An Active Learning Approach to Floorplan Image Annotation for Energy Assessment

Dhoyazan Al-Turki‡, Marios Kyriakou‡, Shadi Basurra, Mohamed Medhat Gaber, and Mohammed M. Abdelsamea*

Abstract—Floorplan energy assessments offer a highly effective way of assessing the energy efficiency of residential homes without actually needing to be present. Through computer modelling, the heat loss or gain of the building can be accurately determined, allowing planners and homeowners to create energy-efficient renovations or redevelopment plans. Floorplan assessments use a variety of data, such as CAD drawings and visualisation tools, to calculate the air infiltration rate and thermal transmittance through the housing structure. By identifying areas where there is potential for improvement in the way the home uses energy, floorplan assessments allow an accurate examination of risk factors for occupants and for measures that could improve the comfort level. However, manual annotation of floorplans can be a daunting task. In this paper, we propose a novel active learning model to detect and annotate the main elements inside floorplan images. Our model is designed to help energy assessors automatically analyse floorplan images, which is a challenging problem due to the time-consuming annotation process. Our active learning approach was initially trained on 500 annotated images and gradually learnt from a large dataset of unlabelled 4500 images, achieving mAP score of 0.833, precision score of 0.972, and recall score of 0.950. The annotated dataset will be publicly available under a Creative Commons licence.

Index Terms—Floorplan image analysis, active learning, object detection, deep learning, energy efficiency.

1 Introduction

The UK government has identified 6 million houses built with inefficient solid walls that require improvement. The government has tried to tackle this issue by creating policy initiatives, such as the Green Deal Home Improvement Fund (GDHIF), the Birmingham Energy Savers (ended in October 2015) and the Green Deal (ended in July 2015), to retrofit domestic housing and other buildings (with a target of energy efficiency improvements to be made to 60,000 homes and 1,000 other buildings by 2020). Usually, these schemes, or any retrofit project, start with an assessment phase to determine whether - or not - the building will materialise its potential savings, and if the cost going towards the improvements will be paid back within a predefined time span. Given the large scale of any scheme related to increasing energy efficiency, both existing and new built buildings, and the capacity required to undertake energy efficiency assessments, it is inevitable to use non-professionals to perform the assessments.Energy assessors currently use the Standard Assessment Procedure (SAP, RdSAP). This is a methodology recommended by the government to assess and compare the energy and environmental performance of dwellings. Its purpose is to provide accurate and reliable assessments of dwelling energy performances that are needed to underpin energy and environmental policy initiatives.

However, SAP and the RdSAP suffer from various limitations. They are 'steady-state models' run using monthly parameter inputs. SAP uses steady-state monthly average heat transfer calculations, effectively 12 sets of numbers, and does not consider any dynamic heat transfer over time. SAP assessment is based on standardised assumptions for occupancy and behaviour, enabling a like-for-like comparison of dwelling performance. SAP predictions are based on 'standard' climatic conditions and assumptions about 'standard' rather than necessarily average occupant behaviour, such as heating patterns for a given type of heating system. Such occupant influences on aspects of demand can vary with demographics, lifestyle of occupancy patterns throughout UK households. It is feared that the implementation of poor advice given on the basis of these simplistic calculations will influence the energy performance of buildings and carbon emissions for many years to come. On the other hand, more complex tools, such as building energy simulation, can offer more comprehensive testing than SAP. They provide a testing platform to measure any interaction between the thermal zones and the environment and the occupants. The model system with fast dynamics while trading off simulation speed requires huge computational power for precision. These tools can model the combined heat and mass transfer that counts for air movement between zones, including the effect of window blinds and building orientation in response to current and future weather data. The main challenge to using these tools is that Non-engineers can struggle to understand the concept of simulation, especially since it involves creating a 3D building model that replicates the physical structure of the existing building. This requires comprehensive understanding of drawing, building geometry, scaling and The ability to absorb three-dimensional

- *D. Al-Turki, M. Kyriakou, S. Basurra, M. Gaber, and M. Abdelsamea are with the School of Computing and Digital Technology, Birmingham City University, Birmingham, B4 7BD, UK.*
- *M. Gaber is also with Faculty of Computer Science and Engineering, Galala University, Suez 435611, Egypt and M. Abdelsamea is also with the Faculty of Computers and Information, Assiut University, Assiut, 71515, Egypt.*
‡these authours contributed equally to this work.
** corresponding author: mohammed.abdelsamea@bcu.ac.uk.*

Manuscript received April 19, 2005; revised August 26, 2015.

spaces; The EcRoFit project aims to scale up the use of the currently available building simulation and optimisation techniques beyond professional users such as engineers, architects and researchers and bring them to a wider range of potential users. The key output will be the development, testing and application of a new tool – iRet. This paper proposes a technique that will allow energy assessors to generate 3D building models from 2D floor plans in a matter of minutes by utilising various AI and computer vision techniques.

Computer vision provides an effective solution for the automatic detection of objects directly from digitised floorplan images. Several algorithms have been developed for this purpose such as R-CNN, SSD, and YOLO, just to mention a few. YOLO-V5 is one of the modern versions of YOLO, which demonstrates its efficiency in providing accurate predictions. It uses bounding boxes to define the required element inside the image or video.These models are limited to the availability of enough annotated images. In computer vision, labelling involves adding tags to raw data, such as images and videos [1]. There are many tools that could be used for this purpose, for example "LabelImg" which is compatible with state-of-the-art object detection models such as YOLOv5. More precisely, these object detection models require a bounding box labelling technique to define four points which are x,y coordinators for the box centre and height and width of that box surrounding the target object of interest.

One of the main challenging problems of using such manual annotation tools is the efforts needed from operators (humans) with previous knowledge of the problem context. The need to optimise the time, speed of training and prediction, and improve the accuracy of the model is a crucial concept for machine learning models. However, obtaining labelled data for new tasks is difficult and expensive. Motivated by such issues, active learning [2] is providing an efficient solution to allow the minimum user intervention to achieve the annotation task more robustly. Active learning is the subset of machine learning in which a learning algorithm can interactively query a user to label the data with the desired outputs. In active learning, the algorithm proactively selects the subset of examples to be labelled next from the unlabelled data pool. The fundamental idea behind the active learner concept is that a machine learning algorithm could reach a higher accuracy level while using a smaller number of training labels if it were allowed to choose the data from which it needs to learn.

In this paper, we propose an uncertainty-aware active learning workflow that has been trained on a small number of well-annotated images (500 floorplan images) to iteratively correct the annotation of 4500 challenging images, resulting in a dataset of 5000 well-annotated floorplan images. The image has been processed and annotated in a way that provides a benchmark data set for energy assessors to facilitate a robust and unbiased solution for energy assessment from the 3D models of the building constructed by the detected and required objects of the associated floorplan image.

The following are the main contributions of this work.

- A dataset of floorplan images that is annotated for a direct application to energy performance assessment.
- A novel uncertainty-aware active learning model for object detection.
- A novel computer vision pipeline for floorplan image analysis for the energy performance assessment purpose.

