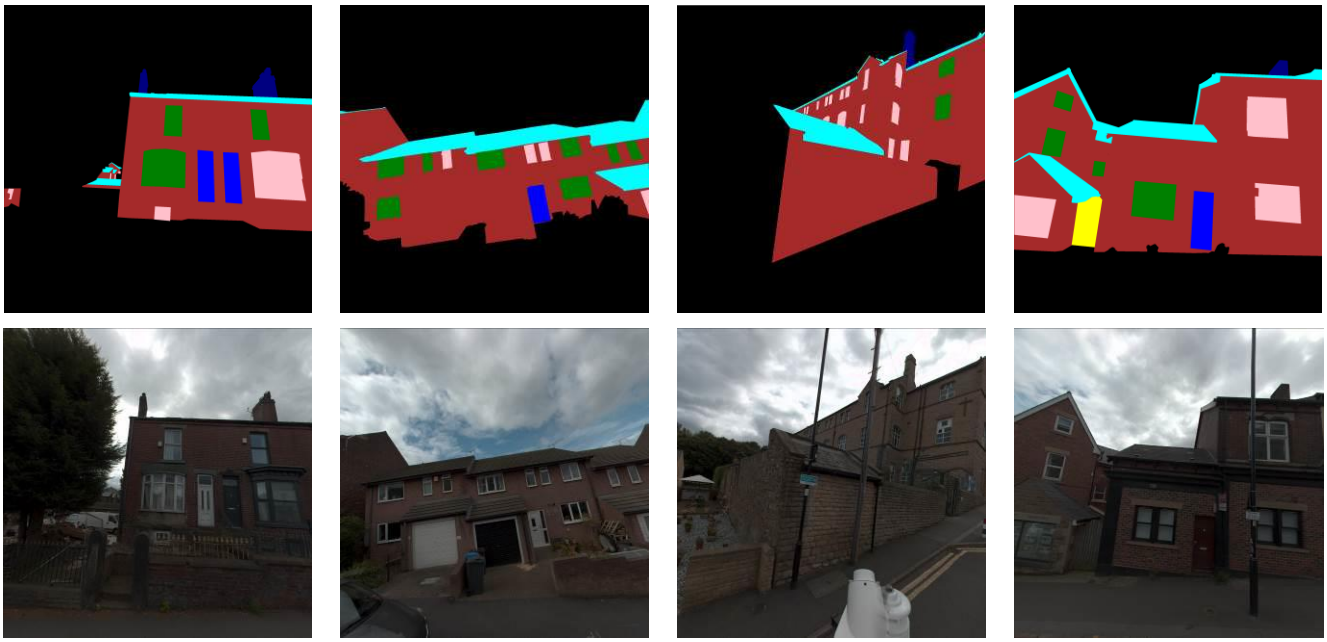

Figure 1. Image annotation examples. The colour coding of the images are as follows: alt door pixels are yellow; alt window pixels are pink; door pixels are blue; window pixels are green; chimney pixels are navy, roof pixels are cyan; wall pixels are brown; and background pixels are black.

The ‘wall’ class contains the whole wall area including neglectable and unavoidable objects. The neglectable and unavoidable objects are defined as things which are relatively small to the whole wall area and attached on walls, e.g. pipe, antenna, CCTV equipment etc. Only structural walls are included in the ‘wall’ class; other walls like a fence and boundary walls are not considered. Occlusions with a high density, such as vehicles and waste bins, are avoided during labelling. However, occlusions with a low density which means targeted objects can still be seen through the occlusions (ex. sparse bushes, metal bars) are included in the annotation. In the ‘roof’ class, we include all parts belonging to a roof (ex. rafter, joist eaves, rakes). The two classes with ‘alt’ prefix are for corresponding partially visible door and window objects which the occlusion annotation scheme is applied. These two classes also include opened and highly distorted objects.

Both of the two datasets are randomly split into 3 parts for training, validating and testing the deep learning model. The choice of the split ratio is very important and empirical. The split ratio varies in different datasets, MS COCO[24] uses the ratio (50%, 25%, 25%) for their training, validation and testing dataset, respectively; Cityscape[25] uses the ratio (60%, 10%, 30%) for their dataset split. Considering our datasets have a much smaller size and the tradition of (80%, 20%) training and validation datasets split ratio in the machine learning area, a unique ratio (80%, 5%, 15%) is adopted. Thus, we have 192 images for the training dataset, 12 images for the validation dataset and 36 images for the testing dataset. The size varies in individual class dataset due to availability.

