Journal of Building Engineering 98 (2024) 111355

Contents lists available at ScienceDirect

Journal of Building Engineering

journal homepage: www.elsevier.com/locate/jobe

Deep learning models for vision-based occupancy detection in high occupancy buildings

Wuxia Zhang [a,*], John Calautit [a], Paige Wenbin Tien [a], Yupeng Wu [a], Shuangyu Wei [b]

[a] Department of Architecture and Built Environment, University of Nottingham, Nottingham, NG7 2RD, UK
[b] Welsh School of Architecture, Cardiff University, Cardiff, CF10 3NB, UK

ARTICLE INFO

Keywords:
Deep learning
Computer vision
Energy efficiency
Building energy simulation
Occupancy detection
You Only Look Once (YOLO)
Faster Region-based Convolutional Neural Networks (Faster R-CNN)
Single Shot MultiBox Detector (SSD)

ABSTRACT

Accurate occupancy information is crucial for enhancing energy efficiency and reducing carbon emissions in buildings. However, the inherent unpredictability of occupants introduces uncertainties in energy analysis and control strategy development. To address these challenges, this study proposes a vision-based method employing state-of-the-art deep learning models to capture real-time occupancy profiles in crowded indoor spaces. Utilising a self-collected image dataset, various deep learning models, including Single Shot MultiBox Detector (SSD), Faster Region-based Convolutional Neural Networks (Faster R-CNN), and different versions of You Only Look Once (YOLO) were trained and evaluated. An experiment was conducted in a lecture room equipped with cameras and environmental sensors to evaluate the performance of each model in terms of precision, computational efficiency, and adaptability to varying occupancy levels during a lecture session. The session included varying occupancy conditions: entering (barely occupied), during the lecture (typical occupancy), and leaving the room (again barely occupied). Among the models tested, YOLOv8x exhibited the best performance in terms of accuracy, while SSD lagged notably. The impact on the detection performance of various locations of camera setups was also explored. Energy simulations revealed that deep learning-based model generated occupancy profiles significantly deviated from conventional "fixed" occupancy profiles, resulting in a 13.45 % variation in predicted heating energy demand. However, compared to the ground truth, these profiles showed minimal variation (up to 6.72 %) for the Faster R-CNN and YOLO models, highlighting their accuracy and robustness. Additionally, although the deep learning-based occupancy profiles generally overpredicted the recorded data, the CO_2 concentration trends they predicted aligned closely with the recorded data, unlike the "fixed" occupancy profiles. The findings underscore the importance of realistic occupancy profiles for reliable energy predictions in buildings and demonstrate the potential of the proposed vision-based method for advancing occupancy detection and building energy management.

1. Introduction

Buildings are major energy consumers, responsible for up to 40 % of global energy usage to provide healthy, safe, and comfortable environments for their occupants [1]. Accurate occupancy data is crucial for predicting energy usage in buildings and developing control strategies to improve energy efficiency and performance. Understanding and monitoring occupancy patterns can lead to

* Corresponding author.
E-mail address: wuxia.zhang@nottingham.ac.uk (W. Zhang).

https://doi.org/10.1016/j.jobe.2024.111355
Received 25 July 2024; Received in revised form 30 October 2024; Accepted 14 November 2024
Available online 16 November 2024

List of abbreviations

PIR	Passive Infrared
RFID	Radio Frequency Identification
HOG	Histogram of Oriented Gradients
SVM	Support Machine Vector
SOM	Self Organising Map
CNN	Convolutional Neural Network
R-CNN	Regions with Convolutional Neural Networks
K-NN	K-Nearest Neighbours
GAN	Generative Adversarial Network
DBF	Deepsort, dynamic Bayesian fusion
FCHD	Convolutional Head Detector
YOLO	You Only Look Once
SSD	Single Shot Detector
RPN	Region Proposal Network
mAP	Mean Average Precision
MIT	Mean Inference Time
IOU	Intersection over Union
VOCs	Volatile Organic Compounds
PM 2.5	Particulate Matter 2.5
BES	Building Energy Simulation
IES VE	Integrated Environmental Solutions Virtual Environment
CIBSE	Chartered Institution of Building Services Engineers
TP	True Positive
FP	False Positive
FN	False Negative
TN	True Negative

significant improvements in energy management, helping to reduce overall energy consumption and carbon emissions [2]. A significant issue with current energy management systems is their reliance on static or fixed schedules for predicting occupancy and controlling building systems. These schedules are typically based on assumed occupancy patterns and do not account for the actual, real-time presence of occupants [3].

As a result, static schedules can lead to inefficiencies, such as ventilating, heating, cooling, and lighting spaces that are unoccupied, or failing to adequately condition spaces that are occupied outside of expected times. This mismatch between predicted and actual occupancy not only wastes energy but also reduces the comfort and satisfaction of building occupants [4]. Addressing these inefficiencies requires more dynamic and accurate methods of occupancy detection that can adapt to the real-time presence and movement of people within buildings [5,6].

Traditionally, occupancy detection methods have relied on a variety of sensors such as passive infrared (PIR) sensors [5], carbon dioxide (CO_2) sensors [6], radio-frequency identification (RFID) systems [7], and electricity meters [8]. While these technologies provide useful data, they often lack the accuracy and granularity needed to account for the dynamic nature of human occupancy, especially in high occupancy environments where the number of occupants and their movements can vary significantly. More recent approaches have incorporated WiFi signals [9] and cameras [10] to monitor occupancy. While WiFi-based methods can offer improved coverage, they still struggle with accuracy and real-time responsiveness. Cameras, on the other hand, present a promising solution by using visual data to track occupancy more precisely [11]. Cameras continuously monitor visual cues to track people's presence, movements, and interactions, which is crucial in dynamic, crowded settings. This is particularly valuable in environments where understanding occupancy patterns and behaviours is essential.

Earlier computer vision methods [12] for occupancy detection often required extensive computational resources and were hindered by the limitations of available hardware and algorithms. These methods typically involved complex feature extraction and pattern recognition, which were not robust enough for real-time applications and, hence, were not scalable or efficient for widespread adoption. In recent years, improvements in computational power, alongside the advancements in computer vision and deep learning have opened new avenues for occupancy detection. Computer vision-based approaches offer a minimally disruptive and highly accurate means of monitoring occupancy by analysing visual data captured from cameras. Deep learning, particularly Convolutional Neural Networks (CNNs), has transformed the field of computer vision [13]. CNNs are designed to automatically and adaptively learn spatial hierarchies of features from input images. This ability to learn and extract intricate patterns and features makes CNNs particularly effective for tasks such as object detection and recognition, which are critical for accurate occupancy detection [14].

Studies have shown that the CNN method performs particularly well in high-density scenes, making it suitable for environments where the number of occupants can vary widely and rapidly. This enhanced capability is due to CNNs' proficiency in handling occlusions and complex visual data, allowing for more reliable detection and tracking of individuals in crowded settings. Specifically,

models such as Single Shot MultiBox Detector (SSD), Faster Region-based Convolutional Neural Networks (Faster R-CNN), and You Only Look Once (YOLO) have shown great promise.

These models enhance occupancy detection by processing large amounts of visual data to identify and count occupants in real-time. SSD and YOLO are known for their speed and efficiency, making them suitable for applications requiring real-time analysis. Faster R-CNN, while slightly slower, provides higher accuracy and is effective in complex scenes with varying levels of occupancy. Despite their promising performance in controlled environments, the effectiveness of these methods in real-world building settings has not been thoroughly examined, particularly in situations where the presence of people is essential for building energy management [15]. Further research is needed to validate the performance of deep learning models in real-world building scenarios and to understand their impact on energy efficiency. This leads us to the next section, where we review the existing literature on vision-based occupancy detection methods and their application in building energy management.

2. Literature review

The exploration of vision-based methods for occupancy detection and prediction in buildings has gained considerable attention in recent years. Table 1 outlines examples of recent research efforts in this area, highlighting a variety of algorithms and data collection techniques utilised over the years. For instance, Yang et al. [16] employed a framework that utilises CNN-based density estimation methods to fuse image information from surveillance videos to obtain accurate and high spatial-temporal resolution indoor occupancy information. This framework trains an ML-based ensemble model to predict occupancy schedules based on the occupancy information extracted from the images, achieving 95.67 % accuracy in high-density environments.

Sun et al. [22] proposed a three-level fusion framework based on YOLOX for indoor occupancy estimation in a University office space, which achieved a prediction accuracy of up to 99 %. Furthermore, Choi et al. [24] employed a deep learning model based on YOLOv4 for occupancy counting in small and medium-sized offices, demonstrating high performance (root mean square error (RMSE): 0.883), broad applicability and cost-effectiveness of the method.

Expanding on these advancements, Sun et al. [20] proposed a four-step system combining motion detection and static estimation. This approach filters non-occupied frames, detects entrance and exit events, and uses a Fully Convolutional Head Detector (FCHD). The results are fused using Kalman filtering and Occupancy Frequency Histogram (OFH), achieving 97.8 % accuracy. This fusion method effectively addresses common issues like occlusions and cumulative errors, enhancing occupancy estimation in diverse environments. These studies [25,19] have shown that with accurate real-time occupancy data in building management systems, heating, ventilation, and air conditioning (HVAC) systems can be optimised to match actual occupancy patterns, which leads to improved energy efficiency and significant energy savings. For instance, the study [22] reduced the total energy consumption of fan coil units by 18.43 %, 8.71 %, and 18.97 % in different cities by using the occupancy estimation framework.

Some studies have demonstrated the capability to identify not just the presence of occupants but also specific activities and behaviours, such as using equipment or opening windows [27]. Such activities can influence the heat gains or heat loss in buildings and consequently influence the operation of HVAC systems. For instance, Wei et al. [25] utilised Faster R-CNN models to detect the number of occupants and their specific activities, such as walking, sitting, and standing, achieving an accuracy of up to 88.5 % for activity recognition. However, the accuracy of detecting specific activities is generally lower compared to simply detecting the number of occupants, which can achieve higher accuracy, as evidenced by the 98.9 % accuracy for occupancy counting in the same study.

In another study, Wei et al. [26] introduced a real-time occupancy and equipment usage detection approach using Faster R-CNN for demand-driven controls. The approach achieved 93.60 % for occupancy activity detection but lower accuracy for equipment detection

Table 1
Examples of research works on occupancy detection and prediction using the vision-based method.

Ref.	Year	Data collection	Test building	Participants number	Result	Algorithm
[17]	2017	Surveillance camera	Office	12	The number of occupants	CNN, SVM, and K-means
[18]	2019	Omnidirectional camera	Office	4	The number of occupants	YOLOv2
[19]	2020	Camera	Office	1	The number and activity of occupant	CNN
[20]	2021	Entrance video and interior camera	Office	11	The number of occupants	GMM, CNN-based FCHD, Kalman filter, OFH
[21]	2021	Internet protocol camera	Office	10	The number of occupants	YOLOv5
[10]	2021	Camera	Classroom	2	The number of occupants	Faster R-CNN with Inception V2
[22]	2022	Cameras	Office	11	Head and occupancy detection	YOLOX, Deepsort, dynamic Bayesian fusion (DBF)
[23]	2022	Cameras in large indoor space	Classroom	More than 100	Head and occupancy detection	YOLOv3
[24]	2022	Internet protocol camera	Office	6	The number of occupants	YOLOv4
[25]	2022	Camera	Classroom	7	Occupancy count and activity profiles	Faster R-CNN with Inception V2
[26]	2022	Camera	Classroom	3	Occupancy count and activity profiles	Faster R-CNN with Inception V2
[16]	2023	Surveillance videos	Classroom	More than 100	Occupancy count and schedules	CNN-based density estimation method and ensemble model

(78.39 %). Occupancy activity detection requires the detection of the entire body of the occupants, which can be more challenging than methods that employ head counting or detection. Similarly, Tien et al. [28] employed Faster R-CNN to detect real-time window conditions, achieving a detection accuracy of 97.29 % in tests conducted in a case study building. Building on this, the study [10] developed a real-time occupancy and window operation detection, which achieved 85.63 % for occupancy activity detection and 92.2 % for window operation detection. These studies highlight the potential for reducing energy loss and optimising HVAC systems by incorporating real-time detection of window operations.

The use of cameras for occupancy detection raises privacy issues, as continuous monitoring can be intrusive. A small survey conducted by Choi et al. [21] suggests that people preferred occupancy counting techniques that automatically extract and delete images without human intervention, rather than methods that use video encryption or blur occupants. Callemein et al. [18] addressed privacy concerns by using a low-resolution omnidirectional camera that maintains privacy while still providing accurate occupancy counts. Using YOLOv2, they combined spatial and temporal image data to improve detection performance even with extremely low-resolution images.

Many of the studies highlighted above were conducted with a limited number of participants in small to medium-sized offices and classrooms, indicating a potential area for further exploration and validation of these vision-based methods in more populated building spaces. A potential limitation in larger scenes is that the increased distance between the camera and occupants results in lower-resolution images [29], making it difficult to accurately identify and count individuals. Furthermore, the complexity of indoor environments, such as open-plan offices and classrooms, further poses a challenge due to the presence of obstacles like furniture and equipment, which can hinder accurate person identification [30].

To address these issues, some studies have explored the use of multiple cameras for counting individuals [21]. For instance, one study [31] introduced a 3D self-organising neural network approach utilising multiple cameras to tackle occlusions and visibility issues common in crowded and cluttered scenes. The use of multiple cameras can provide different viewpoints and help overcome occlusion problems by covering blind spots, but these methods often necessitate calibrated and synchronised cameras, introducing a layer of complexity and computational demand. Dino et al. [23] investigated vision-based methods for estimating the number of occupants using multiple video cameras. Their hybrid approach combined counting people in a scene with incrementally counting those entering or exiting a room. Tested in a large, crowded, and occluded classroom with over 100 occupants, it showed high predictive capacity. However, the study focused on head detection and counting occupants, without considering specific activities or the impact on building energy performance. It also noted high computational costs and significant infrastructure expenses for multiple cameras and continuous internet connections. This underscores the need for more efficient algorithms and comprehensive privacy-preserving solutions. Furthermore, Yang et al. [16] proposed a CNN-ML framework for crowd counting and prediction in high-density public buildings, achieving 95.67 % recognition and 83.12 % all-day prediction accuracy. However, detection accuracy decreases in dense scenes, and the method has high computational costs and requires extensive manual labelling. Future research should optimise algorithm efficiency and reduce manual labelling.