The paper is organised as follows. Section 2 discusses the related work and currently available resources and applications linked to floorplan image analysis. Section 3 details our dataset and the proposed models. Section 4 describes the experimental setup and discusses the results. Finally, Section 5 concludes and summarises the findings of this work.

2 Related Work

In this section, we start with a brief description of the existing publicly available floorplan image datasets and then review the state-of-the-art object detection and active learning models.

Among the existing floorplan image datasets, the CVC-FP dataset was introduced in [3], which is annotated for architectural objects and their structural relations. In [3], a tool has been proposed for general-purpose ground truthing. The output of the tool was in standard Scalable Vector Graphics (SVG). The output has mainly been focused on wall segmentation and room detection tasks, and performance evaluation was performed on wall segmentation using the Jaccard Index (JI). ROBIN (Repository Of BuildIng plaNs) is another floorplan dataset that was introduced in [4], where a deep learning framework, called Deep Architecture for fiNdIng alikE layouts (DANIEL), was proposed to retrieve similar floorplan layouts from a repository. In this approach, the authors perform an extensive analysis comparing the performance of individual hidden layers for the proposed floor plan retrieval task. Several deep object detection models have previously been proposed to deal with objects in floorplan images [5], [6], [7]. For instance, [8] utilised an image dataset called CubiCasa5K consisting of 5000 images that have been categorised into more than 80 types of floorplan objects. In [8], an enhanced multitask convolutional neural network (CNN) was proposed to detect objects in the CubiCasa5K dataset. [9] employed several models, including Mask R-CNN with Deformable Convolutional Networks (DCN) [10] and Convolutional Neural Networks (CNNs) to deal with the same problem, concluding that DCN outperforms CNNs in detecting objects in 2D floorplan images. Similarly, [11], employs the Mask R-CNN model to detect and generate the segmentation maps of objects in a small floorplan dataset, showing potentially accurate and reliable results.

2.1 Object Detection

As a common application for floorplan object detection, YOLO (You Only Look Once) [12] has shown great potential to deal with several objects within floorplan images. The most recent version of YOLO, YOLO v7 has shown successful attempts to increase object detection's speed and precision. In order to determine the bounding boxes and class probabilities for objects in an image, YOLO v7 uses a single convolutional neural network. The network was trained to segment a picture into a grid of cells, with each cell charged with determining if there are objects present. The final object detections for the entire image are determined by combining the predictions from the grid. The inclusion of anchor boxes, which are predetermined bounding boxes that help the network make predictions, is one of the noticeable upgrades in YOLO v7. To enhance the precision and speed of object identification, YOLO v7 also makes use of an upgraded post-processing method and a more effective backbone network. YOLO v7 is a robust and quick object detection technology that has been applied in numerous real-world scenarios. However, it is not suitable for real-time applications on devices with limited processing capacity and continues to consume a significant amount of resources.

Another common method, called Faster R-CNN [13], was designed to combine the advantages of region proposal-based object detection algorithms with convolutional neural networks. A region proposal network (RPN) and a fast R-CNN detector make up the algorithm's two primary parts. The RPN produces region proposals, and the objects in these proposals are classified using the Fast R-CNN detector. Faster R-CNN employs an RPN to create region suggestions after using a convolutional neural network to extract features from the input image. The CNN-generated feature maps are used by the RPN to forecast the objectness scores and bounding box regression parameters for each region proposal. The RPN slides a tiny network over the feature maps. The items in each area proposal are then classified by the Fast R-CNN detector utilising the features produced by the CNN.

The single shot multi-box detector (SSD) was another attempt [14], which uses a single forward run over a deep neural network to predict object class and position, which is intended to be quick and effective. The main advancement of SSD is the use of numerous default boxes for object detection that is matched with the ground truth boxes of the training images. The authors demonstrate that compared to earlier object detection techniques, SSD may be taught with fewer negative samples by using numerous default boxes. As a result, the process of training the model is quicker and more effective. Additionally, SSD processes the input image at several scales using a multi-scale feature extraction network, which enables it to detect objects of various sizes. Additionally, this reduces the problem of scale variance in object detection.

Recent object detection models within the EfficientDet [15] family achieved excellent accuracy while being computationally effective. In contrast to the majority existing object detection models, EfficientDet scales from small to large sizes in a single step by combining an efficient neural network architectural design with a cutting-edge scaling technique. EfficientDet's design is based on the DenseNet structure and includes a single stage multi-scale feature pyramid network (FPN) that can predict bounding boxes at various scales. Additionally, it uses weight standardisation, the Swish activation function, and the effective parallelizable anchor-free detection head.

Another method, called the CenterNet method [16], was designed in a way to ensure that object instances are treated as points in an image, with each point serving as the object instance's centre. In addition to predicting the offset of keypoints (such as corners and edges) from the object centres, CenterNet creates heatmaps of the object centres. The shape and location of the object are then determined by using the keypoint triplets.

2.2 Floorplan image analysis

Floorplan images are usually created using tools [8], where annotations are inserted in a certain order to guarantee consistency and accuracy in labelling. A quality assurance (QA) procedure is in place to regulate label accuracy and positioning precision. The QA procedure consists of two rounds, with the annotator performing the first round and a different QA operator performing the second round, respectively, to guarantee that any potential errors have been fixed.

An approach for creating unstructured 3D point clouds into 2D floorplans with topological linkages between walls was proposed in [17]. It consists of two steps: 2D CAD floorplan production and 3D reconstruction. For each floor, floor segmentation based on horizontal planes and wall proposals based on vertical planes is applied in the 3D reconstruction portion to identify the walls. The horizontal projection of wall proposals is utilised to detect walls, and the detected wall points are then used to create an Industry Foundation Classes (IFC) file. The IFC file is used to develop structural parts in the second section, and the 2D floorplan CAD is produced using this data.

A system for identifying and rebuilding floor plans was proposed in [18], which is composed of two components: reconstruction and recognition. The floorplan area, structural features, supplemental text and symbols, and scale information from picture pixels are all recognised by the recognition component. The information acquired is transformed into a vectorised expression in the reconstruction phase. The scientists precisely extracted horizontal, vertical, and inclined walls from floorplan photos using a vectorisation technique based on iterative optimisation. Vectorised floor designs were created using two different approaches to depict the walls. One method involves drawing the wall using the edge line, which is a portion of the optimal room contour polygon, while the other involves drawing the wall using the centre line.

An image analysis pipeline has been developed in [18] for floorplan analysis that makes use of a network with outputs for two segmentation maps and a collection of heatmaps. The location and dimensions of all potential elements were inferred from the points of interest located. Finally, the semantics of the floor plan, including the types of rooms and icons, are acquired using the two segmentation maps.

In [19], a floorplan image analysis method based on a four-module strategy was proposed. First, a CNN encoder uses an input floorplan image to extract features. The second module was a room boundary decoder (RBD), to forecast room boundaries such as walls, doors, and windows using the extracted features. Room Type Decoder (RTD) was the third module to predict room types such as the living room, bedroom, bathroom, and closet. Finally, a boundary attention aggregated model (BAAM) unit was employed to make use of feature maps from the RBD and RTD modules as inputs to produce features in order to identify the type of room.

