energies

MDPI

Article

Low-Cost Thermal Camera-Based Counting Occupancy Meter Facilitating Energy Saving in Smart Buildings

Marek Kraft *, Przemysław Aszkowski, Dominik Pieczyński and Michał Fularz

Institute of Robotics and Machine Intelligence, Poznań University of Technology, Piotrowo 3A, 60-965 Poznań, Poland; przemyslaw.aszkowski@student.put.poznan.pl (P.A.); dominik.pieczynski@put.poznan.pl (D.P.); michal.fularz@put.poznan.pl (M.F.)

* Correspondence: marek.kraft@put.poznan.pl

Abstract: Using passive infrared sensors is a well-established technique of presence monitoring. While it can significantly reduce energy consumption, more savings can be made when utilising more modern sensor solutions coupled with machine learning algorithms. This paper proposes an improved method of presence monitoring, which can accurately derive the number of people in the area supervised with a low-cost and low-energy thermal imaging sensor. The method utilises U-Net-like convolutional neural network architecture and has a low parameter count, and therefore can be used in embedded scenarios. Instead of providing simple, binary information, it learns to estimate the occupancy density function with the person count and approximate location, allowing the system to become considerably more flexible. The tests show that the method compares favourably to the state of the art solutions, achieving significantly better results.

Keywords: heating; air conditioning; deep learning; building management systems; thermal imaging

Citation: Kraft, M.; Aszkowski, P.; Pieczyński, D.; Fularz, M. Low-Cost Thermal Camera-Based Counting Occupancy Meter Facilitating Energy Saving in Smart Buildings. *Energies* **2021**, *14*, 4542. https://doi.org/10.3390/en14154542

Communicated by: Umberto Berardi

Received: 14 June 2021
Accepted: 23 July 2021
Published: 27 July 2021

1. Introduction

Residential and industrial buildings are responsible for a substantial portion of overall energy consumption across the world, with a significant share of the energy being used by the heating, ventilation, and air conditioning (HVAC) [1–3]. The consequence of this fact is the drive to minimise energy consumption, as this brings economic and environmental benefits. This is especially important in light of the fact that building energy consumption is one of the main driving factors behind carbon dioxide emissions [4]. Moreover, energy production is associated with the emission of other pollutants. Therefore, limiting the energy consumption directly translates into reduced emissions. The most common means for the reduction in HVAC-related energy costs include careful building design or modernisation (e.g., improving thermal insulation provided by walls and windows), enabling heat exchange, ensuring proper ventilation, using modern, efficient HVAC equipment, and, last but not least, intelligent control.

The introduction of thermostats and feedback control resulted both in energy savings and increased comfort by enabling precise temperature setting. The thermostats oftentimes enable the users to adjust other temperature levels for the time window outside of office hours or at night-time, enabling further savings when energy expenditure is not justified. However, such global settings lack the granularity that is required to control the HVAC devices with an accuracy up to a single floor or even a single room. To deal with this issue, modern intelligent buildings are fitted with multiple temperature sensors and occupancy sensors to provide accurate, room level control and adjustment to the current user activity levels. The advent and fast adoption of the Internet of Things (IoT) has transformed multiple application areas [5,6]. Among others, it has facilitated the introduction of a new generation of building management systems (BMSs) [7,8]. Fully integrated measurements from vast sensor networks can be easily aggregated and used for intelligent control. Access to a variety of sensor data aside from single point measurements and the addition of

forecast data improves comfort and safety and facilitates energy savings. Aggregation and interpretation of multiple streams of data is made possible thanks to the widespread use of machine learning methods [9,10].

Potential savings from using occupancy sensors to control HVAC and lighting systems are estimated to be around 25% [11]. The most common method for monitoring occupancy in smart buildings is the passive infrared (PIR) sensors [12]. However, presence monitoring can be taken even further by providing information on the number of persons in a room or other monitored space and adjusting the operation of the system accordingly. Successful monitoring beyond simple presence check with PIR sensors was reported in the literature [13,14]. The methods rely on careful spatial arrangement of multiple sensors and further processing to extract spatial information, e.g., by triangulation based on individual sensor measurement. Moreover, PIR sensors' readings may be unreliable as the number of persons in the monitored space increases [15]. An overview of techniques enabling fine-grained occupancy measurement is presented in [16]. The main findings are that the manual evaluation (questionnaires, interviews) is time-consuming and unreliable, with monitoring using various electronic sensors being a more appealing alternative in terms of accuracy and associated workload. Simple, inexpensive methods (e.g., infrared barrier mounted at entry/exit points) tend to be less accurate. More advanced methods, such as video-based surveillance, are also reviewed in the paper. The conclusion is that these methods have potential to achieve a very good accuracy, as image-based analyses are naturally suited for presence detection and activity monitoring. For example, Jazizadeh and Jung [17] present a vision-based system for personalized thermal comfort assessment based on subtle visual cues observed on the surface of the human skin. The system presented in [18] performs fine-grained occupancy measurement and assigns the actions of the observed subjects to one of the predefined actions to enable even more effective HVAC control. In [19], Tien, Wei and Calautit use a deep learning-based computer vision system to not only monitor the activity of persons in the monitored space, but also account for the active office equipment as a potential heat source. It is worth noting, however, that such a level of sophistication comes at a significantly increased computational cost. Moreover, the vision-based methods, as well as methods based on monitoring of wireless transmissions (wireless local area network (WLAN) [20] or Bluetooth [21]) or wearables [22], are often perceived as privacy-invasive.

