Human Factors in Architecture, Sustainable Urban Planning and Infrastructure, Vol. 89,
https://doi.org/10.54941/ahfe1003687

Multi-Source Information Fusion Network for Building Occupancy Estimation

Kailai Sun, Tian Xing, Xinwei Wang, Zhou Yang, and Qianchuan Zhao

Center for Intelligent and Networked Systems, Department of Automation, BNRist,
Tsinghua University, Beijing, China

ABSTRACT

The human dimension information is crucial for efficient building energy saving, comfort conditions, health and productivity, and security management. Existing vision-based building indoor occupancy measurement approaches have achieved remarkable progress, but struggle to achieve high and robust accuracy because of the complex indoor environments. Vision-based methods face many challenges, including background objects and diverse scales, which bring practical problems to indoor applications. In this paper, to address these issues, we propose a Multi-source information fusion network in video head detection for estimating building occupancy. Our method utilizes cameras to capture surveillance videos and analyses them through a deep neural network. We use the multi-source feature to effectively guide the single-frame detector to propose robust head boxes. We apply a multi-source fusion network to extract features. Besides, we extend head detection datasets with multi-source information, including optical flow maps, depth maps, frame difference maps, etc. Our method achieves superior performance through ablation studies compared to existing methods on practical building surveillance videos. Experiments validate its potential for building energy saving and comfort improvement with a high occupancy estimation accuracy.

Keywords: Human dimension, Building energy, Artificial intelligence, Occupancy estimation

INTRODUCTION

The human dimension information plays a significant role in practical building energy saving and comfortable indoor environments. Recent research shows that buildings consume approximately 40% of global energy (Simona et al., 2018), while building control strategies based on occupancy information can save energy by 20%–45% and improve thermal comfort by 29.1% (Xie et al., 2020). Building occupant-centric control (OCC) adopts a closed-loop feedback strategy (Zou et al., 2017) to control heating, ventilation and air-conditioning (HVAC), and lighting systems.

To implement OCC, many sensors have been applied to sense the occupancy information, including passive infrared (PIR) sensors, carbon dioxide (CO2) sensors, temperature and humidity sensors, Wi-Fi, Bluetooth, cameras, and power plugs. Vision-based occupancy information measurement, achieving accurate performance, has recently become a hot research topic (Choi et al., 2021, Sun et al., 2022b). Vision-based methods usually capture images/videos using cameras and then recognize people through computer vision and deep learning technologies.

As a key method of building occupancy information measurement, people detection has achieved remarkable progress (Chen et al., 2021). With the development of deep learning, the methods based on convolutional neural networks (CNNs) and Transformer dominate this field. These methods are mainly divided into three categories: body, face, and head detection.

In buildings, body and face detection struggle to achieve high and robust accuracy because of the complex indoor environments (Sun et al., 2022a, Zou et al., 2017, Trivedi and Badarla, 2020). The limitations of body and face detection methods have gradually been exposed. Instead, head detection has a wider range of applications because human heads are visible and reliable in complex indoor environments. In this paper, we focus on the head detection task.

Although many head detection methods are well-advanced (Vora and Chilaka, 2018, Ke et al., 2021, Zheng et al., 2022), the head detection task is still challenging. Background objects (e.g., black balls, bags) have similar features (color, size, texture) to human heads; small-scale, diverse-pose, and low-illumination heads are hard to be detected in crowd scenes. Besides, in video applications, directly applying head detection to every frame often suffers from an unaffordable cost. Long-range video detection/tracking methods would be inaccurate when the appearance of heads dramatically changes, especially as heads move fast or the interval between two nearby frames is large.

To achieve accurate and fast video head detection, the input information sources are important. Motion information can enhance the head features and suppress background features: optical flow and frame difference (Sun et al., 2022a). A depth map can provide useful complementary information. A density map can highlight the spatial features of heads. But the optical flow and frame difference would be inaccurate when the background changes drastically, or the foreground hardly moves (Zhao et al.). Depth maps and density maps have much inherent noise. However, existing studies focus mainly on head detection in static images, depth, or optical flow maps. How to combine the multi-source information to solve the above challenges problem is not considered by previous methods.