2.3 Active Learning

As a semi-supervised learning method, active learning combines both a large number of unlabelled instances and a limited number of samples that have labels. It can boost the labelling of new data by choosing unlabelled samples estimated to significantly improve the model's speed and accuracy after labelling. Several algorithms have been proposed for active learning, for example, membership query Synthesis [20] is a method of creating/synthesising queries based on a set of membership criteria, where the learner/-model can create a new instance from scratch that satisfies the definition of the instance space because the learning model is aware of the concept of the instance space. Stream-based selective sampling method has been proposed in [21], which focuses on obtaining unlabeled instances efficiently that allows the learner to test the data sample against the actual distribution before deciding whether or not to request its labeling. It should be noted that the concept of stream-based selective sampling has been investigated in several practical activities, including Natural Language Processing (NLP) [22] and Speech Recognition / Speech Classification (SR) [23].

In [24], an active learning approach has been proposed to classify images that can produce a competitive classifier from a small number of labelled training examples. In this framework, deep convolutional neural networks were used, which aims to simultaneously update the feature representation and classifier with useful annotated samples. It also provides a cost-effective sample selection technique to enhance classification performance with fewer hand annotations. The method selects high confidence samples for feature learning from the unlabelled set, and then automatically assigns pseudo-labels to these samples. The tests demonstrate that the "Cost-Effective Active Learning (CEAL)" architecture can produce encouraging results on two difficult picture classification datasets. It selects the high confidence samples from DU, whose entropy is smaller than the threshold, then they assign clearly predicted pseudo labels to them.

In [25], an active learning algorithm called WBetGS (Weighted Batch Ensemble for Global Sampling) was proposed as a way to address class imbalance in the batch while taking into account the aggregation of data from various outputs and batch boxes. The approach reduces the effects of class imbalance between background and object categories by extracting class-balanced information, and a weighted algorithm is introduced to promote mAP more successfully.

In [26], a paradigm for active learning for video-based person re-identification (re-ID) was proposed. This is achieved by incorporating an active learning strategy into a deep learning system to overcome the data-scarce problem. By creating a sample criterion, the technique concentrates on choosing the most insightful tracked-pairs (also known as true positives) for annotation. The adaptive resampling step and view-aware sampling technique assist in choosing the appropriate candidates for annotation. Pseudo labels that are initialised and updated with the gained annotations are used to train the re-ID model. The suggested approach is straightforward and demonstrates its efficacy on three video-based person re-ID datasets.

In [27], a randomly selected set of images for annotation was selected and trained in a Single Shot Multibox Detector (SSD) [28] model. Subsequently, the method picks up a set/batch of images, annotates them, and retrains the model again; this procedure is repeated until the unlabelled pool of images is empty. The key idea of pool-based active learning is to rank the dataset (D_i) samples in some manner; consequently, the algorithm tries to analyse/evaluate the whole dataset before choosing the optimal query or the combination of queries [29]. The next step is to identify the best subset of unlabelled data ($D_{U,i}\ \forall i \in T$, where $T \subseteq D$) to be labelled so that a model is trained on the most informative samples [30] and label the remaining dataset samples as accurately as possible. In particular, the Active Learning model can focus on many data samples simultaneously, in contrast to stream-based selective sampling [31]. Thus, based on sampling techniques or informativeness metrics, the most informative data samples are chosen from the pool of unlabelled data samples. The following sample to be queried is chosen using this new information after determining which of the most informative samples are detected [32].

Unlike the previously proposed active learning models, our proposed model clusters the images according to the confidence and level of certainty given by the backbone object detection model. This is to allow for selecting the most certain samples for (re)training the object detection models to generate more accurate predictions, and more importantly to improve the robustness of the model in handling difficult samples. Low-quality or less certain samples may require some human interaction during the training process, which guides the proposed model to more accurate features to correct its predictions.

3 Material and Methods

3.1 DataSet

In this section, we discuss the dataset used in this work, its sources, and a brief comparison with other existing datasets, see Table 1. The following are the commonly used floorplan image datasets:

3.1.1 ROBIN (Repository Of BuildIng plaNs)

ROBIN [4] consists of 510 floorplans that are used to automatically find existing building layouts from a repository that may help the architect ensure reuse of the design and timely completion of projects. They propose Deep Architecture for fiNdIng alikE Layouts (DANIEL), so an architect can search from the existing project's repository of layouts (e.g., floorplan) and give accurate recommendations to the buyers.

3.1.2 CVC-FP: Database for structural floor plan analysis

CVC-FP [3] has 122 floorplans that are annotated for architectural objects and their structural relations. They have implemented a ground truthing tool, named the SGT tool, which allows us to make this kind of information specific in a natural manner. The tool allows one to define own object classes and properties, multiple labeling options are possible, and grants cooperative work. The dataset is fully annotated for structural symbols: rooms, walls, doors, windows, parking doors, and room separations.

3.1.3 CubiCasa5K: A Dataset and an Improved Multi-Task Model for Floorplan Image Analysis

CubiCasa5K [8] is a large-scale floorplan image dataset containing 5000 samples annotated of over 80 floorplan object categories. The annotations of the dataset are performed in a dense and versatile manner using polygons to separate the different objects.

3.1.4 Our Dataset

Our dataset is created from the above open source datasets and consists of 5000 images, 500 of the images were initially annotated manually, then we propose an active learning model to help with the annotation of the remaining images. The active learning model was used to detect three types of classes, which are rooms, windows and doors, to eventually generate a 3D model of the building. This work is part of the proposed framework for assessing energy performance, which provides construction, installation, and building services related to energy efficiency. In particular, addresses property energy professionals and retrofit professionals. It is worth mentioning that the total of the three sources data sets is 5634 samples, and we have taken 5000 of them randomly as follows:

- 300 samples from ROBIN dataset
- 100 samples from CVC-FP dataset
- 4600 samples from CubiCasa5K dataset

Then, we selected 500 images from these 5000 images randomly for manual annotation.

3.2 Methodology

Our proposed pipeline has been designed in a way to offer an easy-to-use, highly visual interface to create fast and accurate energy assessments, see Figure 1. It applies multi-objective optimisation techniques to compare multiple options for holistic retrofit measures and can compare and contrast variations of product specifications. The main components of our proposed pipeline can be described as