In contrast, another study [18] employed an omnidirectional or wide field-of-view camera mounted in the ceiling, which captures a single 360-degree image without the need for camera repositioning. This method simplifies the setup by using a single camera to cover an area, reducing the complexity and cost associated with multiple cameras. However, it introduces challenges related to image distortion and lower resolution at the edges of the captured image. The study achieved favourable results but required retraining the detector with similar omnidirectional images to account for the distortion effects. While multi-camera systems improve accuracy in complex environments, they pose challenges in cost and complexity. In contrast, single-view camera systems are simpler and more cost-effective. This research will focus on single-view camera systems evaluating their potential for accurate and efficient occupancy detection in building environments.

2.1. Research gaps and novelty

The review of existing literature reveals advancements in vision-based occupancy detection methods, highlighting the effectiveness of deep learning models such as CNN, YOLO, and Faster R-CNN. These studies demonstrate high accuracy in occupancy counting and activity detection, particularly in small to medium-sized environments. Various approaches, including multi-camera systems and omnidirectional cameras, have been explored to address common challenges like occlusion and low-resolution images in larger and more complex settings. Some studies have also focused on detecting specific activities and behaviours, providing insights for optimising building performance.

Despite the progress made, several gaps remain in the current research. Many studies use various types of algorithms, and a comprehensive evaluation of various deep learning models [32] in building environments is yet to be conducted. The lack of a standardised dataset applicable to the building field and the variance in results even when employing the same algorithm underscores a gap in the current body of research [23]. Testing different vision-based deep learning models on a consistent dataset within a building environment would provide valuable comparative insights. Most studies have been conducted in building environments with a limited number of participants, making it necessary to validate these methods in larger, more populated spaces. Additionally, the impact of accurate occupancy detection on building energy performance has not been thoroughly examined.

Various studies have utilised one to several cameras for different applications. Some studies have employed multiple cameras to resolve issues with spatial coverage and detection accuracy. For example, certain studies used cameras to detect occupants in entrances while simultaneously monitoring indoor spaces. However, there is limited analysis comparing the influence of the cameras' positions and the potential of using a single camera to cover the entire space effectively.

As shown in Table 1, CNN and early iterations of YOLO have been extensively utilised in research, yet the latest YOLO algorithms

have not been sufficiently evaluated in a realistic and dynamic building environment. Since the breakthrough success of CNNs in the ImageNet competition in 2012, deep learning has become mainstream for object detection [33]. Object detection algorithms can be divided into two categories: one-stage and two-stage detection. Two-stage algorithms, like R-CNN [34], Fast R-CNN [35], and Faster R-CNN [36], detect objects through a coarse-to-fine process, providing high accuracy but at the cost of speed. These models, while computationally intensive, perform well in high-density and overlapping occupant scenarios. One-stage detection, such as SSD [37] and the YOLO series [38], directly regresses object size, position, and class, offering faster detection speeds with some trade-offs in accuracy. However, with the release of YOLOv5 in 2022, accuracy and detection speed have notably improved [39].

Despite the growing attention towards deep learning models, their effectiveness in real indoor environments has not been sufficiently assessed. The speed of detection, computational requirements, and performance in a real environment should be further investigated and compared with conventional models. For the experiments conducted in this paper, several popular, open-source computer vision models were used: Single Shot Multibox Detector (SSD) MobileNet V2 [40], Faster R–CNN Inception V2 [36,41], YOLOv5-n [42], YOLOv7-w6 [43] and YOLOv8-x [44].

By employing single-view camera systems, this research seeks to balance accuracy and cost-effectiveness, providing a practical solution for real-time occupancy detection. The study will also assess the potential energy savings and improvements in HVAC system efficiency resulting from accurate occupancy data, addressing both technical and practical aspects of implementation.

2.2. Aims and objectives

This study aims to fill these gaps by evaluating the performance of state-of-the-art deep learning models, including SSD, Faster R-CNN, and the latest YOLO series, in a dynamic and realistic building environment using a low-cost camera. The performance of these models will be assessed in terms of speed of detection, computational requirement and capability in complex scenes. This will be achieved by the following objectives.

- Collect and label a dataset of occupants from random images and test it in a dynamic classroom environment.
- Evaluate the performance of various deep learning models (SSD, Faster R-CNN, YOLO series) using this dataset.
- Implement cameras at various angles and positions to evaluate the impact on detection accuracy.
- Compare predicted occupancy data with ground truth measurements in a classroom to examine the differences.
- Assess the impact of accurate occupancy information on energy use and CO_2 concentration through building energy simulations.

Section 3 details the proposed vision-based deep learning method, including the theoretical foundations and practical applications of this approach. It also describes the case study experiments conducted to test the method. Section 4 analyses the results from the experimental work and energy modelling. The article concludes in Section 5, which also explores potential future research directions.

3. Method

This study employs a computer vision and deep learning-based approach aimed at detecting and recognising occupants within the building environment. The results from the detection phase are inputted into building energy simulation, which analyses building

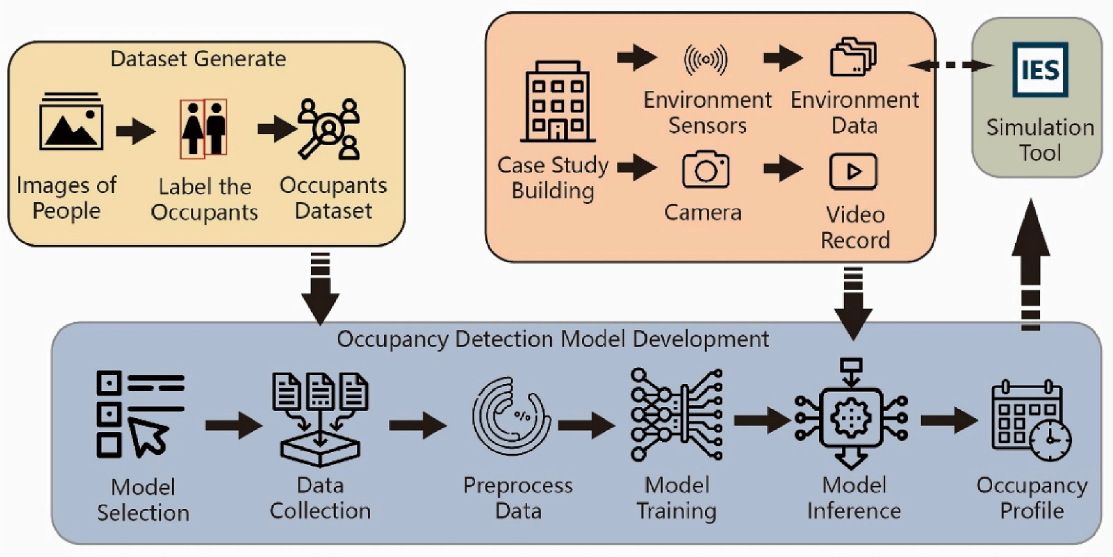

Fig. 1. The workflow of the proposed vision-based deep learning method for occupancy detection.

energy loads and other indices. Fig. 1 shows the general workflow of the vision-based deep learning model employed in this study.

3.1. Dataset generation

This section outlines the process of creating and preparing the dataset used for training the deep learning models. It covers the steps in image collection, annotation, and the splitting of the dataset into training, validation, and testing sets. Detailed information is provided for generating a standardised dataset, which is essential for reproducibility and understanding the foundation of our model training process.

Deep learning object detection requires an image dataset as input. The selection of images should account for various factors to ensure accuracy. For example, variations in indoor lighting, daylight changes, and lighting system operations can all affect image recognition [45]. Therefore, the dataset should include photos from a variety of scenarios, different types of rooms or buildings, varying numbers of people, and multiple angles. These images should be evenly distributed across the test, validation, and training sets to provide comprehensive coverage and improve the model's robustness [46].

Since the occupancy dataset in this study is not based on publicly available datasets, the images were manually gathered, annotated, and then randomly divided into three subsets: training, validation, and testing, with a ratio of 88 %, 8 %, and 4 %, respectively. A total of 377 images were selected, with 330 for the training, 31 for the validation, and 16 for the testing. The small dataset will allow us to test the capability of the models even with limited data, providing valuable insights into their accuracy and computational efficiency. Our training dataset includes images of humans in diverse environments, such as classrooms, offices, and outdoor scenes, to ensure the robustness and generalisation of the deep learning models. This diversity helps improve the model's performance across different scenarios, making it more versatile and reliable for real-time occupancy detection.

The images for training and validation were sourced from publicly accessible image repositories and manually collected by the research team to ensure diversity and robustness. Notably, the images used for training and validation were not collected from the same room as the validation video, which was specifically recorded during a lecture session. This approach helps to avoid overfitting and ensures that the models can generalise well to different environments.

While this study primarily focuses on a lecture room setting, we included images from other environments, to enhance the generalisation capabilities of the models. This broader dataset was used to ensure the models were exposed to varying lighting conditions, room layouts, and occupant behaviours. By incorporating diverse environments during training, we aimed to assess how the models perform beyond a confined lecture room, allowing for better adaptation to real-world scenarios such as open-plan offices, larger buildings, and dynamic occupancy patterns. Although the primary evaluation took place in a lecture room, the training process was designed to prepare the models for broader deployment across different building types.

The collected images were annotated with bounding boxes using the LabelImg tool [47], which generated the necessary label files for the training phase. LabelImg is an open-source graphical tool used to create bounding boxes, crucial for object detection and image classification tasks. It supports outputs in Pascal VOC XML and YOLO text formats, making it compatible with popular machine-learning frameworks. In this project, 377 images were manually annotated by drawing bounding boxes around each occupant, ensuring high-quality training and evaluation of the deep learning models. The annotation process included loading images, drawing bounding boxes, labelling, and saving annotations in the required format.

Each image was annotated with bounding boxes specifying the exact location of humans, using coordinates for x_center, y_center, width, and height. The annotations were saved in.xml files for YOLO input and.txt files that can be easily converted to TF records for TensorFlow models [48]. Preprocessing steps included resizing images to 640x640 pixels and applying auto-orientation to ensure compatibility with the models, preventing memory leaks, poor performance, and imprecise results. Details of the bounding box creation and image preprocessing are listed in Table 2.

To address the potential risk of overfitting, we applied several strategies. Transfer learning was utilised by leveraging pre-trained models based on larger datasets such as COCO (Common Objects in Context). This allowed the models to benefit from pre-learned feature representations, which helped improve their ability to generalise effectively when fine-tuned on our smaller dataset. Transfer learning reduced the training time and prevented overfitting, particularly in the earlier stages of model training.

Furthermore, we employed data augmentation techniques during training to artificially expand the dataset and improve the

Table 2
Dataset creation and preprocessing steps for indoor occupancy detection.

	Tool	Process	Description
Dataset creation	Google Images	Image collection	377 images were collected from Google Images
		Image selection	Images were carefully selected to ensure diversity in postures and arrangements.
Bounding box creation	LabelImg	Loading images	Images were loaded into LabelImg for annotation
		Drawing bounding boxes	Bounding boxes were manually drawn around each occupant in the images using the tool's intuitive interface.
		Assigning labels	Each bounding box was labelled to indicate the number of occupants.
		Saving annotations	The created picture annotations were stored as.xml files, which served as YOLO's input, and.txt files that can be easily converted to TF records for TensorFlow models.
Image preprocessing	Data augmentation	Resizing	All images were resized to a uniform resolution of 640x640 pixels.
		Auto orient	Ensured images are correctly oriented to prevent memory leaks, poor performance, and imprecise results.

models' generalisation ability. These augmentation methods included random rotations, flips, translations, scaling, and brightness adjustments, which introduced variability in the training data. This ensured that the models were exposed to a wider range of visual conditions, enhancing their robustness and reducing the risk of overfitting.

Fig. 2 illustrates the variety of images collected and the manual labelling process used to identify the unique regions of interest in each image. The number of labels applied to each image was determined by its content. The dataset (https://universe.roboflow.com/wuxia-w5dzu/people_small) has been uploaded and is available on Roboflow, a web-based application for object detection datasets [49].

3.2. Deep learning model training and testing

Given that the experiment is conducted in a real environment characterised by high-density, highly variable, and overlapping occupants, YOLO, Faster R-CNN and SSD have been selected for real-time occupancy detection. SSD MobileNet V2 is recognised for its high speed and reasonable accuracy, attributed to its lightweight nature which facilitates quick inference times, thus making it suitable for real-time detection in high-density spaces [40]. YOLOv5 [21] has been employed by many researchers while YOLOv7 [43] and YOLOv8 [44] are the latest version and have shown the best performance to date in terms of accuracy and speed. The YOLO series outpaces Faster R-CNN in speed and has shown improved accuracy over earlier YOLO versions, rendering it well-suited for dynamic, high-density spaces. Moreover, it features enhancements geared towards handling smaller objects, which can be advantageous in situations with overlapping occupants [42].

The SSD and Faster R-CNN models were trained using Tensorflow Object Detection on an NVIDIA GeForce GTX 1080 GPU (2560 CUDA cores, 1607 MHz graphics clock, 320 GB/s memory bandwidth, 8 GB), while the YOLO models were trained with Pytorch in Google Colab [50], which provides free access to NVIDIA T4 Tensor Core GPU (2560 CUDA cores, 1590 MHz graphics clock, 320 GB/s memory bandwidth, 16 GB).

The decision to use two different GPUs is due to the availability of the GPU on Google Collab. Maintaining a valid and insightful comparison across models, despite using varied hardware, is crucial. Both GPUs have the same CUDA core count and nearly identical clock speeds, which are critical for training speed and computational capacity. They also share a 320 GB/s memory bandwidth, ensuring aligned performance. The primary distinction is the NVIDIA T4's larger memory compared to the GTX 1080, allowing for potentially larger batch sizes or more complex models. Despite this, the comparative analysis remains valid since the core specifications affecting training and inference performance are closely aligned.