In this paper, we present a low-cost occupancy sensor capable of counting the persons in its field of view. The sensor is based on a low-power, low-resolution thermal camera. The camera data are processed using a lightweight convolutional neural network, which facilitates the deployment on low-cost, low-power embedded hardware. The solution is easy to deploy and can be classified as privacy-preserving, since extracting the identity of individuals in the field of view is virtually impossible due to the used thermal modality and low resolution of the imager (32×24 pixels). Our system bears some similarities to other state of the art approaches using the same sensor. For example, Abedi and Jazizadeh [23] use the same type of thermal array sensor (MLX90640) to extract the occupancy information. The method also leverages the deep learning approach, but the resulting information is binary, i.e., it does not extract the person count. Another related solution is presented in [24]. The method is also based on deep learning, and the solution is capable of counting the persons in the field of view of the sensor. However, the task the neural network performs is formulated as classification, where each natural number is a unique class corresponding to the person count. Moreover, the aforementioned solution imposes a maximum hard limit on the count equal to five, which certainly is a factor that hinders its flexibility. The solution presented here is based on a lightweight variant of the U-Net neural network, following the similar encoder–decoder architecture [25]. This is because instead of using a simple binary or one-hot output to predict the occupancy or person count, we estimate the occupancy density function. The concept is borrowed from the crowd counting domain [26], in which it is a solution preferred over explicit detection, as it avoids the dependence on the detector

by learning the mapping of images to density maps [27]. The main contributions of the paper are:

- We present a method for fine-grained occupancy assessment using an inexpensive thermal camera. The method can be deployed on low-power, low-cost embedded hardware. Moreover, it is more flexible and accurate than prior art when tested using a more challenging dataset containing realistic monitoring scenarios involving distractors, collected in multiple locations. In addition, unlike other state of the art solutions built around the same sensor, the presented approach provides useful additional information—the location of the persons in the field of view of the camera.
- We investigate the influence of encoder pretraining using low-resolution grayscale images on the training speed and performance of the complete neural network and demonstrate the gains achieved.
- We introduce a public dataset of sequences for neural network testing and training for the fine-grained space occupancy estimation. The sequences are fully annotated, collected in a few different spaces, and reflect the challenges encountered in realistic conditions. Moreover, we also distribute the code enabling the replication of the experiments shown in the paper. We hope that the availability of an open, public dataset and the source code would encourage research in this domain and allow for systematic evaluation of a variety of approaches using a common benchmark.

2. Materials and Methods

2.1. Dataset and Data Collection

In order to streamline the data collection, a portable setup was prepared. The device consists of a thermal camera, an RGB camera, a computational platform with WLAN connectivity and a portable power source. The block diagram of the setup is given in Figure 1.

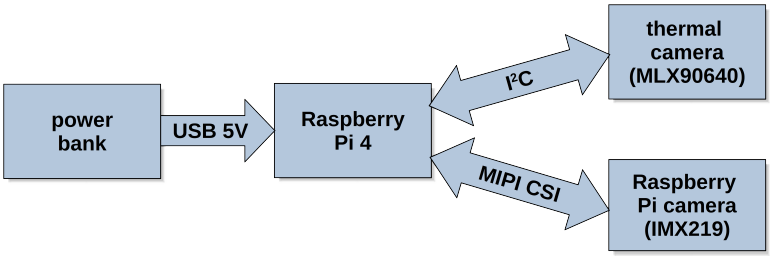

Figure 1. The block diagram of the data collection setup.

The heart of the system is the Raspberry Pi 4 single-board computer (SBC), which is well suited for prototyping and was successfully used in a wide variety of applications [28]. The monitored space is observed by an RGB camera with a IMX219 CMOS sensor with wide angle lens connected to the Raspberry Pi computer with the serial MIPI CSI interface. The data from the RGB camera were used exclusively to make the setup process easier and aid with the experiment control and training data annotation; the data from the RGB camera were not used for neural network training or prediction. The fields of view of the cameras were therefore not calibrated. The most important part of the system is therefore the MLX90640 32 $\times$ 24 pixel resolution, low-cost thermal sensor [29]. The measurement range for each pixel is -40 to 300 °C and typical target object temperature accuracy of 1 °C precision across its full measurement scale. The measurement refresh rates are programmable in the range from 0.5 to 64 Hz, although low operation frequency is is recommended to minimise noise.

The viewing angles of the sensor are 110 and 75 degrees, respectively. The sensor uses the I^2C interface for both control commands and data transmission. No additional calibration of the sensor was performed, e.g., to account for emmissivity, as our system relies on the spatial temperature distribution rather than precise temperature measurements. This makes it easy to integrate it with a variety of embedded systems, including ones based

on low-cost, low-power microcontrollers. The system is powered with a USB cable, using a power bank type battery as a portable power source. Sample images collected using the system are shown in Figure 2. The images were acquired at 2 Hz frequency.