- Floor Designer: This component is used by the performance energy assessor to draw rooms, windows, doors, and floors of a house and add some materials and constructions for a building to be used for the performance energy assessment. The user will start creating a floor, then define zones (rooms) within this floor, then add windows and doors for each zone (room), then add the construction materials used for all walls (internal and external), windows and doors, and repeat this process for each floor of this house to finally define the type of roof for this house and its construction material.
- Smart Sketcher: it is a machine learning model that receives an image of the floorplan as input from the assessor and detects all rooms, windows, and doors, and then converts the 2D floorplan into a 3D model. It consists of four steps which are detecting the floorplan elements (rooms, windows, and doors), then defining the external frame of this house using OpenCV, next generating the 3D model of the house using ThreeJS, and finally generating the external view of the house. This component can interact with web/mobile applications using Flask web API or it could be plugged into the web/mobile application using Google Tensorflow Lite Convertor.
- Light Compass: this component is used to define the house orientation (North, South, East, or West), location, and the weather data to be used as parameters for a performance energy assessment. The user will start defining the orientation with the help of the mobile/tablet built-in orientation sensor, then use the device location sensor to define the house location and city (where based on the city, the user can call a web API to retrieve the weather data for this location).
- Floor Designer component, Smart Sketcher and Light Compass component will store all the data inside .idf(Input Data File) which is the required format for the EnergyPlus engine. IDF is an ASCII file that contains the data that describe the building and HVAC system to be simulated.
- JESS (JEplus Simulation Service): they are a set of web APIs that are used for the online simulation service for EnergyPlus. It receives the .idf file generated from the above components and retrieves a report showing the predicted gas consumption, electricity consumption, and CO2 emissions for the building.
- JEA: is a parametric, sensitivity analysis, and optimisation engine used to generate different results of energy performance based on different situations and then recommend the best option for the building.
- Flask web API is a gateway between the trained AI models and the web version of the solution.
- Tensorflow Lite is a light version of the machine learning model developed to support the mobile version of the trained models.

3.2.1 Image Annotation

In this work, we used the bounding box technique to annotate our image dataset, which is compatible with most of the object detection deep learning modelsTwo independent annotators were recruited to annotate 500 images by drawing visual boxes around the objects of interest in each image and saving the result in a.txt file with the class ID, the coordinate (x, y) for the centre of the boundary, and the length and width of the box. Once the annotation process was completed, only annotations with the highest agreement between the two annotators were accepted. A third expert annotator reviewed all the annotated objects and corrected the mislabeled ones. This process results in accurate annotations for the initial dataset (which consists of 500 images), which is required by our active learning model to complete the annotations of the whole dataset, see Figure 3.

TABLE 1
A comparison between existing floorplan image datasets.

Name	Size	Purpose
ROBIN	512 images	deep learning approaches to automatically analyse building floor plan images and retrieve similar plans from a large-scale repository. The proposed technique can find application in an online property sale/rent scenario where the buyer has preferred features related to the room semantics.
CVC-FP	122 images	The dataset is fully ground-truthed for the structural symbols: rooms, walls, doors, windows, parking doors, and room separations. It not only makes their locations in the images specific but also includes structural relations between them that are of special interest for analysis systems
CubiCasa5K	5000 images	Multi-task learning scheme which uses the 'multi-task uncertainty loss' based on 5000 images and starts detecting 80 different element classes inside the floorplan

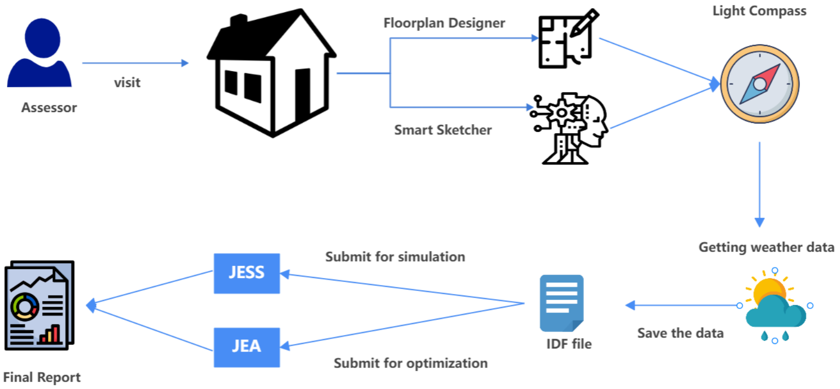

Fig. 1. The generic architecture and main components of our proposed pipeline for energy simulation and optimisation.

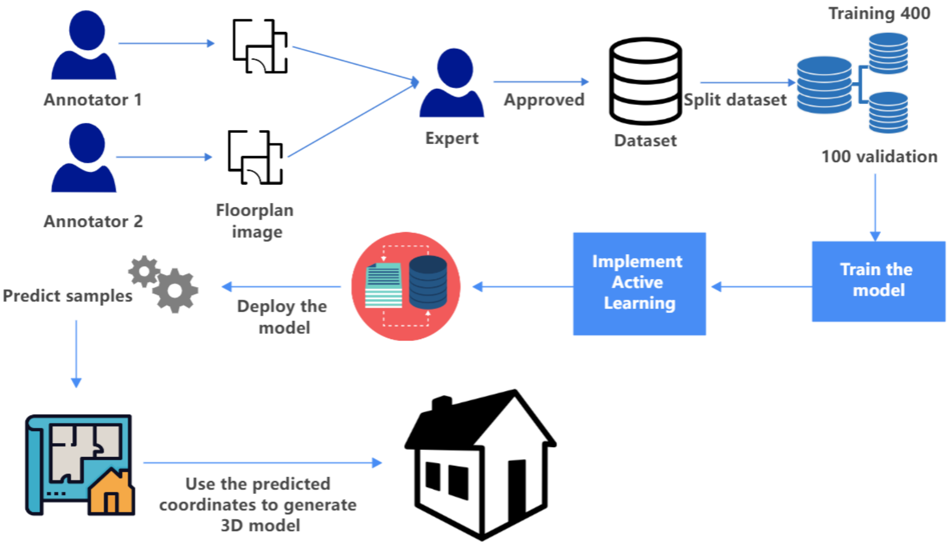

Fig. 2. The Smart Sketcher process includes annotation, training, active learning, 3D generating, and deployment.

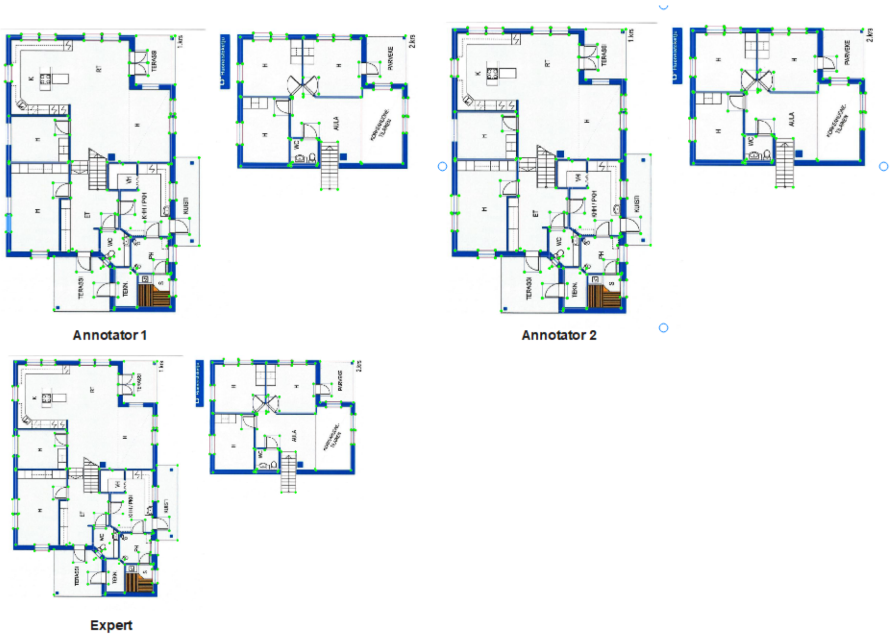

Fig. 3. The first annotator missed 1 door and 2 windows, the second annotator 2 missed 1 window, the last one annotated by the expert.