Applying data augmentation method into deep learning is originally introduced in the AlexNet paper[20]. This method has been applied to medical imagery to intentionally produce more training images from the original ones before feeding the data into the U-Net model[22]. This is realized by performing multiple augmentation methods on original data, e.g. flip, rotate, shift, shear, brightness adjustment, etc. These methods are realized by using the inbuilt functions of TensorFlow[26], an open-source machine learning library. In our dataset, the horizontal flip is adopted. Also, the width and height shift is applied. In addition, the hue is adjusted. Our building façade dataset limits the application of many other data augmentation methods compared to medical image datasets. For example, vertical flip and right-angle rotation cannot be used here since buildings will not be either up-side-down or falling-over. The set of data augmentation methods and corresponding setting values are summarized in Table 1. The horizontal flip is probabilistically implemented with a 50% chance of occurring.

Table 1. Applied data augmentation summary.

Methods	Value
Flip horizontally	50%
Shift	10% on both horizontal and vertical direction
Hue	Δ=0.1

3.2 U-Net Ensemble

The approach adopted here uses multiple models to segment different classes separately and assemble results together in the end. The network architecture shown in Figure 1 is based on the U-Net which contains a contracting path, an expansive path and skip structures. The contracting path of the architecture used here has 6 convolutional blocks. Every block has two convolution layers with a 3×3 size filter with a stride of 1×1, dropout layer, batch normalization and rectifier activation. In addition, zero padding is applied in the convolution process to maintain the feature map dimension. These blocks will increase the number of feature maps from 3 to 1024. Max pooling with a stride of 2×2 is applied to each of these blocks except the last one. These max-pooling layers will decrease the feature map resolution from 256×256 to 8×8. The expansive path will increase the feature map dimension from 8×8 to 256×256 with 3×3 filter and stride of 2×2 deconvolution layer. The deconvolution layer will double the dimension of a feature map by two and decrease its number by two also. In every block of the expansive path, feature maps from the contracting path will be concatenated with the feature maps from the expansive path; and two convolution layers as same as the ones in the expansive path will be applied to reduce the number of feature maps. In the end, a convolution layer with a stride of 1×1 and sigmoid activation will be applied to reduce the number of feature maps to 1 that reflects the probability of the foreground segmentation. Finally, the full segmentation mask is generated by comparing the resulting probability maps of all classes.

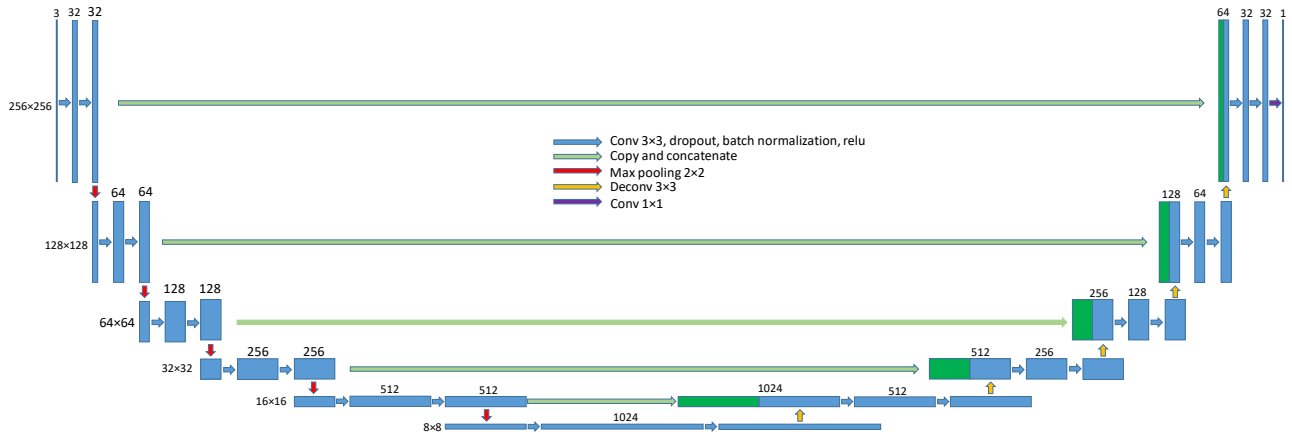

Figure 2. Single developed U-Net model.