All models were trained on the same dataset to ensure consistency. The model loss curves, shown in Fig. 3, illustrate the training process and provide insights into whether the model is underfitting or overfitting. Training stopped either when no further improvement was observed or when the loss consistently fell below a certain threshold. SSD and Faster R-CNN required more than 40,000 epochs to complete training, while the YOLO models took less than 300 epochs due to their different architectures. The training speed and mAP of the different models are summarised in Table 3.

We measured the Mean Average Precision (mAP) at an Intersection over Union (IOU) threshold of 0.5 for each model using our dataset. Additionally, we evaluated the mean inference time (MIT) per image, measured in milliseconds (ms), to compare the performance of the different models. The best results for each metric are highlighted in Table 3. Faster R-CNN outperformed SSD, achieving a mAP of 0.83, ranking as the second-best in terms of mAP among all models. However, as a two-stage detection model, Faster R-CNN's detection speed is slower, resulting in delays that make real-time detection challenging [36]. YOLOv8x achieved the highest mAP among all models in 0.43 h, albeit with a slower inference time. Conversely, YOLOv8n emerged as the fastest model, completing the training process in 0.32 h with a mAP of 0.82, and featuring the shortest inference time among all models. Given these results, YOLOv8n is selected for a detailed evaluation, which involves validation of the method with all four cameras. The detailed results will be discussed in the following sections.

3.3. Case study lecture room, testing and BES modelling

This section describes the setup of the case study room, including the installation of cameras and environmental sensors, the layout, and the conditions under which the experiments were conducted.

It provides context for the practical application and testing of the trained models in a real-world environment, demonstrating the feasibility and effectiveness of our approach.

For the implementation of the proposed vision-based deep learning method, lecture room B5 within the Marmont Centre in the University Park Campus, University of Nottingham, UK (Fig. 4) was selected. The lecture room, located on the first floor of the building, is used by students in the Architecture and Built Environment department for both lectures and tutorial sessions during weekdays. It is also available to students outside of lecture and tutorial periods and on weekends. The room has a capacity of 48 seats and 96.9 m^2 of floor space, measuring 12.75 m $\times$ 7.6 m, with a ceiling height of 2.5 m. Detailed information about the case study room is provided in Table 4.

Four Logitech C920 cameras were installed, one in each corner of the room, capable of recording full-HD 1080p video at 30 frames per second (fps) with a 78-degree field of view. Additionally, two Awair Element environmental sensors were placed both outside and inside the room to monitor temperature, relative humidity, carbon dioxide levels, volatile organic compounds (VOCs), and fine particulate matter ($PM_{2.5}$) [51]. All environmental sensors were set to record data continuously on December 2nd, 2022. The layout and location of the sensors are shown in Fig. 4, and detailed information about the sensors used is listed in Table 5.

The experiment was conducted during a single lecture session, capturing various occupancy conditions: the initial period when

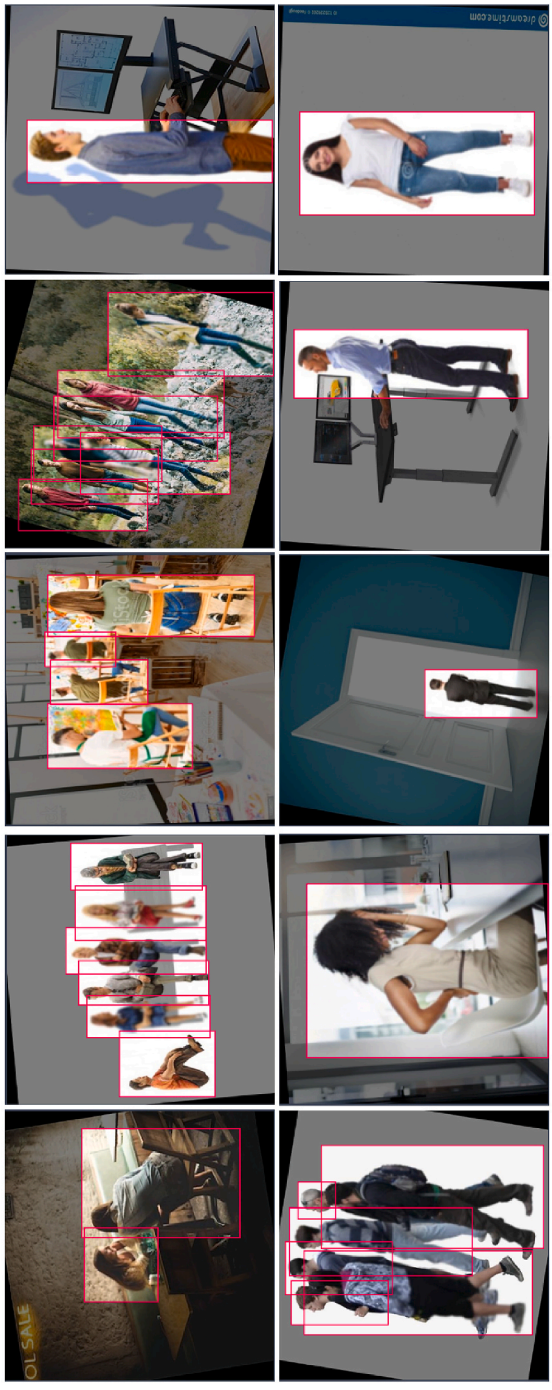

Fig. 2. Example images from the training dataset, showcasing humans in various environments including classrooms, offices, and outdoor scenes. The diversity of the dataset ensures the generalisation of the deep learning models for occupancy detection in different settings.

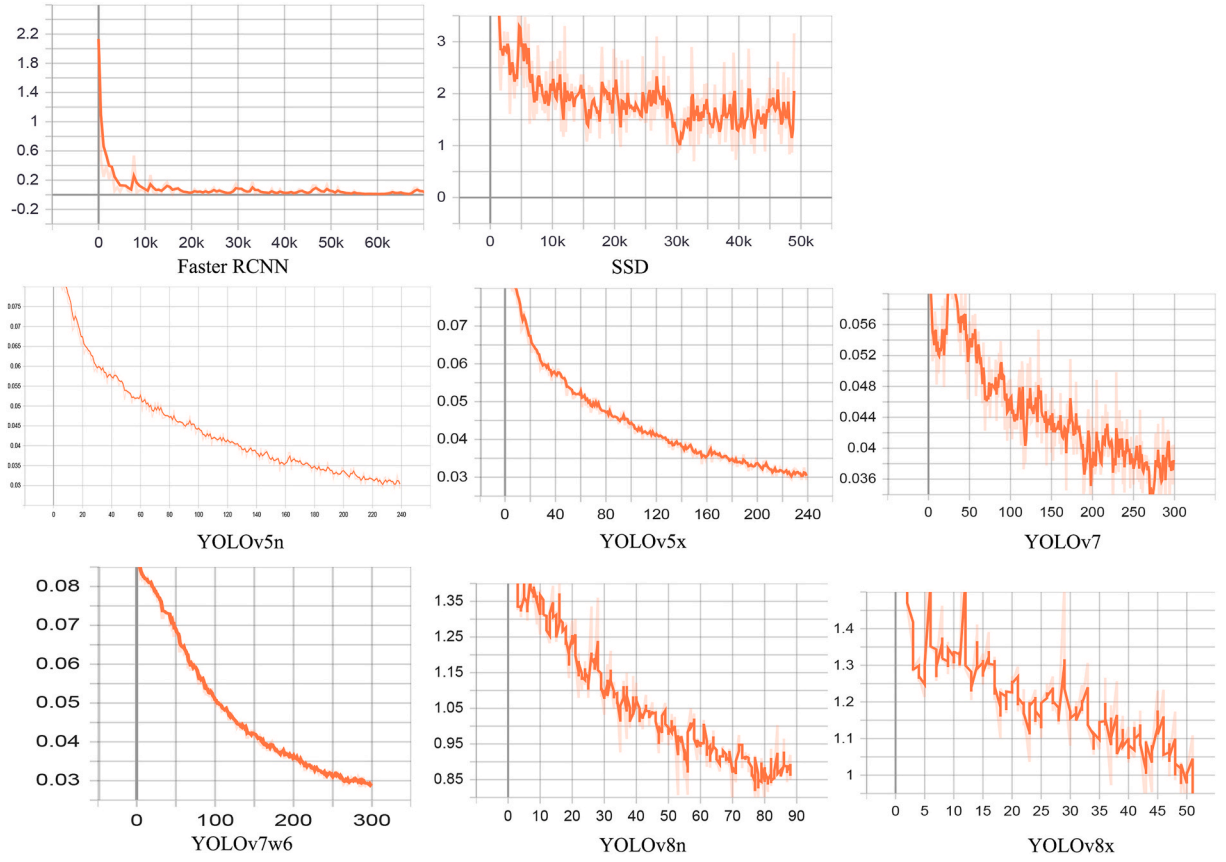

Fig. 3. The training loss curves of Faster R-CNN, SSD, YOLOv5n, YOLOv5x, YOLOv7, YOLOv7w6, YOLOv8n, and YOLOv8x in this study.

Table 3
Comparison of different objection detection models' training performance in this study. The best results for each category are highlighted in bold.

Model	Year	Platform	Training time (hours)	Epochs	mAP^{50} (%)	MIT (ms)	GPU
SSD MobileNetV2 [40]	2020	Tensorflow1.14	8.69	48712	0.22	42	NVIDIA GTX 1080
Faster R-CNN InceptionV2 [36]	2019	Tensorflow1.14	2.9	41901	0.83	79	NVIDIA GTX 1080
YOLOv5n [42]	2020	Pytorch1.7	0.39	240	0.64	28.0	Tesla T4
YOLOv5x [42]	2020	Pytorch1.7	0.42	240	0.63	27.6	Tesla T4
YOLOv7 [43]	2022	Pytorch1.12	1.75	300	0.68	57.5	Tesla T4
YOLOv7w6 [43]	2022	Pytorch1.12	2.03	300	0.76	47.5	Tesla T4
YOLOv8n [44]	2023	Pytorch2.0	**0.32**	88	0.82	**16.4**	Tesla T4
YOLOv8x [44]	2023	Pytorch2.0	0.43	51	**0.87**	292.1	Tesla T4

participants were entering the lecture room, resulting in barely occupied conditions; during the lecture when the room was occupied; and the period when participants were leaving the lecture room, resulting in barely occupied conditions again. By evaluating the models under these different occupancy levels, we were able to assess their performance in adapting to changing conditions within a single session. During the test, the lecture room was mainly occupied from 15:30 to 16:10, with no other activities scheduled for the day. To evaluate the performance of the trained occupancy detection models in a real-world setting, video recording in the room was carried out from 15:15 to 16:20, focusing particularly on the lecture period with a maximum of 25 attendees. Additional environmental factors, such as relative humidity, temperature, and CO_2 levels, were monitored throughout the day to capture the conditions during both occupied and unoccupied periods. All participants were students at the University of Nottingham and were informed about the experiment; they consented to the usage of the footage for this study. The detailed experiment workflow is shown in Fig. 5.

This study initially tests all deep learning models using the recorded video from Camera 1. Subsequent experiments will employ the algorithm demonstrating the best performance, utilising videos from the remaining cameras to assess the influence of different viewing angles. Detailed results from these experiments will be discussed in the following section.

The performance of vision-based models can be significantly affected by various factors, particularly when deployed in complex indoor environments. Although the video footage represents a full lecture scenario, it may not capture all possible occupancy patterns. Factors such as changing lighting conditions, shadows, and the presence of other objects may lead to inaccurate results. Additionally,

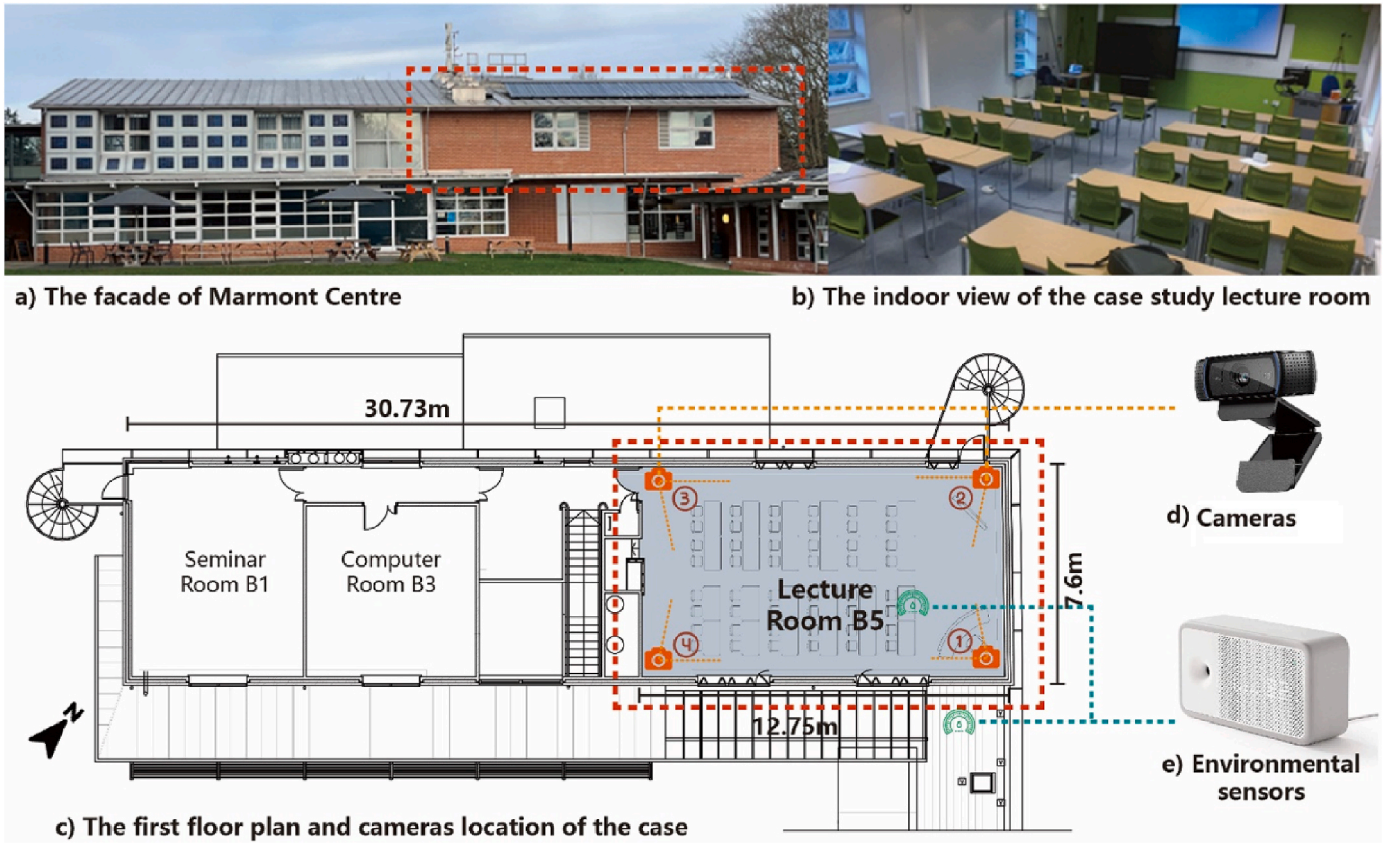

Fig. 4. (a) The Marmont Centre at the University of Nottingham, UK. (b) The indoor view of the case study lecture room. (c) The floor plan and installed sensors layout of the case study building. (d) The picture of the camera in this test. (e) The Awair Element environmental sensors for indoor and outdoor.