3.2.2 Object detection and active learning model

Here we used YOLO-v5 to learn from a small well-annotated dataset of 500 floorplan images. We first used 400 images for training/validation while the remaining 100 images were used to test the trained YOLO-v5 model. We used 500 epochs with 16 batch sizes as a result of the hyperparameters fine-tuning.

Once the YOLO-v5 model was trained, we started the training process of our active learning technique on the remaining unlabelled images of the dataset (the 4500 images). The initial pool of 4,500 images was used by YOLO-v5, where objects were detected and the probabilities of the predictions were calculated accordingly. The probabilities generated by YOLO-v5 were used to measure how confident the model is in detecting and classifying objects within the floorplan images. Hence, such confidence scores were used to break up the initial pool of floorplan images into clusters. For example, in our setting, we divided the 4,500 images into 3 groups, named 'high-confidence', 'middle-confidence', and 'low-confidence', based on the uncertainty of the predictions. Samples in the 'high-confidence' group/-cluster are associated with high probability values obtained by the object detection model and usually represent samples that are well-handeled by the object detection model. On the other hand, samples in the 'low-confidence' cluster are hard samples that the object detection model is usually uncertain to give an accurate prediction. The pool of unlabelled images is updated in a dynamic way, where samples in the high-confidence cluster are used to retrain the object detection model and removed from the pool, while in the next iteration the clustering space is recreated and the process is repeated until the model is converged (where no unlabelled sample is remaining).

To construct the cluster space, we adopted a new metric, called β-score based on Least Confidence (LC) metric to evaluate the overall performance on a unlabelled test image. Since each image contains multiple elements such as doors, windows, and rooms, the b-score was defined as the average of the Least Confidence scores for all the bounding boxes of an image. More precisely, we calculate the least confidence score for each bounding box in an image and then calculate the average of the least confidence scores for all the bounding boxes in the same image, as described by Equation 1:

$$\beta_{avg}(I) = \frac{\sum_{i=1}^{n} LC_i}{n}, \tag{1}$$

where n is the number of elements in a test unlabelled image $I(x)$, x is the pixel location. LC_i is the least confidence score associated with each element i in the image I, which is defined in Equation 2.

$$LC_i(I) = 1 - C, \tag{2}$$

where C is the prediction probability obtained by the trained YOLO-v5 model.

LC_i are measured for all the bounding boxes in an image and then the average least confidence score ($\beta_{avg}(I())$) is calculated. The pool of unlabelled images is clustered into three groups based on the confidence scores of the images (high, medium and low), where we used a trial-and-error method to achieve cluster assignment.

The uncertainty-aware clustering method of the floorplan images was proposed to improve the efficiency and effectiveness of our proposed active learning model. This is by iteratively updating the cluster space where the goal is to maximise the number of samples in the high-confidence cluster and minimise the number of samples in the other clusters. Our active learning model is re-trained iteratively using new samples (of the high-confidence cluster) while the clustering space is updated. Consequently, our active learning model will have the ability to gradually filer out hard samples by moving them from the low/mid-confidence clusters to the high-confidence cluster that will be included in the training set to re-train YOLO-v5. In this way, YOLO-v5 will gradually learn how to deal with difficult samples by eventually moving all samples in the low/mid-confidence clusters. In addition, to improve and ensure the robustness of the model, an expert intervention component was designed to review and correct the predictions before moving samples to the training set. Figure 4 illustrates the evolution of the clustering space using our active learning model. Moreover, Table 2 reports the evolution of the clusters in terms of the number of samples.

In algorithm 1, we describe the implementation flow of our uncertainty-aware active learning model, see also Figure 5.

Algorithm 1 Active Learning Algorithm

```
Require: The YOLOv5 model ($M$) pre-trained on the 5000 floorplan images
Require: The Unlabeled pool of the $4500$ images ($x$) is defined as $U$
for  $x \in U$ do
    Predict the bounding boxes for each $x$
    for $bb \in x$ do
        Calculate the Least Confidence per $bb$
        Store the LC score in a list
    end for
    Calculate the Average Least Confidence ($ALC = b_{score}$) per $x$
    if $b_{score} < 0.4$ then
        Store the $x$ in ”High Confidence” Cluster
    else if $0.4 < b_{score} < 0.6$ then
        Store the $x$ in ”Mid Confidence” Cluster
    else
        Store the $x$ in ”Low Confidence” Cluster
    end if
    Move the Clusters in the training set
end for
Convert the predicted annotations to YOLO format
for High Confidence Cluster, Mid Confidence Cluster, Low Confidence Cluster  do
    Re-train Model ($M$)
end for
```

3.2.3 Generating 3D Model

Once we collected all the coordinates of the bounding boxes from the predictions of the 3 categories (zones, windows, and doors), we fetched the coordinates of the zones and converted them into cubes by adding fixed height; see algorithm 2.

For energy performance assessment, only external doors are required. Therefore, we detected the external door using OpenCV library and manually added the height, which is 75% of the zone height. Then, we used ThreeJS to demonstrate the result on a web page using Java Script. ThreeJS is a cross-browser JavaScript library and application programming interface used to create and display animated 3D computer graphics in a web browser using WebGL, see Figure 6.

Algorithm 2 Collecting Bounding Boxes Coordinates

```
Require: the bounding boxes coordinates(BC) from the trained model prediction
                                        ▷ Loop all the predicted bounding boxes
for  $zonesCor \in BC$ do
                                  ▷ Loop for each element in the prediction result
    for $zc \in zonesCor$ do          ▷ Check if the element is Zone
        if $zc_{class} ==' Zone'$ then
            Calculate the width (xmax - xmin) and length (ymax - ymin) of each bounding box
            Calculate the width (xmax - xmin) and length (ymax - ymin) of each bounding box
            Calculate the center point based on xmin, xmax, ymin, ymax
            Adding all these elements into an array
        end if
    end for
end for
```

4 Experimental Results

In this section, we demonstrate the efficiency and effectiveness of our object detection and active learning models. For the performance evaluation of the object detection model, our initial labelled dataset (500 images) has been divided into two subsets, the training set (80%) and the unlabelled pool of the test set (20%). The remaining 4500 images were used by our active learning model. The following table shows the different iterations for the each experiment with the number of samples for each cluster

4.1 Evaluation Metrics

In this work, we adopted the precision, recall, and mean average precision (mAP) matrices to demonstrate the effectiveness of object detection, which are defined as

- Confusion Matrix (CM) - A matrix that displays the performance of an algorithm.
- Recall (R) - The percentage of real positive values that are correctly identified.
- Precision(P) - The percentage of positive identifications that are actually correct.
- Mean Average Precision (mAP) - The most widely employed metric in research papers; it combines all predictions into a single value. mAP is computed by calculating the Average Precision (AP) for each class and averaging all results for all objects.