The commonly-used binary-entropy loss function is used in each individual model except the chimney segmentation model. A unique cost function is adopted in the chimney segmentation network. Rather than using the cross-entropy based or quadratic cost function, a combined cost function of cross-entropy and dice loss is adopted. Dice loss is initially introduced in the paper of a deep learning model, V-Net[30], and it has better performance in class imbalanced problems. Our final cost function is to add the dice loss function to a binary cross-entropy loss function.

Deep neural networks are trained through a stochastic gradient-based optimizer to minimize the cost function regarding its parameters. The adaptive moment estimator (Adam) optimizer[31] is chosen here. Unlike the traditional stochastic gradient descent (SGD) optimizer in which the learning rate is a constant, the Adam optimizer can update the learning rate by utilizing the first and second moments of gradients. Other hyper-parameters are set as: dropout rate = 0.2, batch size = 3 and max epochs = 50.

3.3 Performance metrics

All the deep learning models are separately trained on the training dataset and validated on the validation dataset to primarily evaluate the performance of single models. Then, the output feature map from isolated models are combined together based on their probability values. The combined façade segmentation results are evaluated to provide a general approach performance. The single model will be assessed through various metrics: accuracy, precision, TPR (true positive rate), TNR (true negative rate) and F1 score. Accuracy and precision is defined as

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{1}$$

And

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{2}$$

In which TP, FP are the true positive, false positive measurements; and TN, FN denotes the true negative and false negative measurements. TPR and TNR are defined as

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{3}$$

And

$$\text{TNR} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{4}$$

Then, the F1 score is calculated by

$$\text{F1} = \frac{2 \times \text{Precision} \times \text{TPR}}{\text{Precision} + \text{TPR}} \tag{5}$$

The output feature maps of different classes are combined through a series of selections. The first selection is to determine whether a single-pixel belongs to the background class. This is realized by setting a threshold in all output feature maps. If the probability is lower than 50% in all feature maps, the pixel is classified as 'background'. If in one or more feature maps, the classification probability of the pixel is higher than the 50%, the pixel is classified as the class with the highest probability.

4 RESULTS AND DISCUSSION

All individual networks have been implemented with TensorFlow library and trained on an NVIDIA Quadro M1200 GPU with 4G memory. It took an average 50 minutes to run through an individual model.

Table 2 shows the results of the performance metrics for each individual U-Net model. While it can be seen that all models achieve very high Accuracy and TNR values, the TPR values for positively detecting doors, alt-doors and alt-windows is low. This highlights that due to the highly imbalanced positive and negative classes, using only Accuracy and TNR metrics would have been unreliable for measuring true model performance here. Instead, the F1 score appears to be a more reliable indicator of model performance. From the F1-score, it can be seen that the 'wall', 'chimney' and 'roof' class models achieved good performance, the 'window' class model achieved satisfactory performance and the 'door', 'alt-door' and 'alt-window' class models do not perform well.

Table 2. Performance metrics for semantic segmentation results.

Model	Accuracy	Precision	TPR	TNR	F1 score
Wall	0.929	0.893	0.846	0.961	0.869
Roof	0.987	0.670	0.883	0.990	0.762
Chimney	0.998	0.813	0.839	0.999	0.826
Door	0.953	0.322	0.143	0.987	0.198
Alt-door	0.985	0.156	0.318	0.989	0.209
Window	0.979	0.705	0.637	0.991	0.669
Alt-window	0.983	0	0	0.983	0

Figure 3 demonstrates prediction comparisons with their corresponding raw image and labelled image from each individual U-Net model. The images from the 'alt window' class model is not shown as the model fails to predict all 'alt window' objects.

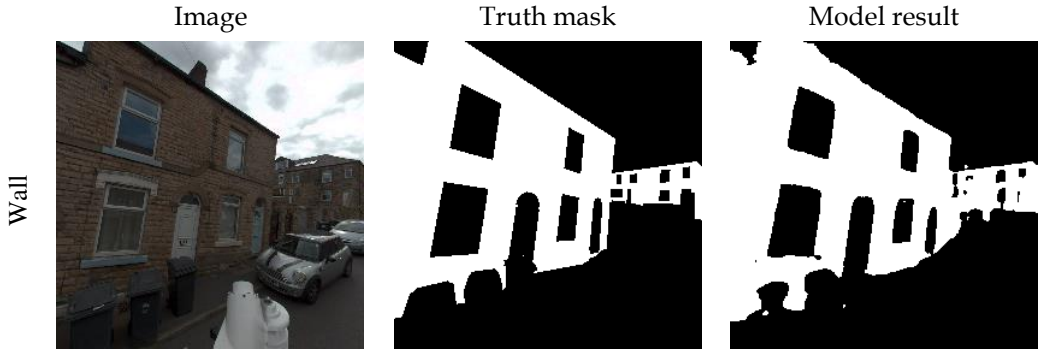

Figure 3. Class prediction examples.