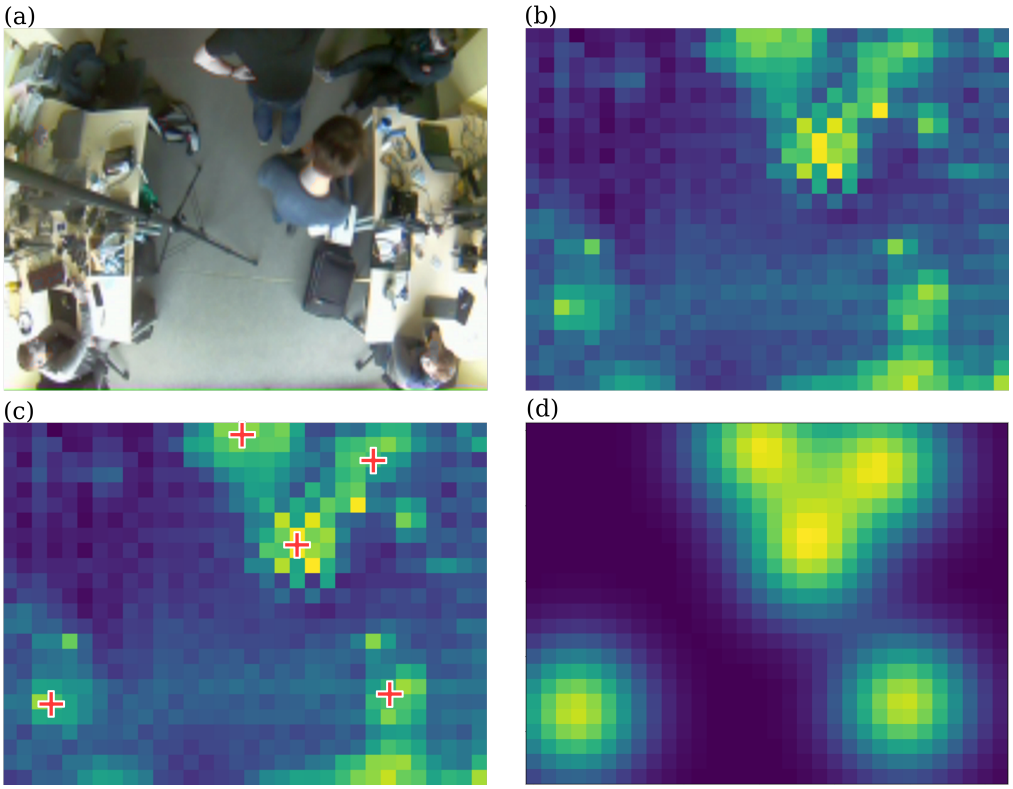

Figure 2. Sample images from the data collection system—RGB image (**a**) and the thermal image (**b**). The locations of the persons in the thermal image as annotated in the dataset (**c**) and the corresponding density map (**d**).

Figure 2 illustrates the method for data annotation. The ground truth starts as an empty image and has the same size as the input image. The locations of the persons in the thermal image are marked by a simple mouse click and applied on the ground truth by setting a single corresponding pixel (in terms of coordinates) in the ground truth image to full brightness. The density map was constructed by convolution of this image with a fixed 2D Gaussian mask, creating a Gaussian mixture distribution with the maximums of the Gaussians corresponding to the locations of the persons. Since the height of the camera does not change drastically between datasets, the Gaussian parameters are the same across all datasets ($\sigma = 3$). Since the sum of the elements of a Gaussian is equal to 1, the person count was computed by summing up all the values in the predicted output density map. The dataset was collected in the office space, with varying conditions (different rooms, locations, etc.). The dataset's characteristics are given in Table 1.

Table 1. Dataset sample counts.

	Number of Persons in a Frame vs. Number of Frames						
	0	**1**	**2**	**3**	**4**	**5**	**Total**
training dataset	99	105	2984	3217	1953	114	8427
validation dataset	0	139	631	1691	225	139	2825
test dataset	162	83	211	341	1235	314	2346

The registered data were divided into short sequences, and the sequences from different locations were distributed into the training, validation and test sets so that the share

of sequences recorded at each location was approximately equal within each set. As a result, the dataset division was not stratified as to the person count. However, such an approach ensures better independence of the training, validation and test sets, as a situation in which two consecutive, highly correlated frames land in different datasets is hindered. Nevertheless, the dataset exhibits much less class imbalance than the dataset used in [24], in which the images in which no person is present correspond to over 80% of the overall dataset. On the one hand, such a situation is probably the most common (for example, in an office outside of the working hours), but for the purpose of training and testing a solution whose main purpose is returning the person count, a more balanced dataset is desirable. The persons moving in the field of view of the sensor were asked to perform a range of standard activities (sitting at the desk, chatting, etc.) to increase the variety of the collected images. The dataset along with the annotations is made public for everyone to download and use (see the 'Data Availability Statement' appendix at the end of the article for the address). We hope it will facilitate the research on this interesting topic and lead to the development of new approaches.

2.2. Neural Network Architecture

As mentioned before, the presented solution enables the reconstruction of the position of the persons in the observed scene, expanding significantly on the functionality enabled by other approaches using the same sensor. To achieve this, an encoder–decoder structure of neural network is used. The block diagram of the architecture is shown in Figure 3.

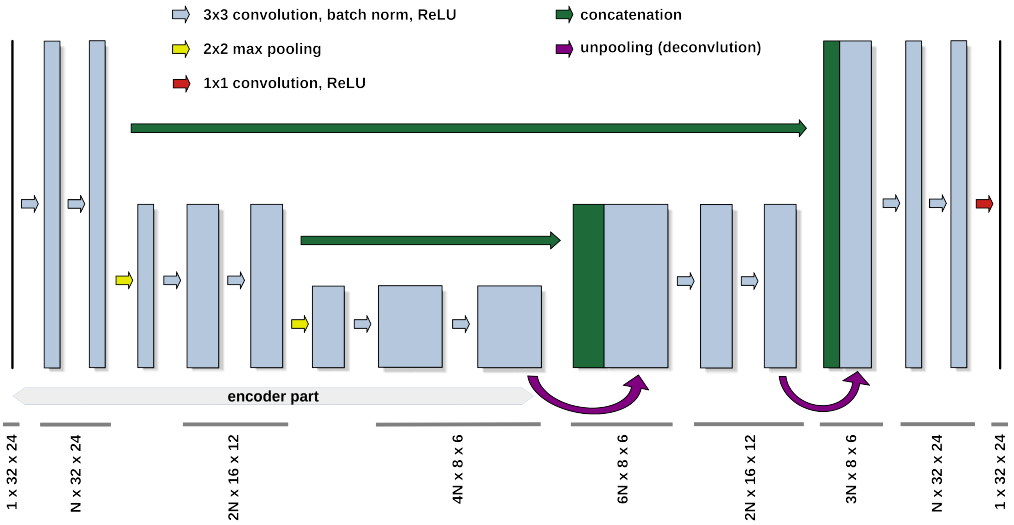

Figure 3. The block diagram of the neural network architecture.

The architecture is based on the U-Net neural network [25], but using small, single channel images as inputs enables significant simplification without compromising the results. Moreover, as the number of parameters of the network is relatively small, it is capable of being used with embedded devices. The simplification is achieved through the reduction in the number of layers in the encoder and decoder stages, and reduction in the number of filters. The network operates with the N set to 32.