Table 4
Information on the case study lecture room and occupancy profiles.

Location	Nottingham, UK
Room area	96.9 m^2
Room dimensions	12.75 m × 7.6 m × 2.5 m
Seats	48
Heating setpoint	21 °C
Occupancy schedule (base case)	08:00–18:00
Ground truth occupancy profile	The observed occupants' number
Occupancy detection profile	Profile generated from vision-based occupancy detection

Table 5
Environmental sensors and camera used in the case study experiment.

Measurement parameters	Sensor	Range	Resolution	Accuracy	Number and location
Air temperature Relative humidity CO_2 concentration	Awair Element environmental sensors	0–90 °C 0%–100 % 400–5000 ppm	0.015 °C 0.01 % 1ppm	±0.2 °C ±2 % 75 ppm or 10 %	2, one inside and one outside
Camera	Logitech C920	78-degree field of view	Full-HD 1080p video 30 frames per second	–	4 in each corner

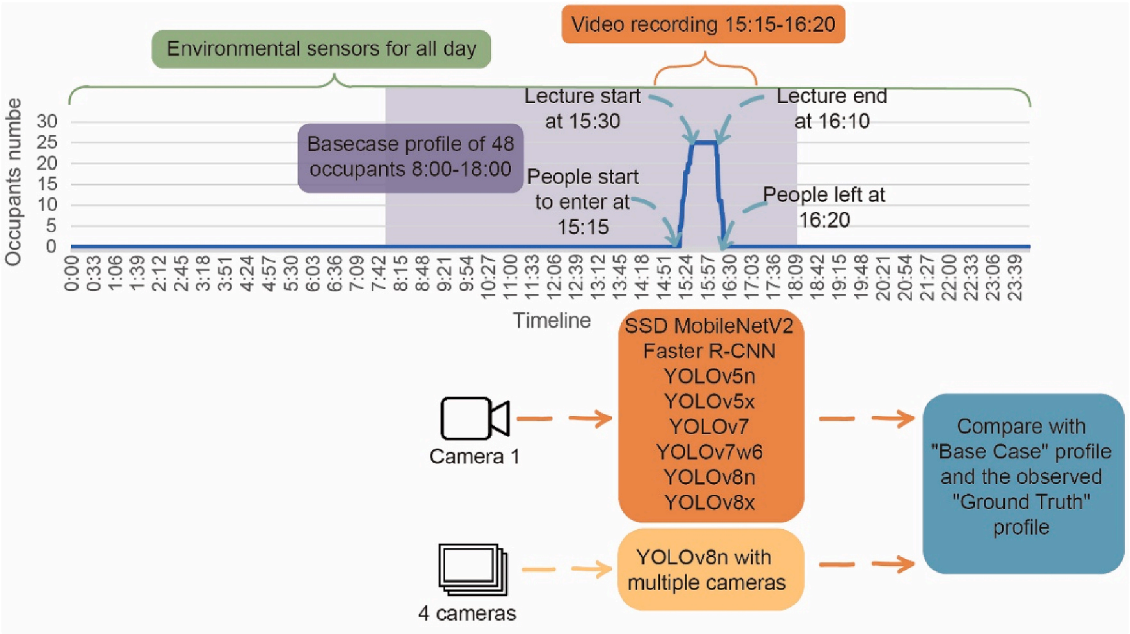

Fig. 5. The workflow for the different cameras employed in the case study experiment on Dec 2nd, 2022.

occupants situated in corners might be easily missed due to the camera's resolution. While the use of low-cost cameras might constrain accuracy in certain scenarios, it enables cost-effective implementation, making it feasible to deploy these systems without heavily investing in new infrastructure. This approach allows for more widespread adoption of occupancy detection technologies within budget constraints.

The deep learning models were tested from four different camera locations, positioned in the corners of the case study room to evaluate the impact of camera angle and location on detection performance. Each camera was tested individually to assess how variations in placement affect occupant detection. This approach allowed us to simulate how different vantage points influence the model's ability to detect occupants in a real-world environment. The case study room provided an ideal setup to evaluate detection accuracy in a moderately sized space. However, we recognise that for larger or more complex rooms, additional cameras or different placement strategies would be required to ensure full coverage, especially in areas that may be obscured or located far from the camera's field of view.

IES VE, a Building Energy Simulation (BES) tool, was used to model the building [52] to evaluate the potential of the proposed

approach and to assess the impact of accurate occupancy information on building energy loads and CO_2 concentration predictions. The building is equipped with a central heating system and has operable windows for natural ventilation. During hours when the building is occupied (08:00–18:00), the heating system was set to maintain an indoor temperature of 21 °C [53]. In this case study, operational hours from 08:00–18:00 on working days [54] were assumed for the base case occupancy profile (fixed schedule). For the simulation, weather data from Nottingham, UK was used. The U-values for the wall, roof, ground, window, and door are detailed in Table 6.

4. Results and discussion

The following section presents the results, analysis, and evaluation of the model detection performance and the impact of the suggested approach on building energy predictions. The trained models are applied in the case study lecture room to evaluate their performance in a real-world environment.

4.1. Comparison of state-of-the-art deep learning models in occupancy detection

Most models except SSD demonstrated the capability to identify, classify, and locate occupants within the lecture room. However, due to varying frame rates and inference speeds among different models, synchronisation across all videos was not achieved. For instance, Faster R-CNN does not facilitate real-time object detection, thereby exhibiting delays, whereas the YOLO series operates in real-time without delays.

Video 1 illustrates the first 15 s of the inference videos from all models, showcasing their performance as individuals began entering the room. As seen in Fig. 6, while the number of occupants was limited, most models captured all individuals present in the video, except for SSD, which only detected one occupant near the camera— these findings align with its low mAP score as shown in Table 3. Occasionally, the deep learning models generated false-positive results, likely due to the presence of objects with patterns resembling the target [55]. For example, YOLOv5n and YOLOv5x falsely identified two objects on the left wall, YOLOv7 positioned a bounding box on the right wall, and YOLOv7w6 misclassified the overhead projector as occupants.

These false positives typically had lower confidence scores, often falling below 0.3, indicating a lack of certainty in those detections and suggesting a potential avenue to enhance model accuracy.

Fig. 7 shows the scene at 15:45 when all students have settled into their seats and exhibit minimal movement, providing a clearer perspective on model performance in a relatively static scenario. **Video 2** compares the detection performance from 15:45:00 to 15:45:15. The SSD model's performance remains suboptimal, detecting only one occupant near the camera. While other models successfully capture the occupants, "flickering" bounding boxes are observed across all models. This flickering, which occurs even in the absence of movement, is a common challenge in object detection algorithms applied to videos. It often arises due to changes in object position, lighting variations, and the underlying algorithmic architecture [56].

Video 1 The inference videos from 15:15:00 to 15:15:15 compare the detection performance of the models when people were entering the room.

Supplementary data related to this article can be found online at https://doi.org/10.1016/j.jobe.2024.111355

Video 2 The inference videos from 15:45:00 to 15:45:15 compare the detection performance of deep learning models when the participants were mostly sitting.

Supplementary data related to this article can be found online at https://doi.org/10.1016/j.jobe.2024.111355

In this scenario, YOLOv5n and YOLOv5x continue to exhibit two false positives on the left side, while YOLOv7 and YOLOv7w6 misidentify the screen and overhead projector as occupants. In contrast, Faster R-CNN, YOLOv8n, and YOLOv8x demonstrate a more accurate capture of almost all occupants, although they still produce a few false positives.

Video 3 and Fig. 8 show the scene as students exit the lecture room, with most occupants congregating near the door and moving out of the camera's view. The SSD model continues to fail to detect any occupants, a consistent issue observed in previous scenarios. Faster R-CNN occasionally misses an occupant near the door in certain frames, possibly due to low resolution. YOLOv5n and YOLOv5x mistakenly label two boxes in the left corner—a recurring error—while YOLOv7 generates several false positives on the left seats. YOLOv7w6 and YOLOv8x successfully capture all occupants, though one false positive occurs for the monitor, likely due to reflection. After discussing occupancy detection in different lecture room scenarios, a detailed analysis of these models' real-world applicability and precision will be conducted. This leads us to a deeper examination using standard metrics, which are discussed in the following section.

Video 3 The inference videos from 16:09:45 to 16:10:00 compare the detection performance of deep learning models when participants are leaving the room.

Supplementary data related to this article can be found online at https://doi.org/10.1016/j.jobe.2024.111355

Table 6
IES VE modelling construction details including U-values (W/m^2K) and thickness.

	Wall	Roof	Ground floor	Window	Door
U-value (W/m^2K)	0.33	0.22	0.32	2.95	2.30
Thickness (mm)	300	290	230	20	40

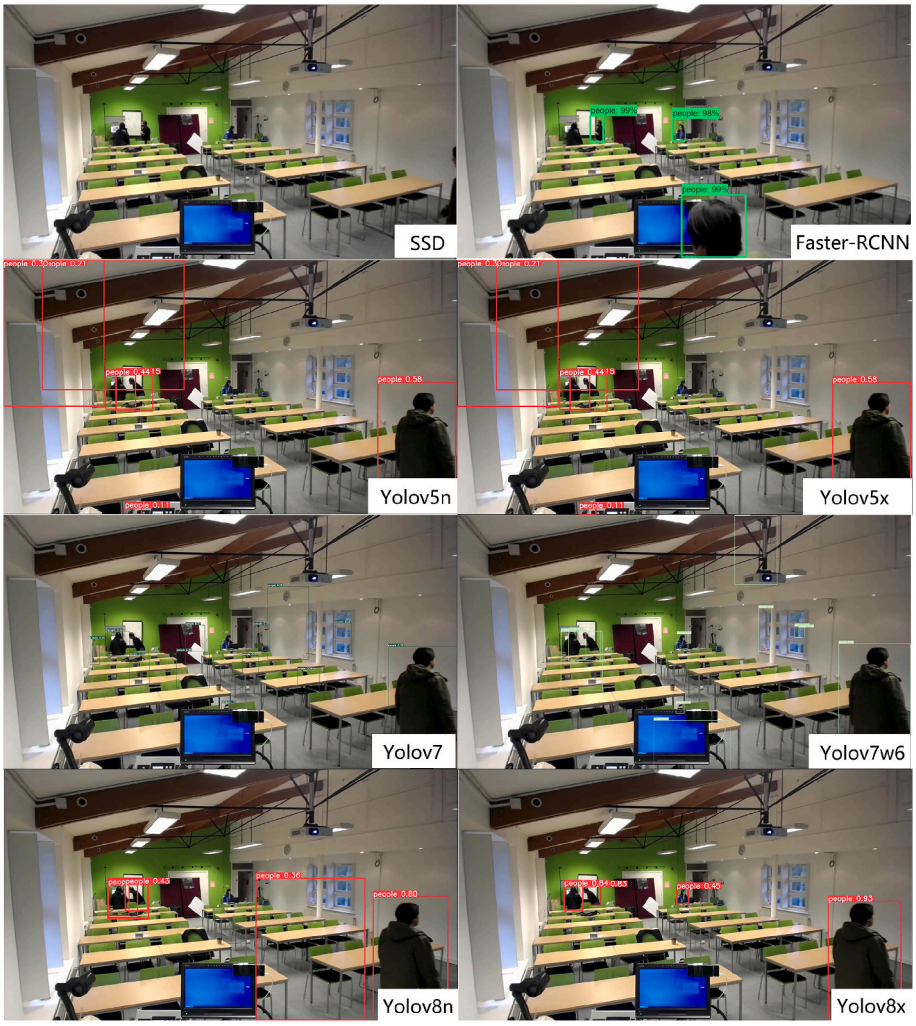

Fig. 6. The frame at 15:15:10 compares the deep learning models' detection of the participants entering the room.

4.2. Evaluation of the model performance in the case study building

Building on the initial findings from Section 4.1, this section conducts a detailed evaluation of the models. We use common metrics such as Accuracy, Recall, Precision, and the F1 Score for a comprehensive analysis of each model's performance. Understanding and applying these metrics is crucial, as they provide a quantitative measure of the models' ability to accurately identify and classify occupants within the building environment. Additionally, we explore the impact of camera positioning and different angles on detection accuracy to gain a more comprehensive understanding of the models' practical performance.

The accuracy metric [55] provides the proportion of correctly classified samples to all samples, offering a broad understanding of the model's performance. The formula for Accuracy is expressed in Eq. (1):

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

Where *TP* (true positive) represents the number of correctly predicted occupants in the video, *FP* (false positive) represents the number of predictions where other objects are regarded as occupants, *FN* (false negative) represents the number of undetected occupants, and *TN* (true negative) represents the number of images without occupants where no prediction is performed. Fig. 9 illustrates examples of True Positive (TP), False Positive (FP), and False Negative (FN) occurrences in a frame captured from the Faster R-CNN, YOLOv5n, YOLOv7 and YOLOv8n model inference videos at 15:41:40.