Motivated by these observations, we propose a **M**ulti-source **I**nformation **F**usion **N**etwork (**MIFN**) for video head detection at the pixel level. To our best knowledge, it is the first to jointly train the RGB frame, the pixel-level motion information (optical flow and frame difference map), the depth map, and the density map into an end-to-end CNN network in video head detection. It uses the four-source feature to effectively guide the single-frame detector to propose robust head boxes. Our contributions are as follows:

(1) We present a novel solution and new insight for video head detection and building occupancy estimation by utilizing pseudo-information fusion to provide a comprehensive location and appearance information of video

heads. (2) We extend an indoor head detection dataset with multi-source information, including optical flow map, depth map, frame difference map, and density map. (3) We design a pseudo-siamese convolutional network and a feature purification method to get multi-source compatible features. (4) Experimental results indicate that the proposed method significantly surpasses the existing state-of-the-art algorithms on the popular Restaurant dataset. (5) We apply MIFN to occupancy counting and building energy-saving, which confirms its potential in practical building control systems.

METHOD

Given an image, we have a feature extraction network N_{feat}, and a detection network N_{det}. The output for input image I is $N_{det}(h)$, where $h = N_{feat}(I)$.

Multi-Source Fusion

The process is shown in Fig 1. Our system processes the original image I without additional sensors. We adopt a transfer learning strategy, using deep neural networks to automatically generate five-source pseudo-data, and input them to the feature fusion network in parallel:

$$\begin{aligned}
\mathbf{I}_{diff} &= \left|\mathbf{I}_{\mathrm{f}} - \mathbf{I}_{\mathrm{f}-1}\right|, \mathrm{f} = 1, \ldots, N, \\
\mathbf{I}_{flow} &= \mathcal{F}_{flow}\left(\mathbf{I}_{\mathrm{f}}, \mathbf{I}_{\mathrm{f}-1}\right), \mathrm{f} = 1, \ldots, N, \\
\mathbf{I}_{dept} &= \mathcal{F}_{dept}\left(\mathbf{I}_{\mathrm{f}}\right), \\
\mathbf{I}_{dens} &= \mathcal{F}_{dens}\left(\mathbf{I}_{\mathrm{f}}\right), \\
\mathbf{I} &= \mathbf{I}_{\mathrm{f}}.
\end{aligned}$$

We will use the pre-trained model to obtain five-source data: $\mathcal{F}_{flow}$ belongs to the optical flow estimation task and we use an approximation function (Teed and Deng); $\mathcal{F}_{dept}$ belongs to the depth estimation task (Yuan et al., 2022); $\mathcal{F}_{dens}$ belongs to the head density estimation task (Liang and Weiand

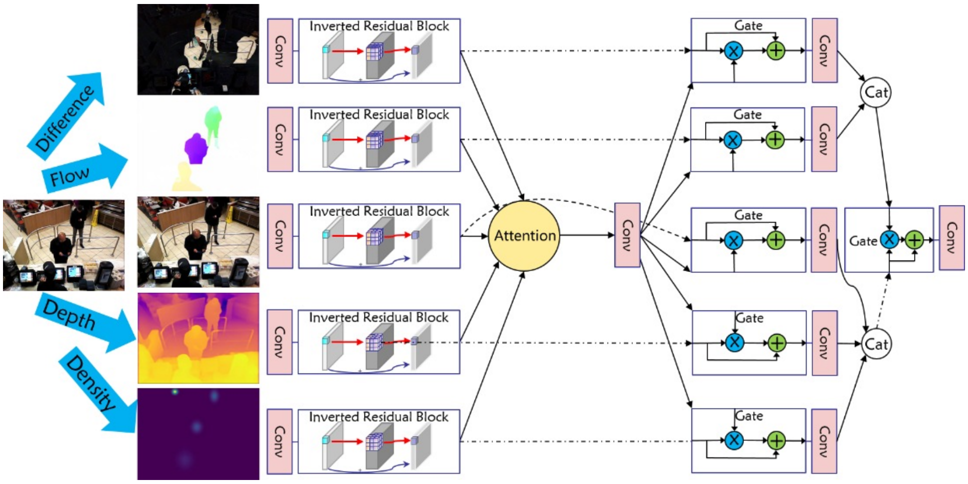


Figure 1: Pipeline of MIFN.