4.2 YOLOv5s

In our initial experiment, we utilised YOLOv5m (medium) with a batch size of 32 and 300 epochs and employed the pre-trained COCO weights. Although the model has a mAp of 0.883, which indicates promising performance throughout the training phase, Figure 7 shows that the model performs noisily in terms of the learning rate.

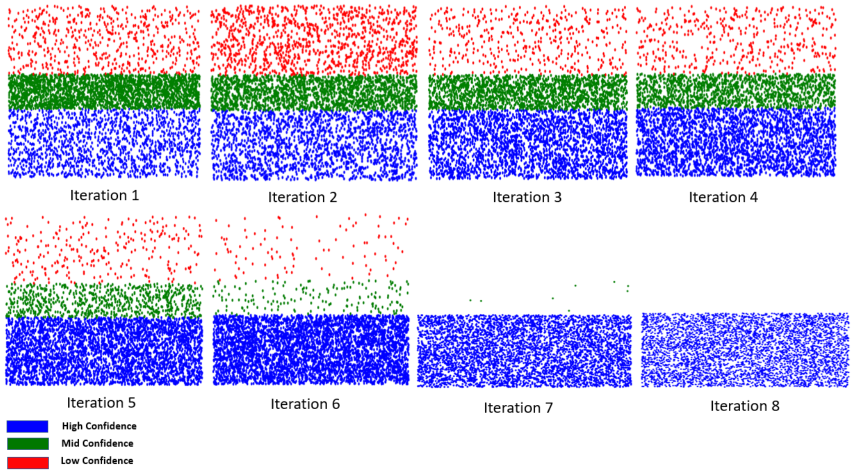

Fig. 4. The evolution of the clustering space of the low- mid-, and high- confidence groups during the training process of our active learning model.

TABLE 2
Clusters evolution in terms of the number of samples.

Iteration number	#Training Samples	#high-conf. samples	#mid-conf. samples	#low-conf. samples
1	400	1194	2707	599
2	1594	291	1932	1083
3	1885	796	1789	430
4	2681	246	1526	447
5	2927	943	772	258
6	3870	771	163	96
7	4641	251	8	0
8	4892	8	0	0

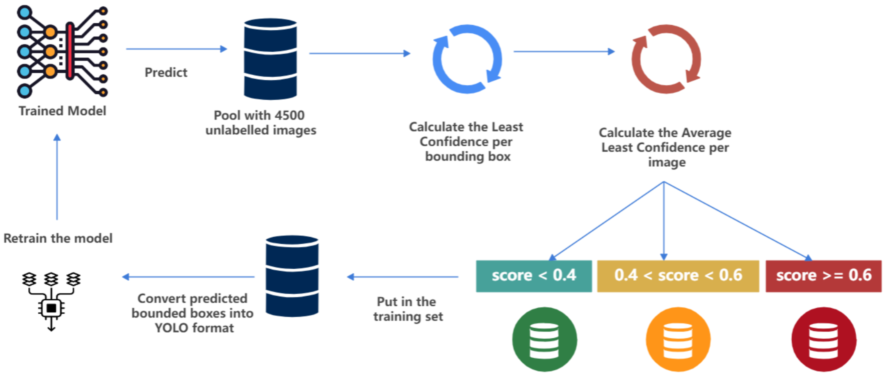
Trained Model
Predict
Pool with 4500 unlabelled images
Calculate the Least Confidence per bounding box
Calculate the Average Least Confidence per image
score < 0.4
0.4 < score < 0.6
score >= 0.6
Put in the training set
Convert predicted bounded boxes into YOLO format
Retrain the model

Fig. 5. The active learning process, including calculating the confidence for each box, the least confidence for the image, and clustering the samples into high, mid, and low confidence.

Figure 7 shows that the YOLOv5 model is struggling to identify the object in an image due to the noise that exists in the model's mAP, precision, and recall. Therefore, we changed the model to the YOLOv5s (small) training model and merely increased the model's initial learning rate ($lr0$) from 0.01 to 0.001. Since the pre-trained weights are trained on the COCO dataset based on real-world objects, we trained YOLOv5s from scratch in an end-to-end way without the pre-trained weights of COCO dataset. As a consequence, YOLOv5s performs slightly lower in terms of mAP scoring, equal to 0.833, while precision and recall scores were 0.778 and 0.824, respectively. Even if the mAP score drops slightly due to the modification of the learning rate, the model performs better during the learning process. Figure 8 demonstrates the new model's performance using 300 epochs.

Figures 9 and 10 demonstrate the overall performance of YOLOv5s on the given dataset in terms of the main evaluation metrics.

Figure 11 represents the precision-recall curve. The precision-recall curve depicts the trade-off between precision and recall at various thresholds. A high area below the

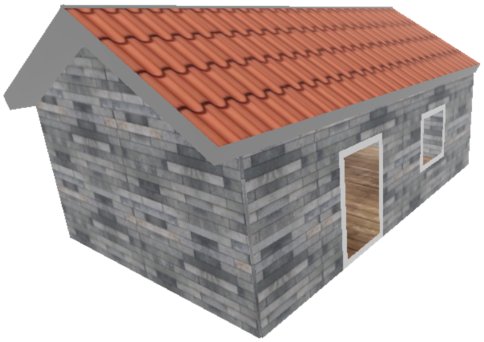

Fig. 6. 3D Model generated using the detected objects from a 2D floorplan image and ThreeJS library.

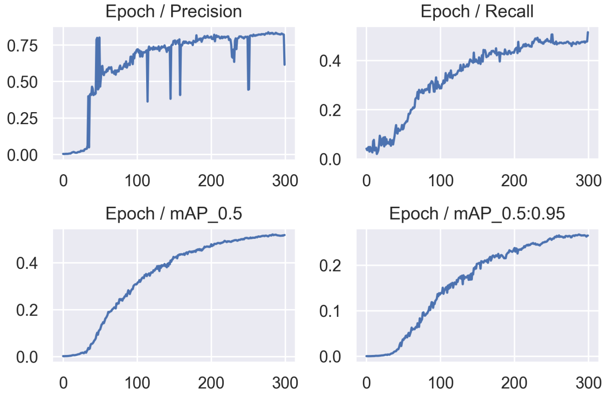

Fig. 7. Precision, Recall and mAP with the default hyperparameters, obtained by YOLOv5m on the small annotated dataset.

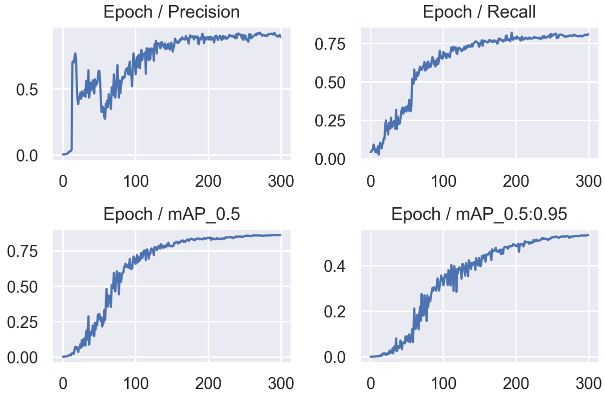

Fig. 8. Precision, Recall and mAP with the modified hyperparameter and epochs=300, obtained by YOLOv5s on the small annotated dataset.

curve indicates both high recall and precision. In addition, high accuracy is correlated with a low false positive rate, and high recall is correlated with a low false negative rate.