The recall is crucial because it displays the proportion of true-positive predictions to all occupants found, which is particularly relevant in scenarios where missing an occupant detection is undesirable. The formula for Recall is expressed in Eq. (2):

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

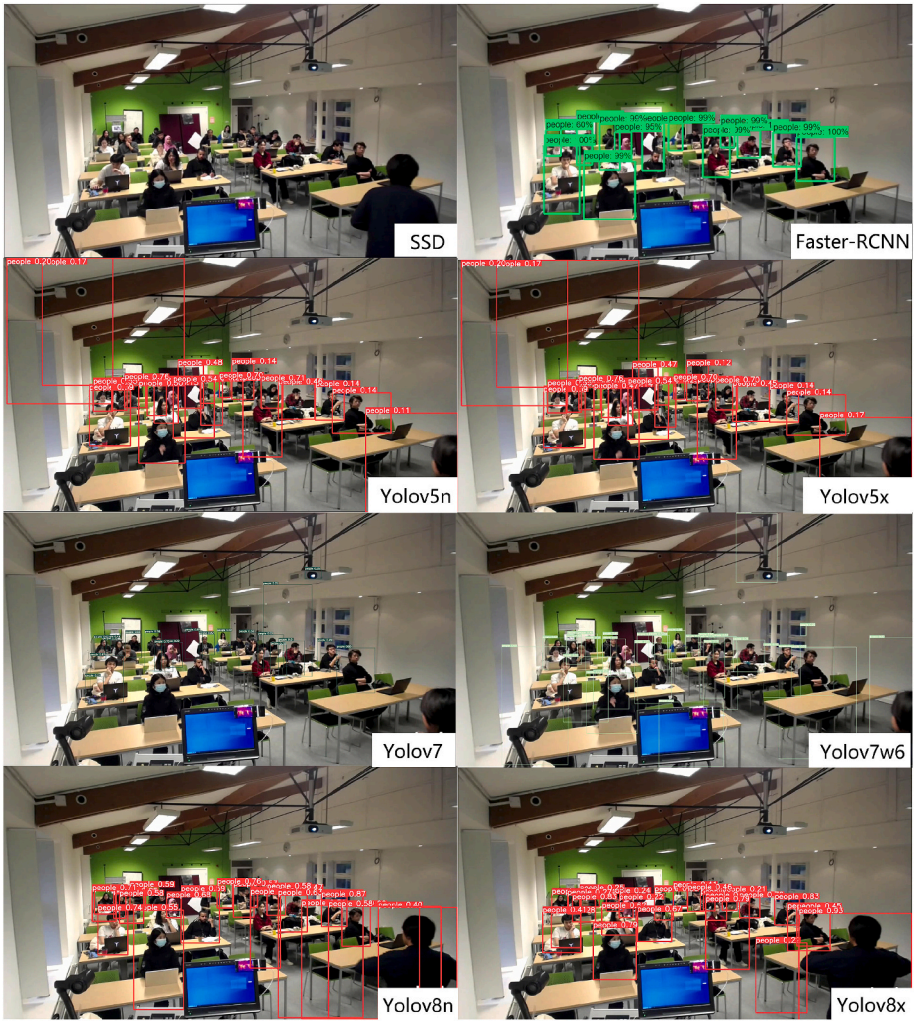

Fig. 7. The frame at 15:45:00 compares the deep learning model detection of the participants in the middle of the lecture.

The ratio of true positive predictions to all positive predictions made is calculated by precision, while the F1 Score provides a balanced measure between precision and recall, which is particularly useful when we want to understand the model's balance between these two metrics. The formulas for Precision and F1 Score are expressed in Eq. (3) and Eq. (4):

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

$$F1\ score = 2^{*}\frac{Precision^{*}Recall}{Precision + Recall} \tag{4}$$

Table 7 compares the performance of eight deep learning models (SSD, Faster R-CNN, YOLOv5n, YOLOv5x, YOLOv7, YOLOv7w6, YOLOv8n, YOLOv8x) across different metrics (Accuracy, Recall, Precision, and F1 score) during three phases: as participants enter the room, during the lecture, and as they exit the room.

During the first 15 min as participants enter the room, YOLOv8x shows the highest accuracy at 0.89, significantly outperforming other models. It also achieves the highest recall at 0.94, indicating its effectiveness in identifying true positive instances. Both Faster R-CNN and YOLOv8x maintain high precision (1.00 and 0.94, respectively), meaning they have the lowest false positive rates. YOLOv8x again stands out with the highest F1 score of 0.94, demonstrating its balanced performance in both precision and recall. In contrast, SSD shows the lowest performance in this phase, with an accuracy of 0.05 and an F1 score of 0.10.

During the lecture, which lasts from 15:30 to 16:10, YOLOv8x and YOLOv8n continue to lead in performance. YOLOv8x achieves an accuracy of 0.75 and a recall of 0.80, while YOLOv8n follows closely with an accuracy of 0.73 and a recall of 0.78. Both models also maintain high precision, with YOLOv8x at 0.93 and YOLOv8n at 0.92. The F1 scores for YOLOv8x and YOLOv8n are 0.86 and 0.84, respectively, indicating their robust performance during the lecture phase. Faster R-CNN, while maintaining perfect precision (1.00),

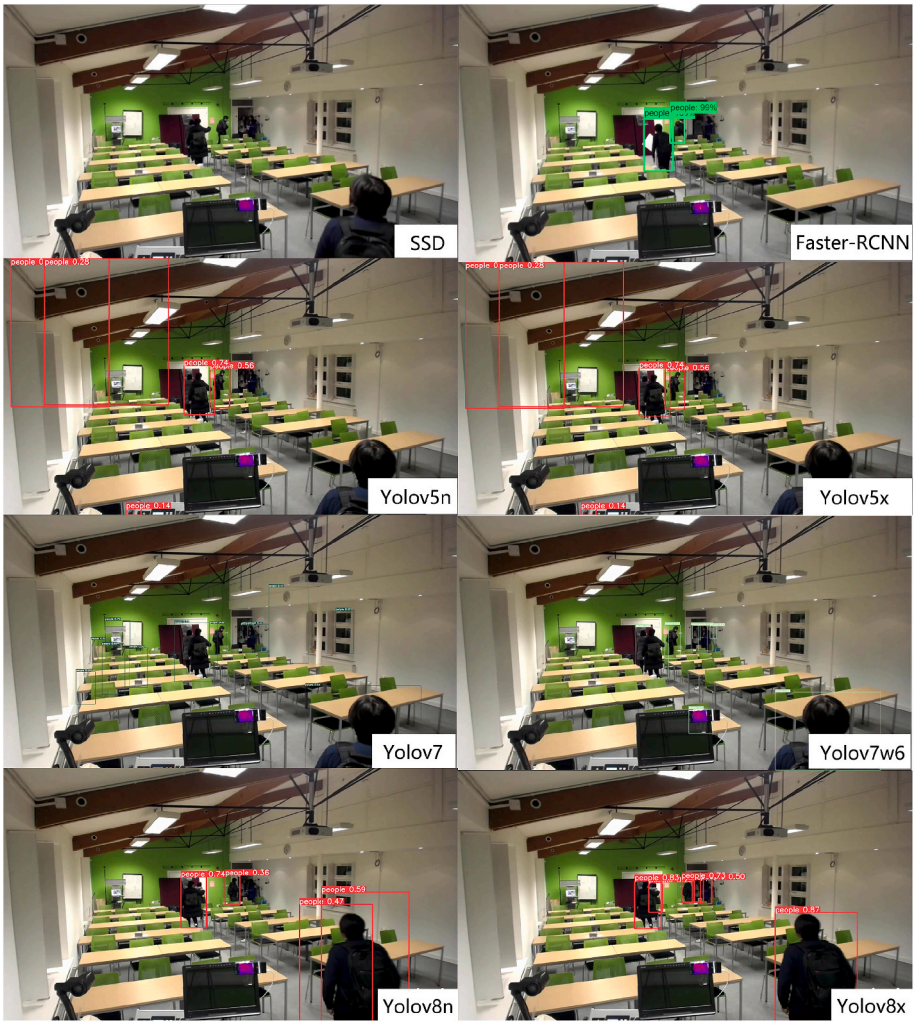

Fig. 8. The frame at 16:09:40 compares the deep learning models' detection of participants leaving the room.

falls behind in recall (0.57) and overall accuracy (0.57).

In the last 10 min as participants leave the room, YOLOv8n achieves the highest accuracy at 0.78 and the highest recall at 0.82. It also maintains a high precision of 0.93 and an F1 score of 0.88, making it the best performer during this phase. YOLOv8x, while slightly behind YOLOv8n, still performs well with an accuracy of 0.63, recall of 0.71, precision of 0.86, and F1 score of 0.77. In this phase, SSD performs the poorest with an accuracy, recall, precision, and F1 score all at 0.00.

When considering the overall performance across all phases, YOLOv8x emerges as the top performer with an accuracy of 0.77, recall of 0.82, precision of 0.93, and an F1 score of 0.87. YOLOv8n also shows strong overall performance with an accuracy of 0.72, recall of 0.78, precision of 0.90, and an F1 score of 0.84. Faster R-CNN excels in precision (1.00) but lags in recall (0.61) and overall accuracy (0.61). SSD, on the other hand, consistently shows the lowest performance metrics.

YOLOv8x consistently outperforms other models across all phases in terms of accuracy, recall, precision, and F1 score. However, YOLOv8n also demonstrates a balanced and strong performance and is particularly noteworthy for its good speed, making it a highly suitable model for real-time applications where both performance and efficiency are crucial. The superior capabilities of these advanced YOLO models in real-time occupancy detection systems provide valuable insights for their application in building and energy management, with YOLOv8n being a promising candidate for future use due to its excellent balance of accuracy and speed.

Fig. 10 compares the accuracy of all models with their training and inference times, where a model with better accuracy occupies a larger circular area. The SSD model's performance is notably poor in most frames, even with the longest training time. YOLOv8x, with an acceptable training time, demonstrated superior performance, achieving the highest overall accuracy of 0.77 and an F1 score of 0.87, although its inference time is lengthy. It can detect most occupants but occasionally misses some at the far end. It is worth noting that Faster R-CNN achieved high precision and confidence scores, as shown in Fig. 10, despite requiring more training time. YOLOv5n, YOLOv5x, and YOLOv8n all exhibited good performance, recognising most occupants, although they sometimes mistakenly identified other objects as occupants initially. YOLOv8n has better accuracy, which is why it was selected for the next stage of testing with

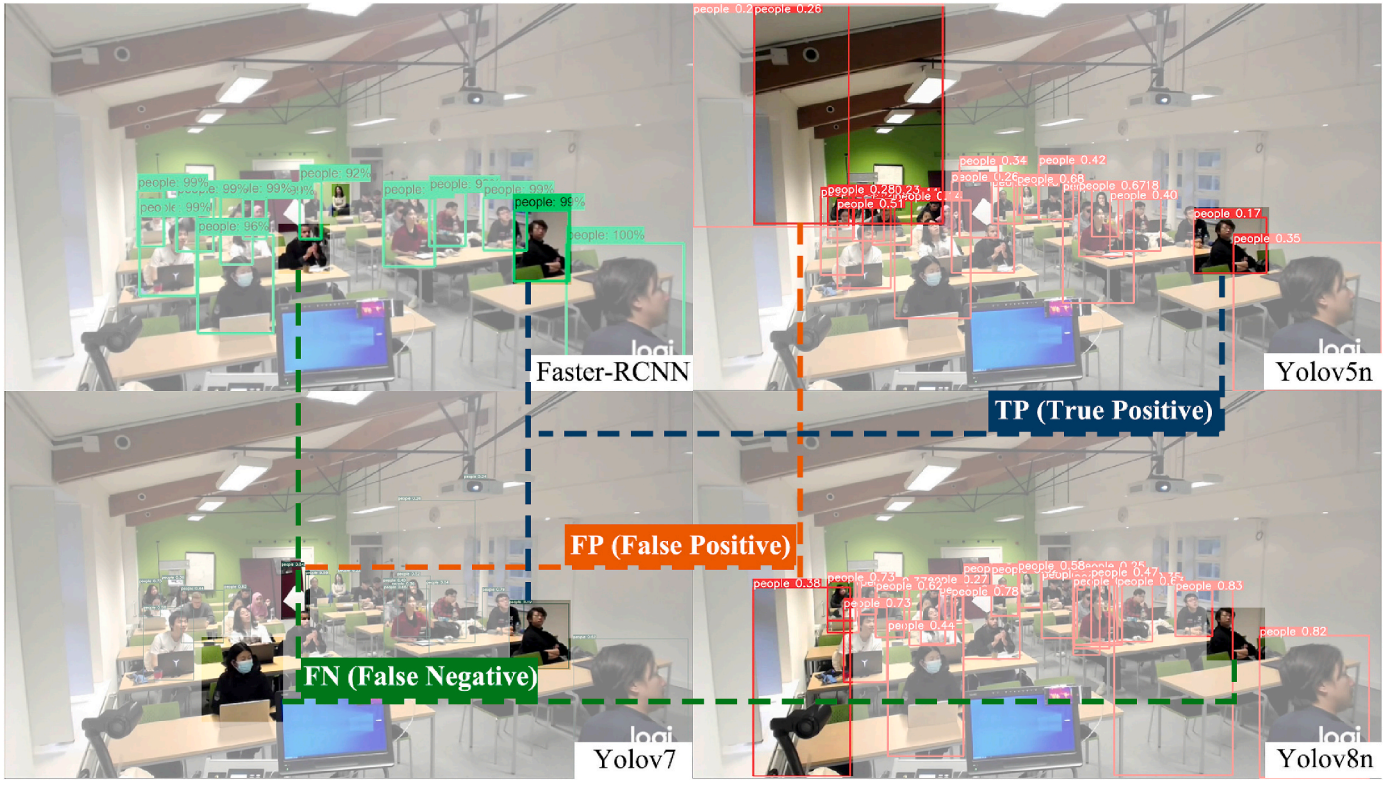

Fig. 9. Examples of TP, FP and FN in a frame taken from the result of Faster R-CNN, YOLOv5n, YOLOv7 and YOLOv8n at 15:35:00.