It can be seen from Figure 3 that the 'window' class model wrongly covers the instance belonging to the 'alt window' class. The two classes of 'door' models produce inaccurate results, either failing to locate the instance or wrongly classifying other unrelated pixels.

To evaluate the effects of classifying window and door objects into alternative and non-alternative classes, the alternative and non-alternative classes are combined together and models retrained. Table 3 shows the resulting performance metrics of the combined window and door class models. In comparison with Table 2, an increase in F1 score and TPR values are seen.

Table 3. Performance metrics for semantic segmentation results for overall results of combined classes.

Model	Accuracy	Precision	TPR	TNR	F1 score
Window	0.981	0.735	0.870	0.986	0.796
Door	0.987	0.307	0.739	0.989	0.434

Figure 4 presents examples of classification results for the two combined classes. Visual inspection of Figure 4 shows that it can also be seen that the combined class models perform better than separate class models.

Figure 4. Combined class model prediction examples.

5 CONCLUSION AND FUTURE WORK

In this paper, a novel system to perform building façade image semantic segmentation on captured house images was presented. Our system consists of urban data capturing and an image semantic segmentation model. The data capturing component is a FLIR Ladybug5+ vehicle-mounted camera that captures images of houses while the vehicle is being driven. From the captured data, an initial building façade segmentation dataset of 240 house images was built. An ensemble of the U-Net deep convolutional neural networks was trained to perform pixel-wise classification for segmentation of building façades. The semantic segmentation model was trained to identify walls, roofs, chimneys, windows and doors from building façade images and differentiate between window and door instances which are partially visible or obscured.

The ensemble results achieved high accuracy in identifying walls, roofs and chimneys, moderate accuracy in identifying windows and low accuracy in identifying doors and instances of windows and doors which were partially visible or obscured. When U-Net models were retrained to identify doors or windows, irrespective of partially visible and obscured instances, a significant rise in door and window identification accuracy was observed. It is believed that a larger training dataset would produce significantly improved results across all classes.

The results presented here prove the operational feasibility in the first part of a process to combine this model with high-resolution thermography and GPS for automating building retrofitting evaluations. The proposed system inspires the potential of a system which can automate the city scale building retrofit plan making process, and it makes it possible to analyses the city scale building material consumptions which is beneficial to potential circular economy researches. Future work will involve significantly increasing the size of the dataset and investigate the prediction interaction between classes.

ACKNOWLEDGEMENT

This research was supported by the EPSRC City Observatory Research Platform for Innovation and Analytics [EP/R013411/1].

REFERENCES

[1] Coyle, E. D. and Simmons, R. A., [Understanding the Global Energy Crisis Recommended Citation, Knowledge], Purdue University Press (2014).

[2] Kelly, M. J., "Retrofitting the existing UK building stock," Build. Res. Inf. **37**(2) (2009).

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

[4] European Commission., "Energy performance of buildings," 2019, <https://ec.europa.eu/energy/en/topics/energy-efficiency/energy-performance-of-buildings> (16 May 2019).

[5] Asadi, E., da Silva, M. G., Antunes, C. H. and Dias, L., "Multi-objective optimization for building retrofit strategies: A model and an application," Energy Build. **44**, 81–87 (2012).

[6] Cetiner, I. and Edis, E., "An environmental and economic sustainability assessment method for the retrofitting of residential buildings," Energy Build. **74**, 132–140 (2014).

[7] Ding, G. K. C., "Sustainable construction—The role of environmental assessment tools," J. Environ. Manage. **86**(3), 451–464 (2008).

[8] Zhao, H., Shi, J., Qi, X., Wang, X. and Jia, J., "Pyramid Scene Parsing Network," 2017 IEEE Conf. Comput. Vis. Pattern Recognit., 6230–6239, IEEE (2017).

[9] Zhang, Y., "Image Segmentation in the Last 40 Years," [Encyclopedia of Information Science and Technology, Second Edition], IGI Global, 1818–1823 (2009).

[10] Mayer, H. and Reznik, S., "MCMC linked with implicit shape models and plane sweeping for 3D building facade interpretation in image sequences," Pcv (2006).