Our model's precision-recall pattern reveals a bowing curve toward the point $(1, 1)$, which denotes a skilled model. The combined precision-recall scores for the door, room, and window classes were 0.841, 0.869, and 0.763, respectively, giving our model a total score of 0.824.

5 Conclusion

This paper proposes a novel technique that will enable energy assessors to generate 3D building models from conventional 2D floor plans. This can be utilised to run building energy simulations to test building energy performance while taking into consideration occupants behaviour, and changing weather conditions over the entire year. Building simulations offer more detailed energy ratings than currently used tools in the UK, such as SAP and RdSAP. We proposed an active learning workflow based on YOLO-v5 to detect three elements inside the floorplan images (zones, windows, and doors) and help the energy assessors to do their job in an automated and efficient fashion. The active learning model was trained on a small number of well-annotated images to iteratively correct the annotation of 4500 challenging images. The dataset has been processed and annotated in a way that provides a benchmark data set to facilitate a robust and unbiased solution for energy assessment from the 3D models of the building constructed by the detected and required objects of the associated floorplan image. This work is part of a proposed framework for assessing energy performance, which provides construction, installation, and building services related to energy efficiency.

Acknowledgments

This work has been partially supported by the EcRoFit project, funded by the European Regional Development Fund $(ERDF)$ and Birmingham City University in the UK.

References

[1] Pedro Miguel Lima de Sousa Reis. Data labeling tools for computer vision: a review. 2022.

[2] Clemens-Alexander Brust, Christoph Käding, and Joachim Denzler. Active learning for deep object detection. *arXiv preprint arXiv:1809.09875*, 2018.

[3] Lluís-Pere de las Heras, Oriol Ramos Terrades, Sergi Robles, and Gemma Sánchez. Cvc-fp and sgt: a new database for structural floor plan analysis and its groundtruthing tool. *International Journal on Document Analysis and Recognition (IJDAR)*, 18(1):15–30, 2015.

[4] Divya Sharma, Nitin Gupta, Chiranjoy Chattopadhyay, and Sameep Mehta. Daniel: A deep architecture for automatic analysis and retrieval of building floor plans. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, volume 1, pages 420–425. IEEE, 2017.

[5] Alessio Barducci and Simone Marinai. Object recognition in floor plans by graphs of white connected components. In *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)*, pages 298–301. IEEE, 2012.

[6] Zahra Ziran and Simone Marinai. Object detection in floor plan images. In *IAPR workshop on artificial neural networks in pattern recognition*, pages 383–394. Springer, 2018.

[7] Shreya Goyal, Chiranjoy Chattopadhyay, and Gaurav Bhatnagar. Knowledge-driven description synthesis for floor plan interpretation. *International Journal on Document Analysis and Recognition (IJDAR)*, 24(1):19–32, 2021.

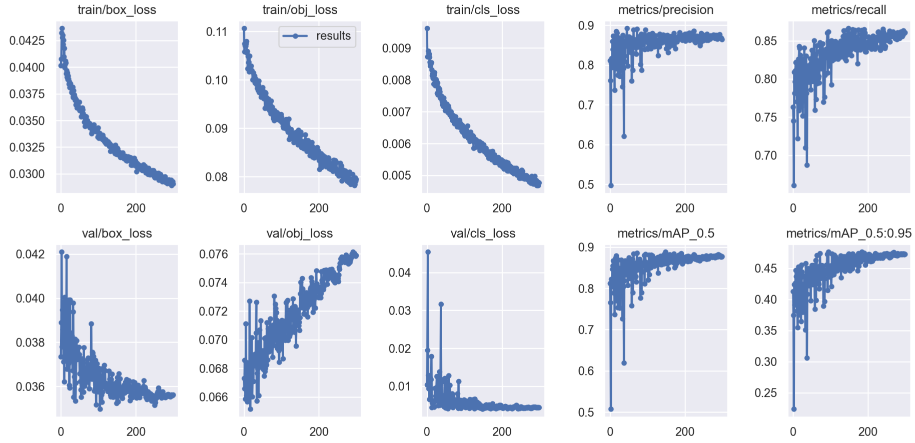

Fig. 9. The results of different metrics obtained by YOLOv5 after 300 epochs on the complete dataset.

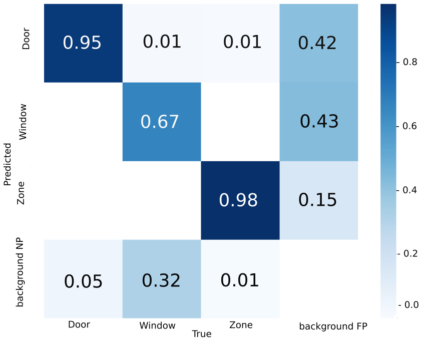

Fig. 10. The confusion matrix of the training set with 300 epochs.

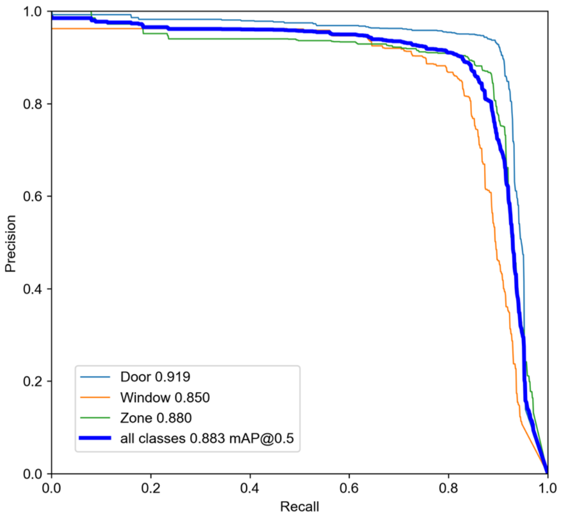

Fig. 11. Precision-Recall Curve of YOLOv5 with 300 epochs.

[8] Ahti Kalervo, Juha Ylioinas, Markus Häikiö, Antti Karhu, and Juho Kannala. Cubicasa5k: A dataset and an improved multi-task model for floorplan image analysis. In *Scandinavian Conference on Image Analysis*, pages 28–40. Springer, 2019.

[9] Shashank Mishra, Khurram Azeem Hashmi, Alain Pagani, Marcus Liwicki, Didier Stricker, and Muhammad Zeshan Afzal. Towards robust object detection in floor plan images: A data augmentation approach. *Applied Sciences*, 11(23):11174, 2021.

[10] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 764–773, 2017.

[11] Fredrik Sandelin. Semantic and instance segmentation of room features in floor plans using mask r-cnn, 2019.

[12] Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. *arXiv preprint arXiv:2207.02696*, 2022.

[13] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015.

[14] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pages 21–37. Springer, 2016.

[15] Mingxing Tan, Ruoming Pang, and Quoc V Le. Efficientdet: Scalable and efficient object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10781–10790, 2020.