Bai). $\mathbf{I}_{diff}$ is directly obtained from the difference between two frames. Due to we generates five-source information, the fused feature can be written as: $h = N_{feat}(I, I_{\mathrm{diff}}, I_{\mathrm{flow}}, I_{\mathrm{dept}}, I_{\mathrm{dens}})$.

Considering the convenience of feature fusion, we use a pseudo-siamese network (same structure with different weights) to extract different data and then perform two fusions at the feature level. First, we concatenate the five-source features in the channel dimension:

$$\begin{aligned}\mathbf{h}_{cat} &= \mathrm{Cat}(\mathbf{N}_{feat1}\left(\mathbf{I}_{diff}\right), \mathbf{N}_{feat2}\left(\mathbf{I}_{flow}\right), \mathbf{N}_{feat3}\left(\mathbf{I}_{dept}\right), \\ &\quad \mathbf{N}_{feat4}\left(\mathbf{I}_{dens}\right), \mathbf{N}_{feat5}(\mathbf{I})).\end{aligned}$$

Then, we propose a hybrid strategy of convolutional attention and self-attention mechanism. For the convolutional attention mechanism, we use the lightweight convolutional attention mechanisms of the network structure:

$$\mathbf{h}_{fuse} = Attention\left(Conv\left(Conv\left(\mathbf{h}_{cat}\right)\right)\right).$$

Where we first perform the attention mechanism, extracting the spatial coordinate information of multi-source coupling. Subsequently, we will implement the self-attention mechanism strategy on the obtained features and the original features:

$$\mathbf{h}_{fuse} = \alpha_1 \cdot Conv\left(Conv\left(\mathbf{h}_{fuse}\right)\right) \odot \mathbf{h}_{cat} + \beta_1 \cdot \mathbf{h}_{cat}.$$

Where we use the convolutional network before fusion to construct a learnable mask and predict the score of the features so that more effective head features can be selected.

Second, we separate channels and re-concatenate the features:

$$\begin{aligned}\mathbf{h}_{motion} &= Cat\left(Conv\left(Conv\left(\mathbf{h}_{fuse1}\right)\right), Conv\left(Conv\left(\mathbf{h}_{fuse2}\right)\right)\right), \\ \mathbf{h}_{static} &= Cat(Conv\left(Conv\left(\mathbf{h}_{fuse3}\right)\right), Conv\left(Conv\left(\mathbf{h}_{fuse4}\right)\right), \\ &\quad Conv\left(Conv\left(\mathbf{h}_{fuse5}\right)\right)\end{aligned}$$

Where we perform convolution operations on the features after channel separation, and then concatenate the corresponding features of motion and static respectively to prepare for the next fusion:

$$\mathbf{h}_{fuse} = \alpha_2 \cdot \mathbf{h}_{static} \odot \mathbf{h}_{motion} + \beta_2 \cdot \mathbf{h}_{static}.$$

Head Detector

After getting the fused features, we can apply many detectors: convolution and Transformer-based detectors. To simplify the process, we use a modified RPN to propose head boxes.

$$\left(c, \mathbf{p}\right) = \mathbf{N}_{\mathrm{det}}\left(\mathbf{h}_{\mathrm{fuse}}\right).$$

Where $\mathbf{c} = \left(\mathbf{c}_{\mathrm{x}}, \mathbf{c}_{\mathrm{y}}, \mathbf{c}_{\mathrm{w}}, \mathbf{c}_{\mathrm{h}}\right)$, is the matrix that contains the parameterized coordinates of anchor boxes. $\mathbf{c}_{\mathrm{x}}, \mathbf{c}_{\mathrm{y}}$ are the predicted center coordinates of head boxes, while $\mathbf{c_w}, \mathbf{c_h}$ are the predicted width and height of head boxes. $\mathbf{p}$ is the probability matrix that predicts the head category.