[11] Reznik, S. and Mayer, H., "Implicit Shape Models, Self-Diagnosis, and Model Selection for 3D Facade Interpretation," Photogramm. - Fernerkundung - Geoinf. PFG (2008).

[12] Rahmani, K. and Mayer, H., "HIGH QUALITY FACADE SEGMENTATION BASED on STRUCTURED RANDOM FOREST, REGION PROPOSAL NETWORK and RECTANGULAR FITTING," ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci. **4**(2), 223–230 (2018).

[13] Rahmani, K., Huang, H. and Mayer, H., "FACADE SEGMENTATION WITH A STRUCTURED RANDOM FOREST," ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci. **IV-1/W1**, 175–181 (2017).

[14] Gadde, R., Marlet, ·Renaud, Paragios, N. and Marlet, R., "Learning Grammars for Architecture-Specific Facade Parsing," Int. J. Comput. Vis. **117**, 290–316 (2016).

[15] Teboul, O., Kokkinos, I., Simon, L., Koutsourakis, P. and Paragios, N., "Shape grammar parsing via reinforcement learning," Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., 2273–2280 (2011).

[16] Koziński, M., Obozinski, G. and Marlet, R., "Beyond Procedural Facade Parsing: Bidirectional Alignment via Linear Programming," D. Cremers, I. Reid, H. Saito, and M.-H. Yang, Eds., Springer International Publishing, Cham, 79–94 (2015).

[17] Cohen, A., Schwing, A. G. and Pollefeys, M., "Efficient structured parsing of facades using dynamic programming," Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., 3206–3213 (2014).

[18] Cohen, A., Oswald, M. R., Liu, Y. and Pollefeys, M., "Symmetry-Aware Façade Parsing with Occlusions," 2017 Int. Conf. 3D Vis., 393–401, IEEE (2017).

[19] Schmitz, M. and Mayer, H., "A CONVOLUTIONAL NETWORK FOR SEMANTIC FACADE SEGMENTATION AND INTERPRETATION," ISPRS - Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci. **XLI-B3**, 709–715 (2016).

[20] Krizhevsky, A., Sutskever, I. and Hinton, G. E., "ImageNet Classification with Deep Convolutional Neural Networks" (2012).

[21] Long, J., Shelhamer, E. and Darrell, T., "Fully Convolutional Networks for Semantic Segmentation," 2015 IEEE Conf. Comput. Vis. Pattern Recognit. **31**(1), 3431–3440 (2014).

[22] Ronneberger, O., Fischer, P. and Brox, T., "U-Net: Convolutional Networks for Biomedical Image Segmentation" (2015).

[23] Korč, F. and Förstner, W., "eTRIMS Image Database for Interpreting Images of Man-Made Scenes" (2009).

[24] Lin, T.-Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., Perona, P., Ramanan, D., Zitnick, C. L. and Dollár, P., "Microsoft COCO: Common Objects in Context" (2014).

[25] Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S. and Schiele, B., "The Cityscapes Dataset for Semantic Urban Scene Understanding," 2016 IEEE Conf. Comput. Vis. Pattern Recognit., 3213–3223, IEEE (2016).

[26] Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G.,

Isard, M., Kudlur, M., Levenberg, J., Monga, R., Moore, S., Murray, D. G., Steiner, B., Tucker, P., Vasudevan, V., Warden, P., et al., "TensorFlow: A System for Large-Scale Machine Learning," 12th USENIX Symp. Oper. Syst. Des. Implement. (2016).

[27] Badrinarayanan, V., Kendall, A. and Cipolla, R., "SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation," IEEE Trans. Pattern Anal. Mach. Intell. **39**(12), 2481–2495 (2015).

[28] Iglovikov, V. and Shvets, A., "TernausNet: U-Net with VGG11 Encoder Pre-Trained on ImageNet for Image Segmentation" (2018).

[29] Li, R., Liu, W., Yang, L., Sun, S., Hu, W., Zhang, F. and Li, W., "DeepUNet: A Deep Fully Convolutional Network for Pixel-Level Sea-Land Segmentation," IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. (2018).

[30] Milletari, F., Navab, N. and Ahmadi, S.-A., "V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation," 2016 Fourth Int. Conf. 3D Vis., 565–571, IEEE (2016).

[31] Kingma, D. P. and Ba, J., "Adam: A Method for Stochastic Optimization" (2014).