Convolutional layers are an essential part of the neural network. The formula for the result of applying the convolutional kernel K to the input feature map I is given in Equation (1), where (x, y) are the feature map coordinates, n_H and n_W are the kernel's height and width, and n_C is the map's and kernel's number of channels.

$$\mathrm{conv}(I, K)_{x,y} = \sum_{i=1}^{n_H} \sum_{j=1}^{n_W} \sum_{k=1}^{n_C} K_{i,j,k} I_{x+i-1,y+j-1,k} \tag{1}$$

The equation demonstrates the result of applying a single convolutional kernel. Applying multiple independent kernels to a single input feature map results in the creation of an output feature map whose depth (number of channels) is equal to the number of the kernels. The coefficients of each kernel and the additional layer-wide bias term are learned during the training phase. Padding is used to keep the size of the output feature maps the same as the size of the input feature maps in terms of width and height. All the layers of the neural network use the rectified linear unit (ReLU) activation function. The function is given by Equation (2).

$$ReLU(x) = max(0, x) \tag{2}$$

Finally, batch normalisation was applied at the output of each layer [30], ensuring zero mean, unit variance distribution of output activations.

The maximum pooling with a 2×2 input window and the stride of 2 wa used to down-sample an input representation. The operator selects the maximum value in the input window and copies the value to the output location. The output of this operation is a feature map with half the width and height of the input feature map. Note that, although the width and height of the feature maps gets smaller, their number increases as the encoder part gets deeper.

The decoder uses the internal representation generated by the encoder to reconstruct the information on the network's output. The deconvolution (understood here as fractional stride convolution) is used to gradually upsample the representation to the original size using trained parameters, with each use doubling the width and height of the image. Also note that the feature maps of corresponding size in the encoder and decoder are concatenated—such information sharing has proven to be useful in terms of accuracy improvement in the original U-Net.

The thermal images are min-max normalized to the range of 0 to 1 before being used for training. Standard training and testing procedure was employed—the training dataset data was used to train the neural network, the validation dataset was used for the control of the training process (key metric value tracking, identification of overfitting), and the test set was used as the holdout data for final testing of the developed solution.

Taking into account the size of the images, the dataset size should be sufficient for training. Nevertheless, additional experiments were performed in which the encoder part of the neural network was trained for classification using low-resolution grayscale images to take advantage of transfer learning [31]. Analogous solution in the form of using an encoder pretrained on ImageNet [32] classification task was successfully used in a similar setting, for example for semantic segmentation [33]. Since ImageNet is comprised of higher resolution, color images, pretraining was performed using the CIFAR-10 dataset converted to grayscale [34] (60,000 32×32 pixel images with 10 classes) and the Fashion MNIST dataset (70,000 28×28 pixel images with 10 classes) [35]. The training was performed using Adam optimizer [36], with the learning rate set to 0.001. Mean average error was selected as the loss function. All data and code used to generate the results along with the trained models is available for download (see Data Availability Statement).

3. Results

The trained network is capable of successfully predicting the occupancy density function as shown in Figure 4.

Note that the local maxima in the predicted output occupancy density image correspond to the actual locations of the persons. The predicted density maxima correspond to the locations of the person in the field of view of the sensor, even though the persons in the thermal image exhibit very different characteristics and temperature distribution. The density function components have a lower value for persons that are only partially visible (see Figure 4, top row for an example). This is due to the neural network being less confident of the partially occluded person's presence, which in turn results in the sum of the distribution components being a fractional number deviating from the true person

count. Nevertheless, obtaining the true person count in the scene is still possible by adding all components in the predicted image and rounding it to the closest natural number. Generally speaking, the presented method would be able to recover the approximate location of the persons present in the room, which can be leveraged to drive a zone temperature control system.