To train the network, labels will be generated as follows. We convert every ground truth head box $H = \left(H_x, H_y, H_w, H_h\right)$ to the parameterized coordinate:

$$\begin{aligned}
\widehat{c_x} &= \left(H_x - A_x\right)/A_w \\
\widehat{c_y} &= \left(H_y - A_y\right)/A_h \\
\widehat{c_w} &= \log\left(H_w/A_w\right) \\
\widehat{c_h} &= \log\left(H_h/A_h\right),
\end{aligned}$$

where $A = \left(A_x, A_y, A_w, A_h\right)$ is one of the anchors. Ground truth parameterized coordinate matrix $\widehat{\mathbf{c}}$ will be generated. Ground truth category matrix $\widehat{\mathbf{p}}$ will be generated by directly distinguishing whether the anchor contains a head or not (1 or 0). We employ a multi-task loss function:

$$L(\mathbf{c}, \mathbf{t}) = \frac{1}{N_s} \sum_n \left[L_{cla}\left(\mathbf{p_n}, \widehat{\mathbf{p_n}}\right) + L_{box}\left(\mathbf{c_n}, \widehat{\mathbf{c_n}}\right)\right].$$

As mentioned above, for each bounding box, $\widehat{\mathbf{c_n}}$ is the category ground-truth, and $\widehat{\mathbf{p_n}}$ is the coordinate ground-truth. n is the index of the bounding box. We wish to minimize the distance between each predicted anchor box ($\mathbf{p_n}$, $\mathbf{c_n}$) and the corresponding ground truth. L_{cla} is the cross-entropy loss, while L_{box} is the smoothed L1 loss. The loss term is normalized by N_s, where N_s is the number of positive samples. Also, the loss function is computed using only positive samples. Positive samples are defined by three strategies: (1) IOU between anchor box and ground truth label ≥ 0.7. (2) Each ground-truth box overlaps with many anchor boxes, and we mark the anchor box corresponding to the largest IoU as a positive sample. (3) We limit the number of positive samples to ≤ 16. After defining the loss function, the backbone network and the detector are jointly trained end-to-end.

EXPERIMENT

To evaluate MIFN, we test the publicly available crowd Restaurant dataset (El Ahmar et al.). For evaluation metrics, we use the standard average precision (AP^{50}). The Restaurant dataset was collected in four different indoor locations at a restaurant. It includes 1610 images, from which the test set contains 123 images. The images are extracted from the video with a large time interval, thus having significant diversity and difference.

Our detection network is described in Section Method. The training hyperparameters are given. Backbone uses the first 11 layers of the MobileNetv2 network pre-trained on the ImageNet dataset. The anchor box sizes are selected as 2 and 4, and the whole model is trained by the SGD optimizer for 50 epochs. The learning rate is 10-2, which decays to 10-3, 10-4, and 10–5 after 15, 35, and 42 epochs, respectively. The detector RPN network consists of 5 convolutional layers, initialized using a standard normal distribution with a standard deviation of 0.01. We set $\alpha_1 = \beta_1 = \alpha_2 = \beta_2 = 1$. The batch size is set to 1.

The results are shown in Tab. 1. Our MIFN is superior to other SOTA algorithms (including detectors based on CNN and Transformer). It is worth noting that we use the MobileNetv2 network (Sandler et al.). Although this network is lightweight, the network performance will decrease. Even in this case, our algorithm is still excellent.

For further analysis, we conduct ablation experiments to explore the importance of multi-source information, as shown in Tab. 2. We found that: the pseudo-multi-source information effectively improves the network performance and does not require additional sensors; motion information, especially frame difference information, plays a key role in head detection; density information can further improve and enhance the head detection network. A partial visualization of the qualitative results is shown in Fig. 2: We show successful results where all heads are accurately detected even when several heads are small. There is an error in the bottom-right image, and the head with the red hat is missing. In conclusion, the experimental results validate the effectiveness of our algorithm.