Table 7
Comparison of the performance of the deep learning models during three distinct phases: as participants enter the room, during the lecture, and as they exit the room.

Model		SSD	Faster R-CNN	YOLOv5n	YOLOv5x	YOLOv7	YOLOv7w6	YOLOv8n	YOLOv8x
Enter (First 15 min)	Accuracy	0.05	0.76	0.51	0.47	0.49	0.50	0.66	0.89
	Recall	0.05	0.76	0.66	0.62	0.68	0.74	0.77	0.94
	Precision	1.00	1.00	0.69	0.66	0.64	0.61	0.82	0.94
	F1 score	0.10	0.86	0.67	0.64	0.66	0.67	0.80	0.94
Lecture (15:30–16:10)	Accuracy	0.03	0.57	0.57	0.46	0.57	0.60	0.73	0.75
	Recall	0.03	0.57	0.65	0.54	0.64	0.73	0.78	0.80
	Precision	1.00	1.00	0.81	0.75	0.84	0.77	0.92	0.93
	F1 score	0.05	0.72	0.72	0.63	0.73	0.75	0.84	0.86
Leave (Last 10mins)	Accuracy	0.00	0.59	0.50	0.41	0.56	0.43	0.78	0.63
	Recall	0.00	0.59	0.65	0.53	0.82	0.67	0.82	0.71
	Precision	0.00	1.00	0.69	0.64	0.64	0.56	0.93	0.86
	F1 score	0.00	0.74	0.67	0.58	0.72	0.61	0.88	0.77
Overall	Accuracy	0.03	0.61	0.55	0.46	0.55	0.56	0.72	**0.77**
	Recall	0.03	0.61	0.66	0.56	0.66	0.73	0.78	**0.82**
	Precision	1.00	**1.00**	0.77	0.72	0.77	0.71	0.90	0.93
	F1 score	0.06	0.75	0.71	0.63	0.71	0.72	0.84	**0.87**

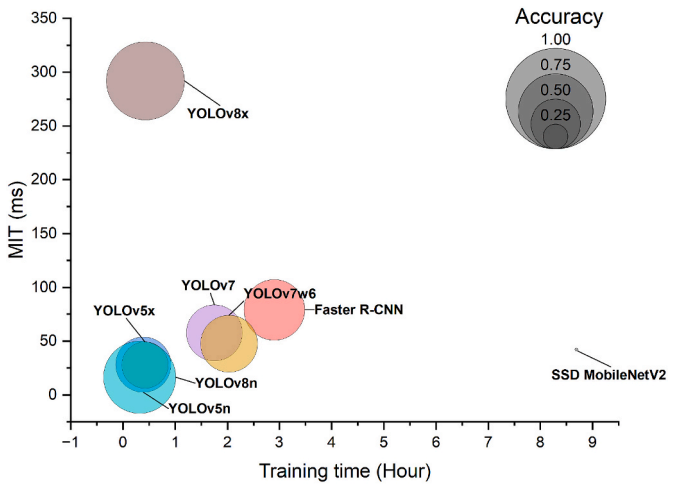

Fig. 10. Performance of deep learning models comparing training time, inference time and accuracy (higher accuracy model occupied bigger circular area).

cameras from various angles. YOLOv7 and YOLOv7w6 provided the same accuracy and F1 score, although their training times and mAP were quite different. These two models occasionally mistook the screen on the table and the overhead projector for occupants.

Fig. 11 presents occupancy profiles predicted by deep learning models compared to the actual number of occupants (ground truth). These profiles will be used as input for subsequent energy simulations. The results from the deep learning models represent the number of detected occupants, including both true positive and false positive detections. Consequently, there may be instances where the results approximate the ground truth but exhibit discrepancies due to missed occupants or false detections. Further work is needed to improve the accuracy, stability, and dependability of the detection models. Based on a comparative analysis of the eight models used in the experiment, YOLOv8n exhibits the least variation in prediction error, yielding the most accurate results when compared to the ground truth.

Moreover, four cameras were strategically positioned at each corner to assess the impact on detection performance, addressing the inevitable occlusions inherent in a singular view. Fig. 12 shows a frame from the inference videos captured by different cameras at 15:45, a time when most students were seated with minimal movement. **Video 4** compares the detection from the four different cameras from 15:45:00 to 15:46:00. The YOLOv8n model was selected for this experiment due to its superior overall performance demonstrated earlier. While some occupants were missed by one camera, they were discernibly captured by others. For instance, camera 2 overlooked several individuals in the right corner, yet they were captured by cameras 1 and 3. This highlights the suboptimal performance of camera 2, which consistently missed occupants in the right corner.

Fig. 13 compares the detection results from each of the four cameras. Detailed model performance is also shown in Table 8. The occupant number detected by cameras 1 and 3 tends to surpass that of cameras 2 and 4. These contrasting outcomes from different cameras underscore the potential uncertainty in detection when relying on a single view. Considering installation costs, enhancing the accuracy of the deep learning model emerges as a more efficient alternative compared to deploying multiple cameras within a room.

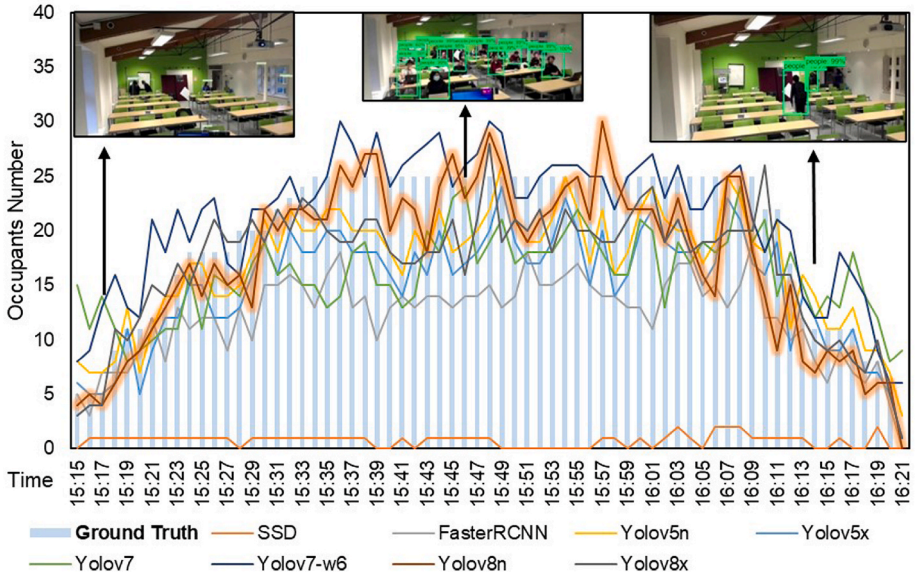

Fig. 11. The occupancy profiles predicted by the deep learning models, compared to the ground truth.

Investing in more advanced algorithms can potentially reduce the need for extensive hardware setups, leading to cost savings and simpler implementations.

Video 4 The inference detection videos comparing 4 different cameras using YOLOv8n from 15:45:00 to 15:46:00.

Supplementary data related to this article can be found online at https://doi.org/10.1016/j.jobe.2024.111355

4.3. Energy and CO_2 simulation results

As illustrated in Fig. 4 and elaborated in Section 3.3, an IES VE model was used to simulate the case study building [27]. The purpose of this simulation was to determine how the accurate occupancy profiles would affect the building's predicted performance, with a focus on energy consumption and CO_2 concentration. Even though the SSD model was wildly off and ill-suited for the needed application, it is nevertheless assessed here to demonstrate its impact on the forecasts. The occupancy profiles generated, encompassing both true and false positive results from the deep learning models, were integrated into the simulation. These profiles were formatted into 5-min intervals to align with the constraints of the building energy software.

A conventional occupancy profile, designating full occupancy (48 individuals) from 8:00 to 18:00 on weekdays, was established as the "Base Case" for comparison. Fig. 14a shows the CO_2 concentration trends predicted using the occupancy profiles based on the eight deep-learning models compared with the recorded CO_2 sensor data during the experiment. The CO_2 concentration trend predicted using detections from the deep learning models aligned reasonably well with the recorded data, showcasing a prompt response to occupancy changes from 15:15 onwards, although it generally overpredicted the measurements. This could be an issue with the IES VE modelling, as the ground truth (actual occupancy) also showed discrepancies. In contrast, the "Base Case" simulation exhibited a pronounced discrepancy with the actual data, and the SSD model profile failed to capture the trend.

A notable lag of approximately 15 min was observed in the recorded CO_2 data from the sensor as occupants began entering at 15:15, with the sensor reflecting changes only around 15:30. This delay is typical for CO_2 sensors, highlighting a temporal limitation in capturing occupancy variations [57]. The vision-based deep learning approach demonstrated a promising capacity to mitigate this lag, signifying its potential to enhance real-time responsiveness in monitoring and controlling building environments.

Fig. 14b displays the internal gains profiles predicted based on the deep learning models, compared to the "Base Case" profile, which assumes full occupancy from 8:00 to 18:00. The actual observed "Ground Truth" profile is highlighted. There is a significant gap between the actual occupancy profiles and the fully occupied profiles commonly used [58]. Most deep learning models in this study can identify and locate people in the case study room, demonstrating the potential for improving model performance and energy demand prediction accuracy.

Fig. 14c displays the heating energy load results in the modelled room from 8:00 to 17:00. The "Base Case" simulation showed a significant discrepancy from the "Ground Truth" simulation, indicating that the conventional fixed profile can be inaccurate in certain scenarios. At the beginning of the lecture during the heating period at 15:15, a substantial amount of heating energy was required to maintain the room at the setpoint temperature of 21 °C for the duration of the occupancy period. This demand rapidly decreased as occupants entered the room and generated internal heat gains. Compared to the actual profile ("Ground Truth"), the YOLOv8n and YOLOv8x models achieved the most accurate heating energy load predictions, while SSD showed the worst performance.

Fig. 15 illustrates the predicted heating energy consumption in the case study room on the test day, comparing different models, the "Ground Truth" and the "Base Case" profiles. The results reveal a 13.45 % discrepancy between the conventional "Base Case" profile

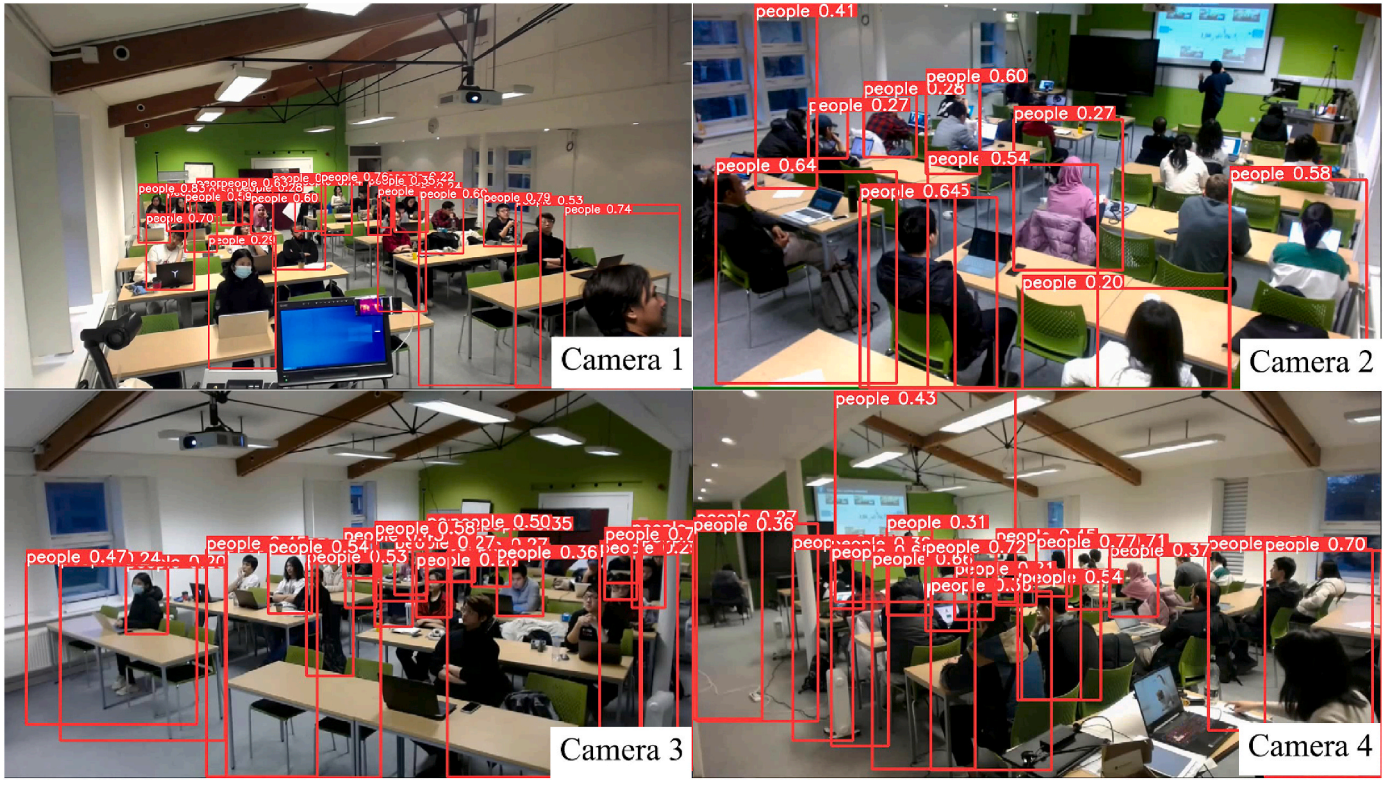

Fig. 12. A frame taken from the YOLOv8n inference detection videos comparing 4 different cameras at 15:45:00.