[16] Kaiwen Duan, Song Bai, Lingxi Xie, Honggang Qi, Qingming Huang, and Qi Tian. Centernet: Keypoint triplets for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6569–6578, 2019.

[17] Uuganbayar Gankhuyag and Ji-Hyeong Han. Automatic 2d floorplan cad generation from 3d point clouds. *Applied Sciences*, 10(8):2817, 2020.

[18] Xiaolei Lv, Shengchu Zhao, Xinyang Yu, and Binqiang Zhao. Residential floor plan recognition and reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16717–16726, 2021.

[19] Zhongguo Xu, Cheng Yang, Salah Alheejawi, Naresh Jha, Syed Mehadi, and Mrinal Mandal. Floor plan semantic segmentation using deep learning with boundary attention aggregated mechanism. In *2021 4th International Conference on Artificial Intelligence and Pattern Recognition*, pages 346–353, 2021.

[20] Hardik Dave. Active learning sampling strategies.

[21] Joseph Nelson. What is active learning ?

[22] Cuong Van Tran, Tuong Tri Nguyen, Dinh Tuyen Hoang, Dosam Hwang, and Ngoc Thanh Nguyen. Active learning-based approach for named entity recognition on short text streams. In *Multimedia and Network Information Systems*, pages 321–330. Springer, 2017.

[23] Wenjing Han, Eduardo Coutinho, Huabin Ruan, Haifeng Li, Björn Schuller, Xiaojie Yu, and Xuan Zhu. Semi-supervised active learning for sound classification in hybrid learning environments. *PloS one*, 11(9):e0162075, 2016.

[24] Keze Wang, Dongyu Zhang, Ya Li, Ruimao Zhang, and Liang Lin. Cost-effective active learning for deep image classification. *IEEE Transactions on Circuits and Systems for Video Technology*, 27(12):2591–2600, 2016.

[25] Soumya Roy, Asim Unmesh, and Vinay P Namboodiri. Deep active learning for object detection. In *BMVC*, volume 362, page 91, 2018.

[26] Menglin Wang, Baisheng Lai, Zhongming Jin, Xiaojin Gong, Jianqiang Huang, and Xiansheng Hua. Deep active learning for video-based person re-identification. *arXiv preprint arXiv:1812.05785*, 2018.

[27] Soumya Roy, Asim Unmesh, and Vinay P Namboodiri. Deep active learning for object detection. In *BMVC*, page 91, 2018.

[28] Sheping Zhai, Dingrong Shang, Shuhuan Wang, and Susu Dong. Df-ssd: An improved ssd object detection algorithm based on densenet and feature fusion. *IEEE access*, 8:24344–24357, 2020.

[29] Dongrui Wu. Pool-based sequential active learning for regression. *IEEE transactions on neural networks and learning systems*, 30(5):1348–1359, 2018.

[30] Liping Yang, Alan M MacEachren, Prasenjit Mitra, and Teresa Onorati. Visually-enabled active deep learning for (geo) text and image classification: a review. *ISPRS International Journal of Geo-Information*, 7(2):65, 2018.

[31] Pengzhen Ren, Yun Xiao, Xiaojun Chang, Po-Yao Huang, Zhihui Li, Brij B Gupta, Xiaojiang Chen, and Xin Wang. A survey of deep active learning. *ACM computing surveys (CSUR)*, 54(9):1–40, 2021.

[32] Daniel Gissin and Shai Shalev-Shwartz. Discriminative active learning. *arXiv preprint arXiv:1907.06347*, 2019.

Dhoyazan Al-Turki. Dhoyazan is an IT professional with 16+ years of experience in the field of web, mobile and AI apps development, for many governmental and private sectors. He was the official speaker for Microsoft in Jordan (2016, 2017 and 2018) and was invited as speaker in Tedx Jordan on 2020 (AI and Mysterious Future). Currently, he is a PhD student in Birmingham City University and his main research interests are in computer vision and AI, including deep learning, active learning, and image processing.

Marios Kyriacou. Marios is a Data Scientist at the CYENS-Centre of Excellence in Cyprus. He studied Mathematics and Statistics at the University of Cyprus, with a specialisation in Applied Mathematics. He recently completed his MSc in Artificial Intelligence at Birmingham City University. His main research area revolves around machine learning and deep learning models as well as computer vision and image processing applications.

Shadi Basurra. Shadi Basurra received his Bsc (Hons) degree in Computer Science from Exeter University, the UK, and MSc in Distributed Systems and Networks from Kent University at Canterbury, UK. He obtained his Ph.D. from the University of Bath in collaboration with Bristol University. After completing his Ph.D., Shadi worked at Sony Corporation developing Goal Decision Systems, he then moved on to work as a Research Fellow at the Zero Carbon Lab – Birmingham City University. He recently joined the Computer Science Centre as a Senior Lecturer in Software Engineering at Birmingham City University. Shadi research interests include simulation and emulation of networks (vehicular, mesh and sensor ad hoc network), game theory, multi-agent systems, multi-objective optimisation, model calibration and dynamic simulation of zero-carbon design and retrofit of buildings.

Mohamed Medhat Gaber. Mohamed is a Professor in Data Analytics at the School of Computing and Digital Technology, Birmingham City University. Mohamed received his PhD from Monash University, Australia. He then held appointments with the University of Sydney, CSIRO, and Monash University, all in Australia. Prior to joining Birmingham City University, Mohamed worked for the Robert Gordon University as a Reader in Computer Science and at the University of Portsmouth as a Senior Lecturer in Computer Science, both in the UK. He has published over 200 papers, co-authored 3 monograph-style books, and edited/co-edited6 books on data mining and knowledge discovery. His work has attracted well over eight thousand citations, with an h-index of 43. Mohamed has served in the program committees of major conferences related to data mining, including ICDM, PAKDD, ECML/PKDD and ICML. He has also co-chaired numerous scientific events on various data mining topics. Professor Gaber is recognised as a Fellow of the British Higher Education Academy (HEA). He is also a member of the International Panel of Expert Advisers for the Australasian Data Mining Conferences. In 2007, he was awarded the CSIRO teamwork award.

Mohammed Abdelsamea. Abdelsamea is currently a Senior Lecturer in Data and Information Science at the School of Computing and Digital Technology, Birmingham City University. He is also a Fellow of the British Higher Education Academy. Before joining BCU, he worked for the School of Computer Science at Nottingham University, Mechanochemical Cell Biology at Warwick University, Nottingham Molecular Pathology Node (NMPN), and Division of Cancer and Stem Cells both at Nottingham Medical School, as a Research Fellow. In 2016 , he was a Marie Curie Research Fellow at the School of Computer Science at Nottingham University. Before moving to the UK, he worked as a Lecturer of Computer Science for Assiut University, Egypt. He was also a Visiting Researcher at Robert Gordon University, Aberdeen, UK. Abdelsamea has received his Ph.D. degree (with Doctor Europaeus) in Computer Science and Engineering from IMT - Institute for Advanced Studies, Lucca, Italy. His main research interests are in computer vision including: image processing, deep learning, data mining and machine learning, pattern recognition, and image analysis.