Applications: Building Occupancy Estimation and Energy-Saving

Building occupant-centric control is an important application of occupancy estimation for indoor comfort and energy-saving (Naylor et al., 2018). Our

Table 1. Comparison of MIFN against other SOTA methods on restaurant dataset.

Method	Backbone	AP^{50}
HTC++(Chen et al., 2019)	Swin-B	0.68
SSD (Ke et al., 2021)	ResNet18	0.51
FCHD (Vora and Chilaka, 2018)	VGGNet16	0.75
CrowdDet (Chu et al., 2020)	ResNet50	0.61
Iter-E2EDET (Zheng et al., 2022)	ResNet50	0.61
MPSN(Sun et al., 2022a)	MobileNetv2	0.84
MIFN	MobileNetv2	0.86

Table 2. Ablation study on validation and test sets of restaurant data sets.MN2: MobileNetv2.

Method	Module	val AP^{50}	test AP^{50}
MN2+RPN	RGB	~	0.785
MN2+RPN	RGB+Flow	0.759	0.790
MN2+RPN	RGB+Diff	0.752	0.838
MN2+RPN	RGB+Diff+Flow	0.749	0.826
MN2+RPN	RGB+Diff+Flow+Depth	0.804	0.841
MN2+RPN	RGB+Diff+Flow+Depth+Density	0.811	0.860
MN2+RPN	RGB+Diff+Flow+Depth+Density	0.830	0.856

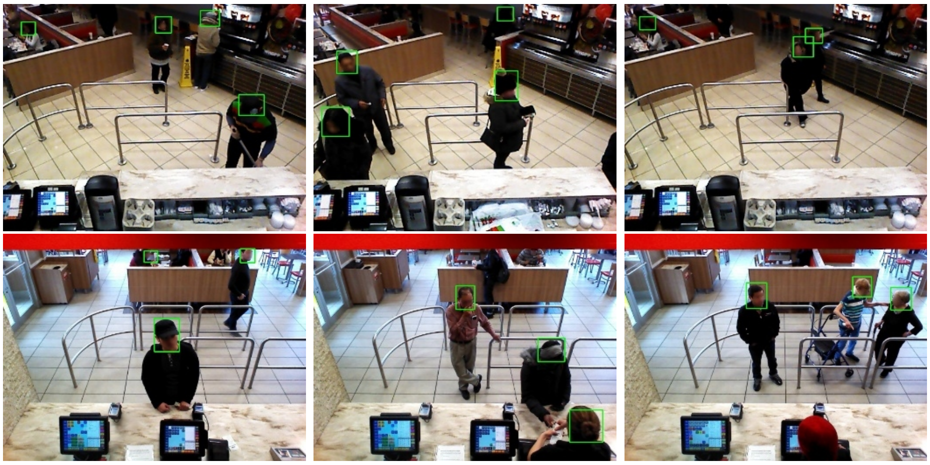


Figure 2: Visualization of MIFN on the test set of restaurant dataset.

previous sections have demonstrated an accurate indoor occupancy estimation method utilizing multi-source information fusion. In this section, we will use the occupant estimations of the restaurant dataset to design the OCC algorithm for energy saving.

Two different areas in the restaurant dataset are selected as the application scenarios. These two scenarios detected 104 and 102 frames in total, respectively. The indoor HVAC equipment is considered to be fan coil units (FCU), which is widely used for public buildings like offices, restaurants, etc. (Lu et al., 2009). With the increment of indoor occupants, the cooling load and the amount of fresh air required to maintain indoor air quality will both increase. Therefore, the FCU should enlarge the supply air volume as the number of people in the room increases. Our OCC algorithm is designed based on the occupant estimations' distributions. Thus, we counted the distribution of the estimated results of the restaurant dataset, as Fig. 3 illustrates.