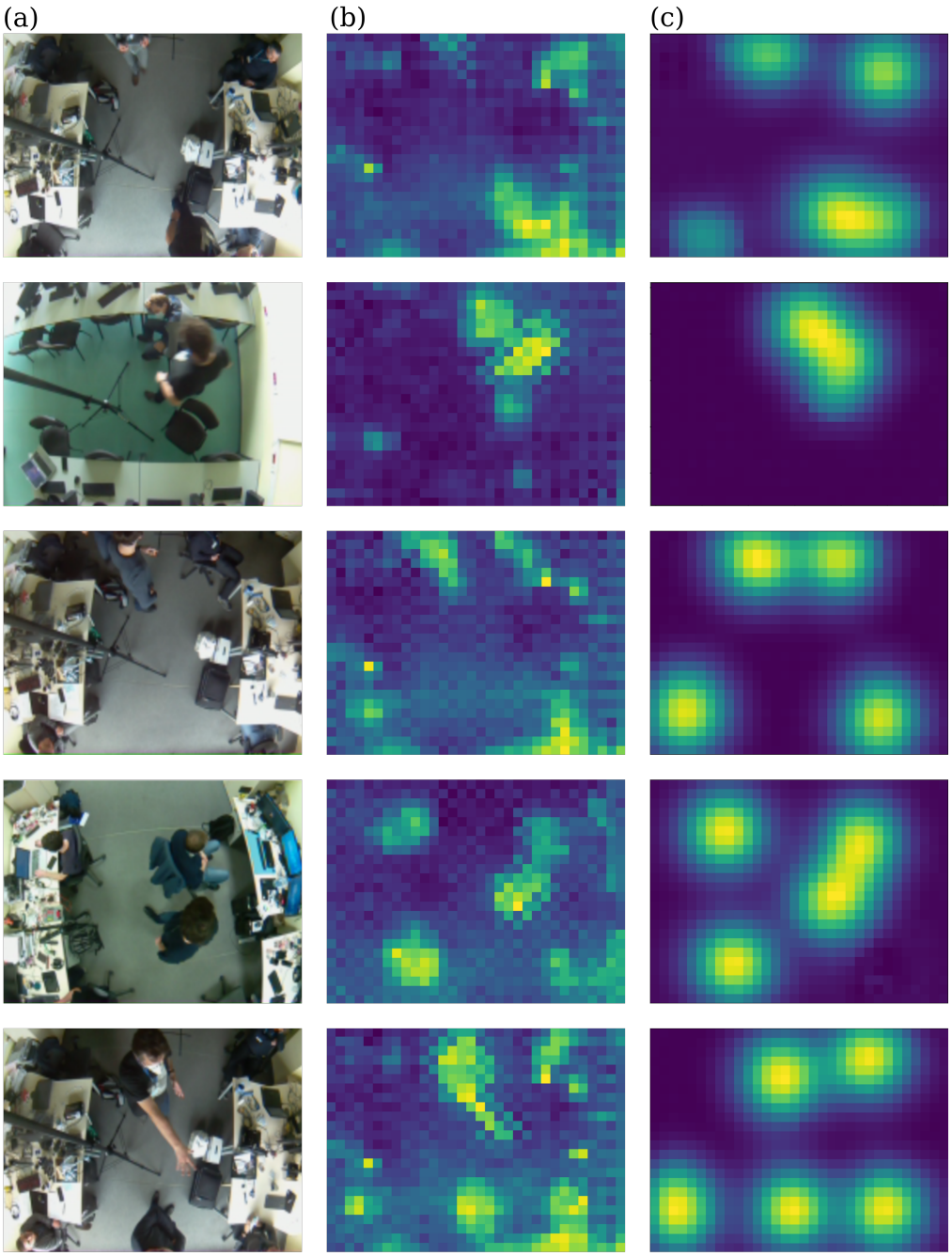

Figure 4. Sample inputs with predictions made by the neural network. Column (**a**)—RGB images, column (**b**)—thermal images (neural network input), column (**c**)—neural network predictions (occupancy density function). Different frames in consecutive rows.

The time plot of predictions is shown in Figure 5. The test dataset sequences were concatenated to form a single plot.

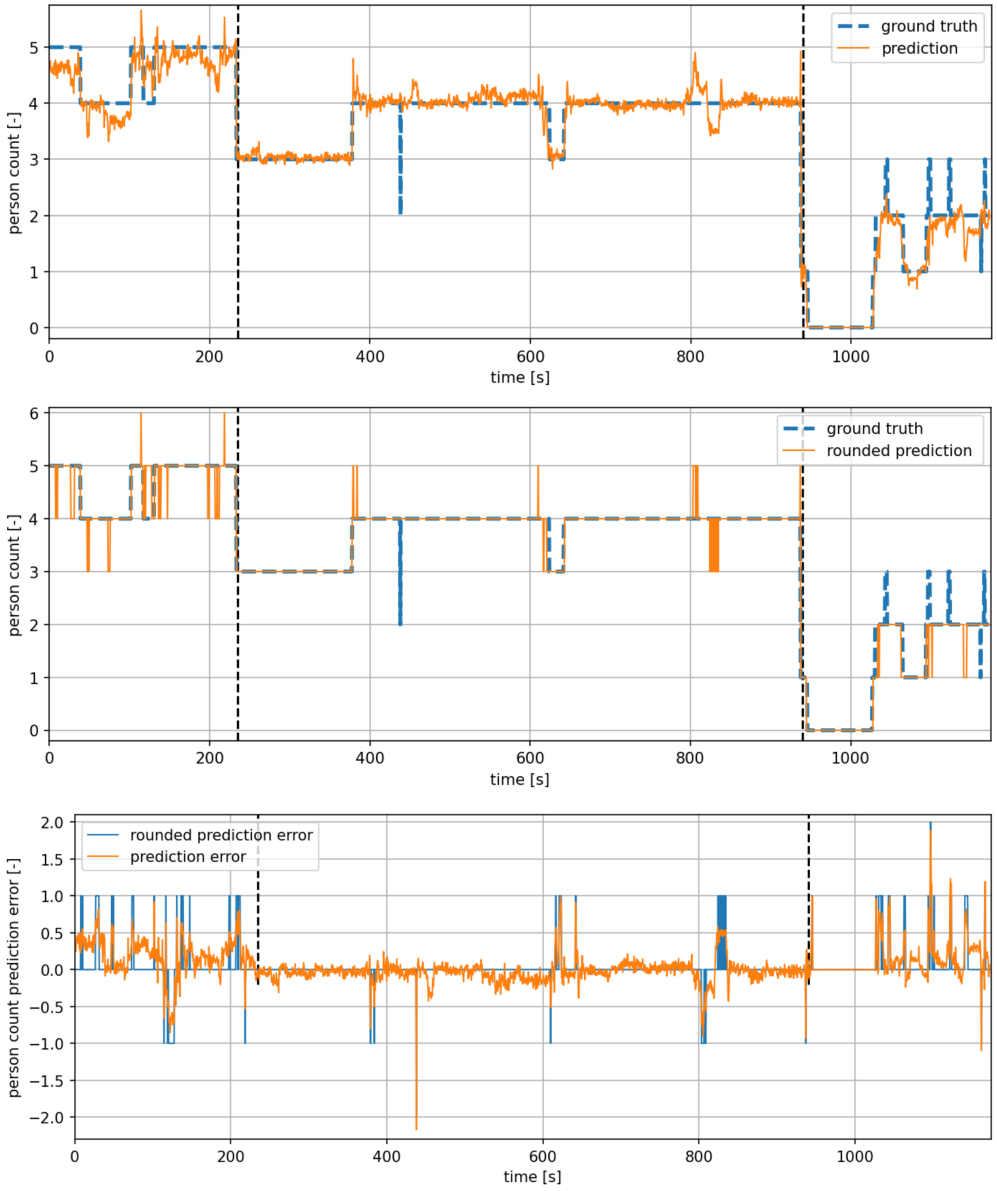

Figure 5. Predictions of person count as a function of time. The black, dashed vertical lines denote the point of concatenation of sequences coming from different scenes. Top row—raw prediction vs. ground truth. Middle row—rounded prediction vs. ground truth. Bottom row—raw and rounded prediction error.

The prediction values are the result of adding all the components of the distribution function, so the raw values can be fractional and deviate from the natural numbers. Simple rounding towards the nearest integer can be used to improve the quality of results in terms of person count. As the result clearly shows, the rounded prediction error is mostly zero and the estimated count is very rarely off by more than one.