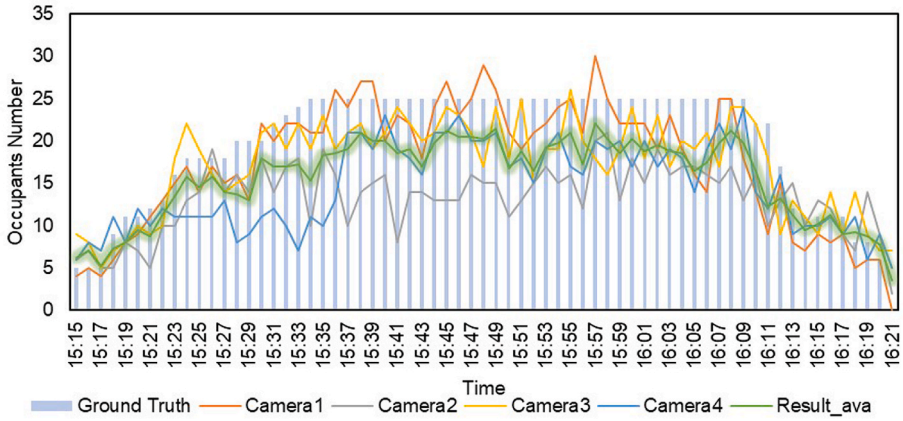

Fig. 13. Occupancy profiles predicted by the deep learning models based on the different detection camera locations using YOLOv8n compared to ground truth.

Table 8
The model performance across all four cameras.

Metric	Camera 1	Camera 2	Camera 3	Camera 4
Accuracy	0.72	0.51	0.69	0.65
Recall	0.78	0.71	0.90	0.82
Precision	0.90	0.64	0.74	0.76
F1 score	0.84	0.67	0.81	0.79

and the “Ground Truth” heating energy consumption. In contrast, the deep learning models show a narrower variation, with deviations from the "Ground Truth" results reaching up to 6.72 %, except for the SSD model, which barely detects any occupants. The conventional fixed profile sets the heater on continuously, even when there are no occupants, leading to higher energy consumption. The data highlights the potential of deep learning models to accurately capture occupancy changes, improving energy consumption predictions compared to traditional methods. These findings emphasise the shift from static to real-time occupancy detection models.

4.4. Potential and limitations of deep learning-based real-time occupancy detection

Generally, the occupancy profiles generated using the deep learning models showed good performance, particularly when compared to the conventional fixed occupancy profile regarding energy and CO_2 predictions. These models were adept at quickly capturing occupancy variations, faster than the CO_2 sensor, indicating promising avenues for future developments. Although the configuration of the model influenced detection performance, other elements like lighting conditions and obstructions also played a role. These factors contributed to the observed changes in predictions and detection performance over the detection period. In this context, YOLOv8x emerged as the most proficient in detection performance, providing the most accurate predictions on occupant profiles, while YOLOv8n excelled in speed and maintained good accuracy overall.

The vision-based method does present certain limitations, with obstructions being a notable challenge due to its inherent nature. This study evaluated the impact of the position and angle of the cameras on detection performance. The effectiveness of the vision-based method is constrained by the camera’s field of view, necessitating strategic placement and perspective adjustments. Additionally, camera resolution could potentially hinder model performance, especially in detecting minute movements or small objects. Future efforts should focus on a more rigorous assessment and verification of the detection technique to ensure its robustness and reliability in diverse real-world scenarios.

This study highlights the potential for real-time occupancy detection to enhance HVAC system performance. By dynamically adjusting HVAC based on actual occupancy patterns, building systems can be tailored more precisely to meet demand, reducing wasteful overconsumption during periods of low or no occupancy. For example, ventilation rates can be lowered in sparsely occupied rooms; while heating and cooling are allocated to spaces where people are present. In buildings with large energy footprints, such as universities, shopping malls, or office buildings, even modest improvements in energy efficiency can lead to significant cost savings over time, reinforcing the economic and environmental benefits of integrating real-time occupancy data into building management systems.

In addition to HVAC optimisation, real-time occupancy detection can also be applied to automated lighting systems, where lights are turned off in unoccupied areas. This is especially relevant in large, irregularly occupied spaces such as libraries, lecture halls, and open-plan offices. By automating lighting based on occupancy data, reductions in electricity use can be achieved, particularly in facilities that operate on extended schedules or have a mix of occupancy patterns throughout the day. Building on the promising results

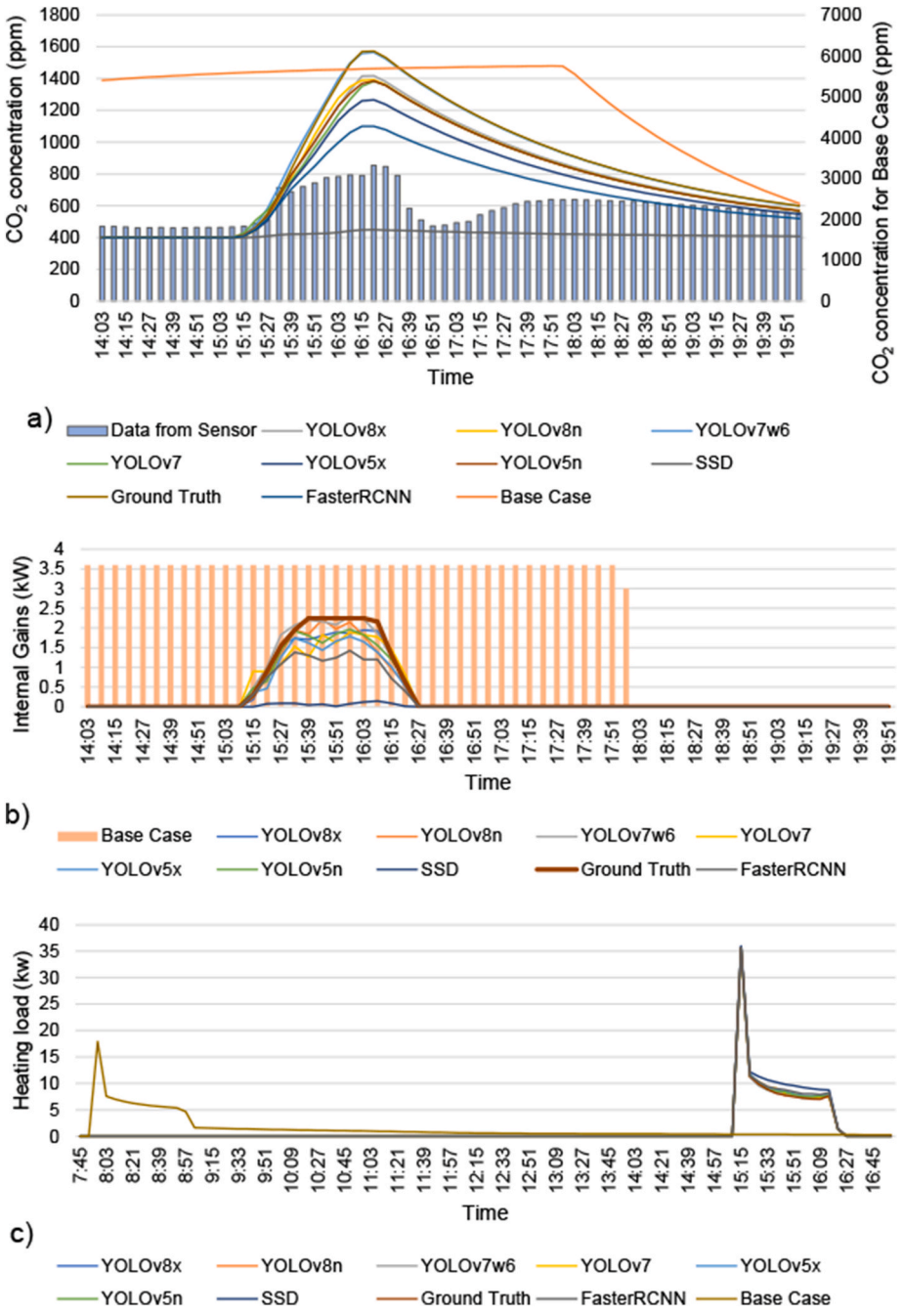

Fig. 14. Predicted vs. recorded data for a) CO_2 concentration, b) internal gains, and c) heating loads.

of this study, future work should focus on expanding the application of the proposed occupancy detection models to open-plan offices, larger buildings, and environments with more diverse occupancy patterns. These settings typically feature more dynamic and dispersed occupancy than lecture rooms, necessitating further evaluation of the models' adaptability to these conditions.

While our vision-based deep learning approach demonstrates the potential for real-time occupancy detection, several limitations must be noted. In this study, we ensured that the cameras covered most of the occupied area, but we observed that the distance of occupants from the camera impacted detection accuracy, with occasional misses for individuals seated at the far end of the room, especially when they appeared smaller in the frame or were partially occluded. Although a multi-camera setup may help mitigate these issues by capturing occupants from multiple angles, this study did not explicitly explore the impact of room size and geometry on model performance. Larger or more complex room layouts with partitions could require additional cameras or alternative placements to maintain detection accuracy. Additionally, the influence of dynamic lighting conditions, such as changing natural or artificial light,

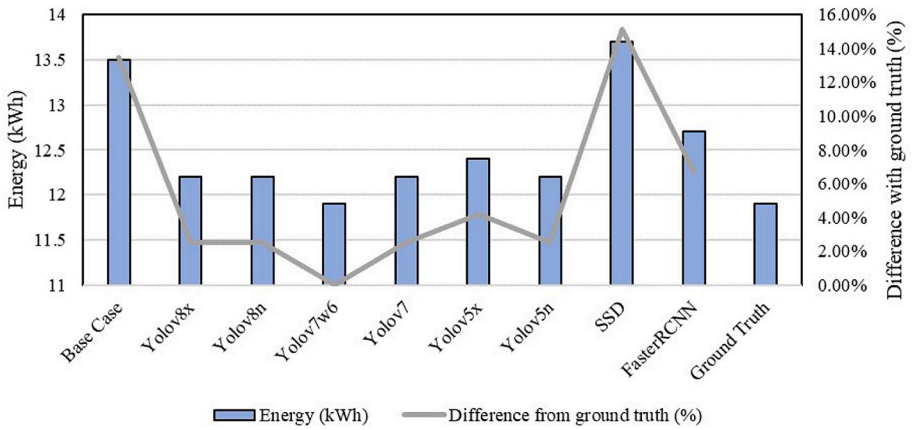

Fig. 15. The predicted heating energy of the case study room on Dec 2nd, 2022, based on the simulation of deep learning model profiles, the "Ground Truth" profile and the "Base case" profile.

on detection accuracy was not explored and may affect accuracy due to shadows or glare.

A comprehensive economic evaluation of deep learning-based occupancy detection systems should be pursued in future work. However, to provide a preliminary economic perspective, we conducted a simplified cost-benefit analysis using a Raspberry Pi 4 unit (£52.8) with a camera module (£11.5), which we plan to deploy in future implementations. The initial investment per setup is approximately £64.3, with minimal installation and maintenance costs, and no licensing fees due to the use of open-source software (TensorFlow, PyTorch). Operational costs are estimated at around £8.58 per year per unit, assuming an average power consumption of 4W at 24.5 pence per kWh.

For this analysis, we assumed daily energy savings of 2 kWh to 3 kWh during the heating period (approximately six months), yielding estimated annual savings between £21.98 and £32.98 at a gas cost of 6.05 pence per kWh for heating. Based on these assumptions, the payback period for each setup is projected to range from approximately 4.95 to 2.63 years.

While these initial estimates are encouraging, they are based on conceptual models and assumptions. Detailed cost analyses and economic evaluations will become more accurate and relevant once a working prototype is developed and tested in real-world settings. This future work will provide empirical data, enabling more comprehensive economic assessments and validating the practical viability of the proposed system. It will also be essential to ensure that the hardware maintains low power consumption in future iterations to maximise these benefits and ensure long-term cost-effectiveness. Low power consumption can be achieved through energy-efficient hardware, optimised algorithms, and power management strategies during low-activity periods.

5. Conclusion and future works

A vision-based deep learning approach for real-time occupancy detection in crowded environments is presented in this work, with a specific focus on a lecture room at a university. We evaluated eight deep learning models, including SSD, Faster R-CNN, YOLOv5n, YOLOv5x, YOLOv7, YOLOv7w6, YOLOv8n, and YOLOv8x, using a self-compiled dataset, during a university lecture experiment. The performance of these models was evaluated in terms of speed of detection, computational requirement, and complexity of the scene. The evaluation revealed varying performance levels among the models. YOLOv8x emerged as the most accurate, with an overall accuracy of 0.77 and an F1 score of 0.87, albeit with a longer inference time. YOLOv8n also demonstrated commendable speed in both training and inference phases while maintaining good accuracy, making it a suitable choice for scenarios prioritising both speed and accuracy. The SSD model, on the other hand, trailed behind significantly, showing a subpar detection ability, particularly struggling to identify occupants unless they were near the camera.

The results from this study demonstrate the effectiveness of the vision-based deep learning models for occupancy detection in a lecture room setting. However, the method is designed to generalise well across different environments, including offices, open-plan spaces, and larger buildings, and this will be tested in future works. The diverse dataset used during training, which included images from offices and outdoor areas, can allow the model to detect accurately in other settings with varying occupancy patterns and room configurations.

In open-plan spaces or larger offices, for instance, the models would need to handle a more dynamic range of occupancy, where individuals may move more freely or gather in specific areas. Given the capability of the models to detect overlapping and occluded individuals, we expect they will maintain similar performance levels in tracking occupancy changes in these environments; however, this will need to be validated in future studies.

Furthermore, the use of data augmentation and transfer learning can help reduce overfitting to specific room types, enhancing the model's ability to generalise across different building environments. Data augmentation introduces variability by altering the training images, while transfer learning allows the model to leverage patterns from larger, diverse datasets, improving its robustness when applied to new spaces with varying layouts and conditions. Future work will focus on validating this approach in larger and more

complex environments. For instance, multi-building campuses, shopping malls, or transportation hubs represent environments with a high potential for occupancy-driven energy management.