To compare the results against the state of the art approach presented in [24], we also performed an experiment using a neural network with the same fully connected structure as the one that was reported to achieve the best accuracy in this work. Since the dataset used to train the network is not publicly available, we trained it using the data from the dataset introduced in our paper, with the same split as we used to train the encoder–decoder network. Using the same data and the same splits enables fair side by side comparison. Comparative results in terms of mean average error (MAE) and mean square error (MSE), as well as the accuracy and the number of neural network parameters are given in Table 2.

Table 2. The comparison of key metrics on the test set—the mean average error and mean squared error, along with the rounded version for direct comparison, the accuracy and the number of trainable parameters of both neural networks.

Metric Name	Our Result	Metwaly et al. Result [24]
MAE	0.145	-
MAE rounded	0.060	0.304
MSE	0.057	-
MSE rounded	0.062	0.470
Accuracy	0.941	0.777
No. of parameters	130,193	396,806

Aside from the global metrics, the confusion matrices providing additional insight into the performance of both approaches were also computer and are shown in Figure 6. The matrices are plotted for the rounded variants of the predictions.

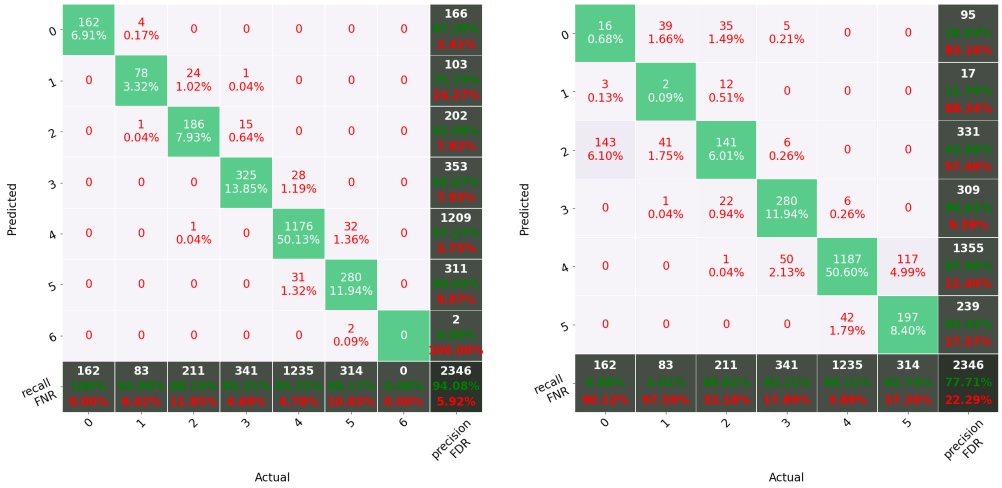

Figure 6. Confusion matrices for the presented solution (**left**) and the solution reproduced from [24]. The rows correspond to the predicted class, while the columns correspond to the ground truth class. The diagonal cells correspond to observations that are correctly classified. The off-diagonal cells correspond to incorrect classifications. Each cell contains the number of samples and its corresponding percentage relative to the total test dataset sample count. The rightmost column of the plot shows the total number of predicted samples, the precision, and false discovery rate for each class. The bottom row shows the total number of ground truth samples, the recall, and false negative rate for each class. The bottom (**right**) cell shows the overall accuracy in terms of the percentage of the correct and incorrect classifications.

4. Discussion

The implemented neural network architecture and the occupancy measurement method have achieved competitive result in comparison to the state of the art when evaluated using the dataset presented in this paper, outperforming the competing solution by a significant margin. As demonstrated by the resulting metrics, fine-grained occupancy estimation is a viable option, even using relatively low-cost and low-power hardware. Due to the small input and output image size, and the overall low computational complexity of the neural network, training can be performed even on relatively low-end hardware. Nevertheless, we have found that using a pretrained encoder speeds up the process roughly twice (final loss function value is reached in about half the time), although offers no benefits in terms of accuracy improvement. Doubling the number of filters in the neural network's layers also does not affect the final accuracy, yet halving the number of filters results in performance degradation. The average single prediction time using the Raspberry Pi 4 SBC is 27 ms.

The areas where the method struggles most is the prediction of the presence and position of persons near the border of the registered area (acquired thermal image), as presented in the top row of Figure 4. With the persons only partially visible, the confidence of the corresponding prediction becomes lower, which is reflected in the estimated density. As a result, the summed up components yield a value below the real number of persons, which can be alleviated quite simply, although only to some degree, by using nearest integer rounding. The problem would, however, become more prominent with the increasing number of persons present at the edges of the field of view.

The approach presented in this paper achieved significantly better results on the presented dataset than the approach described in [24]. Although we have no access to the dataset used in [24], we can draw a highly probable conclusion—that our dataset is more diverse, e.g., due to the background (floor) temperature nonuniformity or the presence of distractors (office equipment, hot beverage mugs, etc.). Moreover, we also collected the data in multiple locations. Furthermore, the use of fully connected network and flattening the input image to a set of independent input features in [24] makes it much harder for the neural network to reason about the underlying spatial relations and structure of the objects observed in the scene. The convolutional neural network introduced in this paper has an inherent capability to deal with two-dimensional data, and convolutions have been proven to be very effective feature extractors for image processing applications [37]. Interestingly, the solution based on the fully connected neural network struggles with dealing with low person counts and seems to be gravitating towards prediction of majority classes. This might be seen as another indication, that preserving spatial information is an important functionality for this task. The classes for this person count are somewhat underrepresented, yet the encoder–decoder network still yields much better results.

Overall, in contrast to [23] the presented system leverages the deep learning techniques used with conjunction with the MLX90640 sensor to provide information beyond simple binary occupancy data. While the capability to estimate the number of persons using similar sensor and set of techniques was shown in [24], our proposal fares notably better in direct comparison using more challenging data. Moreover, it provides the end user with more flexibility and extra information—there is no hard limit on the number of persons as opposed to classification-based approach. Using the presented approach, the approximate location of the persons in the field of view of the camera is known, which facilitates zone-based control [38], and paves the way to extensions such as inclusion of activity profiles in the climate control feedback loop [39]. All this is achieved using a simpler neural network architecture—the parameter count of the closest competing solution (in terms of functionality) is roughly triple.

5. Conclusions

The article presents an approach for fine-grained occupancy monitoring, enabling person counting with relatively high accuracy, which is a clear advantage over simple binary occupancy sensors. The solution is based on a low-power, low-resolution inexpensive thermal camera. As a result, a relatively simple convolutional neural network is capable of performing this task, which facilitates the use of inexpensive, low-power embedded systems as the target platform. Although a universal SBC was used in the experiments for its flexibility, the solution would run on a microcontroller such as the one used in [24], since it uses significantly less parameters than the implementation described therein. Additional benefits include more flexibility (no hard upper limit for the person counter) and providing the end user with additional useful information in terms of person location. This paves the way for the use of more sophisticated control algorithms, further enhancing user comfort and potential for energy savings. The solution preserves the privacy of the end users, as identity recognition using the input images is impossible due to the utilised modality and low resolution. The approach is tested on a new, dedicated dataset, which consists of sequences collected in a few office environment locations. Both the dataset and the code used to replicate the experiments are available for download. Opportunities for

future work include the assessment of thermal parameters of detected persons, e.g., by the measurement of the radiated temperature weighted by the occupancy density function, or monitoring user activity, e.g., by movement-related information extraction for use in lighting or appliance control.

Further development directions include, but are not limited to an exploration of alternative methods for person detection by the density image analysis (e.g., by blob counting) and employment of dedicated tracking approaches and tests with a range of target embedded platforms to optimise the processing speed vs. power consumption trade-off.

Author Contributions: Author Contributions: Conceptualization, M.K.; methodology, M.K. and P.A.; software, P.A. and D.P.; validation, M.K., P.A., D.P. and M.F.; formal analysis, M.K.; investigation, P.A., D.P. and M.K.; resources, M.K.; data curation, P.A.; writing—original draft preparation, M.K.; writing—review and editing, M.K., P.A., D.P. and M.F.; visualization, P.A.; supervision, M.K. and D.P.; project administration, M.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data and code enabling the replication of the presented experiments can be found at https://chmura.put.poznan.pl/s/0JtlXwpqmSFWMNA?path=%2FDatasets%2Fthermo-presence%2Fthermo_presence_article_files, accessed on 26 July 2021.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Spandagos, C.; Ng, T.L. Equivalent full-load hours for assessing climate change impact on building cooling and heating energy consumption in large Asian cities. *Appl. Energy* **2017**, *189*, 352–368. [CrossRef]
2. Tsemekidi Tzeiranaki, S.; Bertoldi, P.; Diluiso, F.; Castellazzi, L.; Economidou, M.; Labanca, N.; Ribeiro Serrenho, T.; Zangheri, P. Analysis of the EU residential energy consumption: Trends and determinants. *Energies* **2019**, *12*, 1065. [CrossRef]
3. U.S. Energy Information Administration. How Much Energy Is Consumed in US Residential and Commercial Buildings? Washington, DC, USA. 2018. Available online: https://www.eia.gov/tools/faqs/faq.php?id=86&t=1 (accessed on 26 July 2021)
4. Lin, T.P.; Lin, F.Y.; Wu, P.R.; Hämmerle, M.; Höfle, B.; Bechtold, S.; Hwang, R.L.; Chen, Y.C. Multiscale analysis and reduction measures of urban carbon dioxide budget based on building energy consumption. *Energy Build.* **2017**, *153*, 356–367. [CrossRef]
5. Weber, R.H.; Weber, R. Internet of Things as Tool of Global Welfare. In *Internet of Things*; Springer: Berlin/Heidelberg, Germany, 2010; pp. 101–125.
6. Asghari, P.; Rahmani, A.M.; Javadi, H.H.S. Internet of Things applications: A systematic review. *Comput. Netw.* **2019**, *148*, 241–261. [CrossRef]
7. Yu, J.; Kim, M.; Bang, H.C.; Bae, S.H.; Kim, S.J. IoT as a applications: Cloud-based building management systems for the internet of things. *Multimed. Tools Appl.* **2016**, *75*, 14583–14596. [CrossRef]
8. Minoli, D.; Sohraby, K.; Occhiogrosso, B. IoT considerations, requirements, and architectures for smart buildings—Energy optimization and next-generation building management systems. *IEEE Internet Things J.* **2017**, *4*, 269–283. [CrossRef]
9. Pasini, D.; Ventura, S.M.; Rinaldi, S.; Bellagente, P.; Flammini, A.; Ciribini, A.L.C. Exploiting Internet of Things and building information modeling framework for management of cognitive buildings. In Proceedings of the 2016 IEEE International Smart Cities Conference (ISC2), Trento, Italy, 12–15 September 2016; pp. 1–6.
10. Tushar, W.; Wijerathne, N.; Li, W.T.; Yuen, C.; Poor, H.V.; Saha, T.K.; Wood, K.L. Internet of things for green building management: Disruptive innovations through low-cost sensor technology and artificial intelligence. *IEEE Signal Process. Mag.* **2018**, *35*, 100–110. [CrossRef]
11. Beltran, A.; Erickson, V.L.; Cerpa, A.E. Thermosense: Occupancy thermal based sensing for hvac control. In Proceedings of the 5th ACM Workshop on Embedded Systems for Energy-Efficient Buildings, Roma, Italy, 11–15 November 2013; pp. 1–8.
12. Kaushik, A.R.; Celler, B.G. Characterization of passive infrared sensors for monitoring occupancy pattern. In Proceedings of the 2006 International Conference of the IEEE Engineering in Medicine and Biology Society, New York, NY, USA, 30 August–3 September 2006; pp. 5257–5260.
13. Zhang, Z.; Gao, X.; Biswas, J.; Wu, J.K. Moving targets detection and localization in passive infrared sensor networks. In Proceedings of the 2007 10th International Conference on Information Fusion, Quebec, QC, Canada, 9–12 July 2007; pp. 1–6.
14. Kemper, J.; Linde, H. Challenges of passive infrared indoor localization. In Proceedings of the 2008 5th Workshop on Positioning, Navigation and Communication, Hannover, Germany, 27–27 March 2008; pp. 63–70.