Additionally, the study explored the impact of the location and angle of the camera, to assess occlusion challenges often encountered in single-view setups. Specifically, the experiment demonstrated that when one camera missed certain occupants due to obstructions or limited field of view, other cameras positioned at different angles could successfully detect those individuals. For example, individuals missed by camera 2 were detected by cameras 1 and 3, illustrating how a multi-camera setup can compensate for the limitations of a single viewpoint. However, implementing such a multi-camera system would also increase costs and complexity, necessitating more extensive infrastructure and maintenance.

It is important to note that this study did not explore the direct impact of room size on detection performance. The model was tested in a room with a fixed size using cameras placed at various locations. While the results provide insights for moderate-sized spaces, future studies should investigate the scalability of this approach in larger rooms. Specifically, larger spaces could require additional cameras and the use of wide-angle lenses to increase coverage. However, wide-angle lenses may introduce image distortion, which would need to be addressed to maintain detection accuracy. Further testing in diverse room layouts and dimensions is needed to determine the ideal number of cameras, lens types, and their optimal placement for different environments. Additionally, integrating more advanced camera technologies, such as depth sensors or high-resolution cameras, could further improve detection accuracy in larger spaces.

Examining the impact on the predicted energy consumption, our findings revealed a substantial daily heating energy demand difference of approximately 13.45 % when comparing the conventional occupancy profile (Base Case) and the Ground Truth (actual occupancy number). In contrast, deep learning models except the SSD model showed much smaller variations, with a maximum difference of 6.72 % compared to the “Ground Truth”. This highlights the potential of the approach to reduce the gap between actual and predicted energy consumption and improve precise, demand-driven building management systems. Although the deep learning models generally overpredicted the recorded data, the CO_2 concentration trends they predicted aligned closely with the recorded data, unlike the Base Case profile. This alignment demonstrates the potential of the proposed method not only to improve the accuracy and reliability of energy performance predictions but also to respond more quickly to occupancy changes than CO_2 sensors. These findings suggest a promising future for demand-driven building management systems. Future efforts, including assessments of scalability, comparative studies with other technologies, and user feedback, could offer comprehensive insights into the practical deployment and effectiveness of the proposed system in real-world settings. These steps aim to support the development of a more robust and efficient real-time occupancy detection system for improved energy management in buildings.

CRediT authorship contribution statement

Wuxia Zhang: Writing – original draft, Visualization, Validation, Methodology, Data curation, Conceptualization. **John Calautit:** Writing – review & editing, Supervision, Project administration, Conceptualization. **Paige Wenbin Tien:** Writing – review & editing, Software, Investigation. **Yupeng Wu:** Writing – review & editing, Supervision. **Shuangyu Wei:** Writing – review & editing, Software.

Declaration of competing interest

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

Appendix A. Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.jobe.2024.111355.

Data availability

Data will be made available on request.

References

[1] S. D'Oca, T. Hong, J. Langevin, The human dimensions of energy use in buildings: a review, Renew. Sustain. Energy Rev. 81 (2018) 731–742.
[2] T. Kitzberger, J. Kotik, T. Pröll, Energy savings potential of occupancy-based HVAC control in laboratory buildings, Energy Build. 263 (2022) 112031.
[3] C. Wang, K. Pattawi, H. Lee, Energy saving impact of occupancy-driven thermostat for residential buildings, Energy Build. 211 (2020) 109791.
[4] M. Pappalardo, T. Reverdy, Explaining the performance gap in a French energy efficient building: persistent misalignment between building design, space occupancy and operation practices, Energy Res. Social Sci. 70 (2020) 101809.
[5] D. Sheikh Khan, J. Kolarik, C. Anker Hviid, P. Weitzmann, Method for long-term mapping of occupancy patterns in open-plan and single office spaces by using passive-infrared (PIR) sensors mounted below desks, Energy Build. 230 (2021) 110534.
[6] A. Franco, F. Leccese, Measurement of CO2 concentration for occupancy estimation in educational buildings with energy efficiency purposes, J. Build. Eng. 32 (2020) 101714.
[7] N. Li, G. Calis, B. Becerik-Gerber, Measuring and monitoring occupancy with an RFID based system for demand-driven HVAC operations, Autom. ConStruct. 24 (2012) 89–99.

[8] R. Razavi, A. Gharipour, M. Fleury, I.J. Akpan, Occupancy detection of residential buildings using smart meter data: a large-scale study, Energy Build. 183 (2019) 195–208.
[9] N. Alishahi, M.M. Ouf, M. Nik-Bakht, Using WiFi connection counts and camera-based occupancy counts to estimate and predict building occupancy, Energy Build. 257 (2022) 111759.
[10] P.W. Tien, S. Wei, J.K. Calautit, J. Darkwa, C. Wood, Real-time monitoring of occupancy activities and window opening within buildings using an integrated deep learning-based approach for reducing energy demand, Appl. Energy 308 (2022) 118336.
[11] J. Gao, F. Zuo, K. Ozbay, O. Hammami, M.L. Barlas, A new curb lane monitoring and illegal parking impact estimation approach based on queueing theory and computer vision for cameras with low resolution and low frame rate, Transport. Res. Pol. Pract. 162 (2022) 137–154.
[12] R. Tong, D. Xie, M. Tang, Upper body human detection and segmentation in low contrast video, IEEE Trans. Circ. Syst. Video Technol. 23 (2013) 1502–1509.
[13] N. O'Mahony, S. Campbell, A. Carvalho, S. Harapanahalli, G.V. Hernandez, L. Krpalkova, et al., Deep learning vs. Traditional computer vision, in: K. Arai, S. Kapoor (Eds.), Advances in Computer Vision, Springer International Publishing, Cham, 2020, pp. 128–144.
[14] V.A. Sindagi, V.M. Patel, A survey of recent advances in CNN-based single image crowd counting and density estimation, Pattern Recogn. Lett. 107 (2018) 3–16.
[15] Li Y, Zhang X, Chen D. CSRNet: dilated convolutional neural networks for understanding the highly congested scenes. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition2018. p. 1091-1100.
[16] Y. Yang, Y. Yuan, T. Pan, X. Zang, G. Liu, A framework for occupancy prediction based on image information fusion and machine learning, Build. Environ. 207 (2022) 108524.
[17] J. Zou, Q. Zhao, W. Yang, F. Wang, Occupancy detection in the office by analyzing surveillance videos and its application to building energy conservation, Energy Build. 152 (2017) 385–398.
[18] T. Callemein, K.V. Beeck, T. Goedemé, Anyone here? Smart embedded low-resolution omnidirectional video sensor to measure room occupancy. 2019 18th IEEE International Conference on Machine Learning and Applications (ICMLA), 2019, pp. 1993–2000.
[19] P.W. Tien, S. Wei, J.K. Calautit, J. Darkwa, C. Wood, A vision-based deep learning approach for the detection and prediction of occupancy heat emissions for demand-driven control solutions, Energy Build. 226 (2020) 110386.
[20] K. Sun, Q. Zhao, Z. Zhang, X. Hu, Indoor occupancy measurement by the fusion of motion detection and static estimation, Energy Build. 254 (2022) 111593.
[21] H. Choi, C.Y. Um, K. Kang, H. Kim, T. Kim, Application of vision-based occupancy counting method using deep learning and performance analysis, Energy Build. 252 (2021) 111389.
[22] K. Sun, P. Liu, T. Xing, Q. Zhao, X. Wang, A fusion framework for vision-based indoor occupancy estimation, Build. Environ. 225 (2022) 109631.
[23] I. Gursel Dino, E. Kalfaoglu, O.K. Iseri, B. Erdogan, S. Kalkan, A.A. Alatan, Vision-based estimation of the number of occupants using video cameras, Adv. Eng. Inf. 53 (2022) 101662.
[24] H. Choi, J. Lee, Y. Yi, H. Na, K. Kang, T. Kim, Deep vision-based occupancy counting: experimental performance evaluation and implementation of ventilation control, Build. Environ. 223 (2022) 109496.
[25] S. Wei, P.W. Tien, T.W. Chow, Y. Wu, J.K. Calautit, Deep learning and computer vision based occupancy CO2 level prediction for demand-controlled ventilation (DCV), J. Build. Eng. 56 (2022) 104715.
[26] S. Wei, P.W. Tien, Y. Wu, J.K. Calautit, A coupled deep learning-based internal heat gains detection and prediction method for energy-efficient office building operation, J. Build. Eng. 47 (2022) 103778.
[27] P.W. Tien, S. Wei, J.K. Calautit, J. Darkwa, C. Wood, Enhancing the detection performance of a vision-based window opening detector, Clean, Energy Syst 3 (2022) 100038.
[28] P.W. Tien, S. Wei, T. Liu, J. Calautit, J. Darkwa, C. Wood, A deep learning approach towards the detection and recognition of opening of windows for effective management of building ventilation heat losses and reducing space heating demand, Renew. Energy 177 (2021) 603–625.
[29] G. Gao, J. Gao, Q. Liu, Q. Wang, Y. Wang, Cnn-based Density Estimation and Crowd Counting: A Survey, 2020 arXiv preprint arXiv:200312783.
[30] Y. Zhang, D. Zhou, S. Chen, S. Gao, Y. Ma, Single-image crowd counting via multi-column convolutional neural network. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 589–597.
[31] L. Maddalena, A. Petrosino, F. Russo, People counting by learning their appearance in a multi-view camera environment, Pattern Recogn. Lett. 36 (2014) 125–134.
[32] J. Dridi, M. Amayri, N. Bouguila, Transfer learning for estimating occupancy and recognizing activities in smart buildings, Build. Environ. 217 (2022) 109057.
[33] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, Commun. ACM 60 (2017) 84–90.
[34] D. Konstantinidis, V. Argyriou, T. Stathaki, N. Grammalidis, A modular CNN-based building detector for remote sensing images, Comput. Network. 168 (2020) 107034.
[35] Girshick R. Fast r-cnn. Proceedings of the IEEE International Conference on Computer Vision2015. p. 1440-1448.
[36] S. Ren, K. He, R. Girshick, J. Sun, Faster r-cnn: towards real-time object detection with region proposal networks, Adv. Neural Inf. Process. Syst. 28 (2015).
[37] J. Pincott, P.W. Tien, S. Wei, J.K. Calautit, Indoor fire detection utilizing computer vision-based strategies, J. Build. Eng. 61 (2022) 105154.
[38] L. Monti, S. Mirri, C. Prandi, P. Salomoni, Smart sensing supporting energy-efficient buildings: on comparing prototypes for people counting. Proceedings of the 5th EAI International Conference on Smart Objects and Technologies for Social Good, Association for Computing Machinery, Valencia, Spain, 2019, pp. 171–176.
[39] G. Jocher, A. Chaurasia, A. Stoken, J. Borovec, Y. Kwon, J. Fang, et al., ultralytics/yolov5: V6. 1-TensorRT, TensorFlow Edge TPU and OpenVINO Export and Inference, Zenodo, 2022.
[40] Y.-C. Chiu, C.-Y. Tsai, M.-D. Ruan, G.-Y. Shen, T.-T. Lee, Mobilenet-SSDv2: an improved object detection model for embedded systems. 2020 International Conference on System Science and Engineering (ICSSE), IEEE, 2020, pp. 1–5.
[41] O. Karaman, A. Alhudhaif, K. Polat, Development of smart camera systems based on artificial intelligence network for social distance detection to fight against COVID-19, Appl. Soft Comput. 110 (2021) 107610.
[42] G. Jocher, A. Chaurasia, A. Stoken, J. Borovec, Y. Kwon, ultralytics/yolov5: V6. 1-TensorRT TensorFlow edge TPU and OpenVINO export and inference, Zenodo 2 (2022) 2.
[43] C.-Y. Wang, A. Bochkovskiy, H.-Y.M. Liao, YOLOv7: Trainable Bag-Of-Freebies Sets New State-Of-The-Art for Real-Time Object Detectors, 2022 arXiv preprint arXiv:220702696.
[44] Jocher G, Chaurasia, A., & Qiu, J. YOLO by Ultralytics. 8.0.0 Ed2023.
[45] K. Sun, Q. Zhao, J. Zou, A review of building occupancy measurement systems, Energy Build. 216 (2020) 109965.
[46] D. Kang, Z. Ma, A.B. Chan, Beyond counting: comparisons of density maps for crowd analysis tasks—counting, detection, and tracking, IEEE Trans. Circ. Syst. Video Technol. 29 (2019) 1408–1422.
[47] Tzutalin. LabelImg, 2018.
[48] T. Developers, TensorFlow, Zenodo, 2022.
[49] Wuxia, People_small dataset, Roboflow Universe (2022).
[50] E. Bisong, E. Bisong, Google colaboratory. Building Machine Learning and Deep Learning Models on Google Cloud Platform: a Comprehensive Guide for Beginners, 2019, pp. 59–64.
[51] T.Y. Hsu, Q.V. Pham, W.C. Chao, Y.S. Yang, Post-earthquake building safety evaluation using consumer-grade surveillance cameras, Smart Structures and Systems, An International Journal 25 (2020) 531–541.
[52] I.E. Solutions, IES Virtual Environment (IESVE)[Computer software] (2020).
[53] CIBSE, Chartered Institution of Building Services Engineers, CIBSE, 2021.
[54] T. Cibse, Energy benchmarks, The Chartered Institution of Building Ser-vices Engineers (2008).

[55] M. Sokolova, N. Japkowicz, S. Szpakowicz, Beyond accuracy, F-score and ROC: a family of discriminant measures for performance evaluation. AI 2006: Advances in Artificial Intelligence: 19th Australian Joint Conference on Artificial Intelligence, Hobart, Australia, December 4-8, 2006 Proceedings 19, Springer, 2006, pp. 1015–1021.
[56] A. Azulay, Y. Weiss, Why Do Deep Convolutional Networks Generalize So Poorly to Small Image Transformations? arXiv Preprint arXiv:180512177, 2018.
[57] X. Liang, J. Shim, O. Anderton, D. Song, Low-cost data-driven estimation of indoor occupancy based on carbon dioxide (CO2) concentration: a multi-scenario case study, J. Build. Eng. 82 (2024) 108180.
[58] E. Azar, C.C. Menassa, A comprehensive analysis of the impact of occupancy parameters in energy simulation of office buildings, Energy Build. 55 (2012) 841–853.