15. Wahl, F.; Milenkovic, M.; Amft, O. A distributed PIR-based approach for estimating people count in office environments. In Proceedings of the 2012 IEEE 15th International Conference on Computational Science and Engineering, Paphos, Cyprus, 5–7 December 2012; pp. 640–647.
16. Yang, J.; Santamouris, M.; Lee, S.E. Review of occupancy sensing systems and occupancy modeling methodologies for the application in institutional buildings. *Energy Build.* **2016**, *121*, 344–349. [CrossRef]
17. Jazizadeh, F.; Jung, W. Personalized thermal comfort inference using RGB video images for distributed HVAC control. *Appl. Energy* **2018**, *220*, 829–841. [CrossRef]
18. Tien, P.W.; Wei, S.; Calautit, J.K.; Darkwa, J.; Wood, C. A vision-based deep learning approach for the detection and prediction of occupancy heat emissions for demand-driven control solutions. *Energy Build.* **2020**, *226*, 110386. [CrossRef]
19. Tien, P.W.; Wei, S.; Calautit, J. A Computer Vision-Based Occupancy and Equipment Usage Detection Approach for Reducing Building Energy Demand. *Energies* **2021**, *14*, 156. [CrossRef]
20. Balaji, B.; Xu, J.; Nwokafor, A.; Gupta, R.; Agarwal, Y. Sentinel: Occupancy based HVAC actuation using existing WiFi infrastructure within commercial buildings. In Proceedings of the 11th ACM Conference on Embedded Networked Sensor Systems, Roma, Italy, 11–15 November 2013; pp. 1–14.
21. Tekler, Z.D.; Low, R.; Gunay, B.; Andersen, R.K.; Blessing, L. A scalable Bluetooth Low Energy approach to identify occupancy patterns and profiles in office spaces. *Build. Environ.* **2020**, *171*, 106681. [CrossRef]
22. Deng, Z.; Chen, Q. Development and validation of a smart HVAC control system for multi-occupant offices by using occupants' physiological signals from wristband. *Energy Build.* **2020**, *214*, 109872. [CrossRef]
23. Abedi, M.; Jazizadeh, F. Deep-learning for Occupancy Detection Using Doppler Radar and Infrared Thermal Array Sensors. In Proceedings of the International Symposium on Automation and Robotics in Construction (IAARC), Banff, AB, Canada, 21–24 May 2019.
24. Metwaly, A.; Queralta, J.P.; Sarker, V.K.; Gia, T.N.; Nasir, O.; Westerlund, T. Edge computing with embedded ai: Thermal image analysis for occupancy estimation in intelligent buildings. In Proceedings of the INTelligent Embedded Systems Architectures and Applications Workshop 2019, New York, NY, USA, 13–18 October 2019; pp. 1–6.
25. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, 5–9 October 2015; pp. 234–241.
26. Li, B.; Huang, H.; Zhang, A.; Liu, P.; Liu, C. Approaches on crowd counting and density estimation: A review. *Pattern Anal. Appl.* **2021**. [CrossRef]
27. Perko, R.; Klopschitz, M.; Almer, A.; Roth, P.M. Critical Aspects of Person Counting and Density Estimation. *J. Imaging* **2021**, *7*, 21. [CrossRef]
28. Saari, M.; bin Baharudin, A.M.; Hyrynsalmi, S. Survey of prototyping solutions utilizing Raspberry Pi. In Proceedings of the 2017 40th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), Opatija, Croatia, 22–26 May 2017; pp. 991–994.
29. Melexis MLX90640 32x24 IR array—Datasheet. Available online: https://mel-prd-cdn.azureedge.net/-/media/files/documents/datasheets/mlx90640-datasheet-melexis.pdf (accessed on 26 July 2021).
30. Ioffe, S.; Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In Proceedings of the International conference on Machine Learning, PMLR, Lille, France, 7–9 July 2015; pp. 448–456.
31. Zhuang, F.; Qi, Z.; Duan, K.; Xi, D.; Zhu, Y.; Zhu, H.; Xiong, H.; He, Q. A comprehensive survey on transfer learning. *Proc. IEEE* **2020**, *109*, 43–76. [CrossRef]
32. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 248–255.
33. Orsic, M.; Kreso, I.; Bevandic, P.; Segvic, S. In defense of pre-trained imagenet architectures for real-time semantic segmentation of road-driving images. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 12607–12616.
34. Krizhevsky, A. *Learning Multiple Layers of Features from Tiny Images*; Technical Report; University of Toronto: Toronto, ON, Canada, 2009.
35. Xiao, H.; Rasul, K.; Vollgraf, R. Fashion-mnist: A novel image dataset for benchmarking machine learning algorithms. *arXiv* **2017**, arXiv:1708.07747.
36. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
37. Hertel, L.; Barth, E.; Käster, T.; Martinetz, T. Deep convolutional neural networks as generic feature extractors. In Proceedings of the 2015 International Joint Conference on Neural Networks (IJCNN), Killarney, Ireland, 12–16 July 2015; pp. 1–4.
38. Goyal, S.; Ingley, H.A.; Barooah, P. Occupancy-based zone-climate control for energy-efficient buildings: Complexity vs. performance. *Appl. Energy* **2013**, *106*, 209–221. [CrossRef]
39. Budiman, F.; Rivai, M.; Raditya, I.G.B.P.; Krisrenanto, D.; Amiroh, I.Z. Smart Control of Air Conditioning System Based on Number and Activity Level of Persons. In Proceedings of the 2018 International Seminar on Intelligent Technology and Its Applications (ISITIA), Bali, Indonesia, 30–31 August 2018; pp. 431–436.