The control bounds of FCU are determined by the α-upper quantiles of occupant distributions, which satisfies:

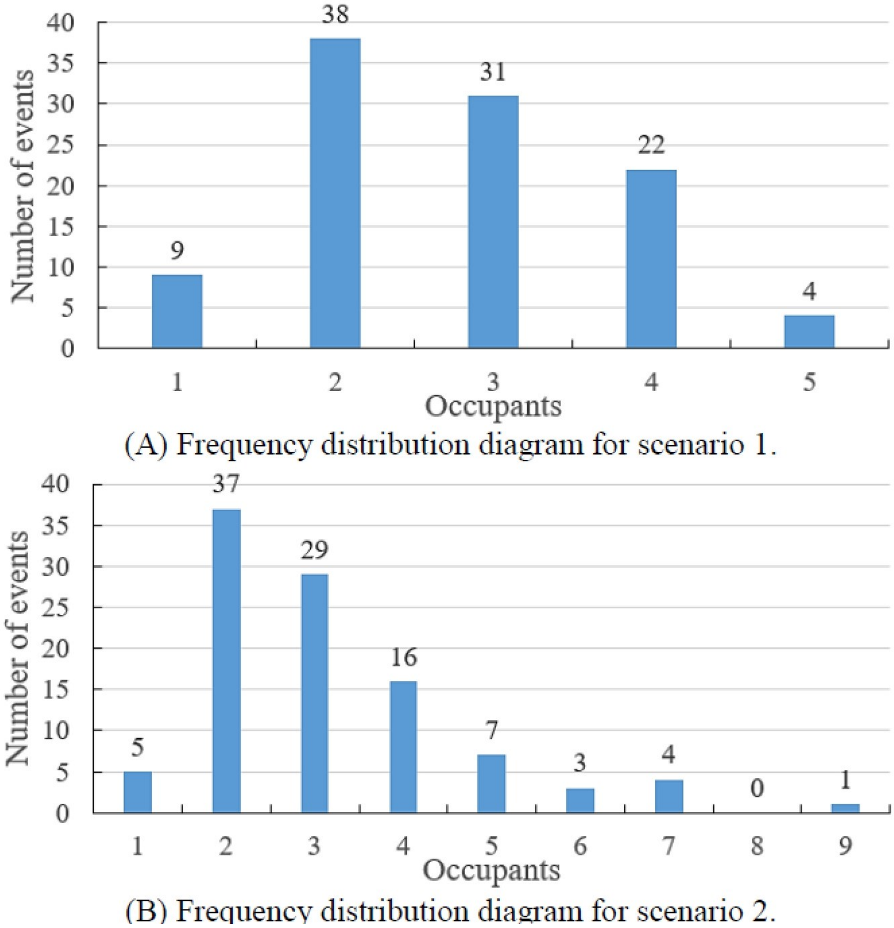
$$F(\alpha) = x_\alpha, \quad s.t. \quad P(X > x) = \alpha, \tag{1}$$

where x_α is the occupant number for the α-upper quantile, X is the indoor occupant number, and $\alpha \epsilon [0, 1]$. Since the FCU device only has three fan speed levels (low, medium, and high), we choose the occupant numbers for 0.33, 0.67, and 1-upper quantiles as the control bounds for the fan speed control. The OCC algorithm for an FCU is:

$$C(FCU) = \begin{cases} high, & if\ X > x_{0.33} \\ medium, & if\ x_{0.67} < X \leq x_{0.33} \\ low, & if\ X \leq x_{0.67} \end{cases}. \tag{2}$$

According to Eq. (1) (2), the control bounds for scenarios 1 and 2 are listed in Tab 3:

We followed the work (Atienza Márquez et al., 2017) to evaluate the energy cost of an FCU at different fan speed levels. We compared the power



(A) Frequency distribution diagram for scenario 1.

(B) Frequency distribution diagram for scenario 2.

Figure 3: Occupant frequency distributions for the application scenarios.

Table 3. Control bounds for our OCC algorithm.

Fan speed level	Occupant range	
	Scenario 1	Scenario 2
high	>3	>4
medium	3	3 and 4
low	<3	<3

expectations of our OCC algorithm and the baseline full-open strategy, which can be calculated by:

$$E(w) = \sum_{i=1}^{n} \frac{w(i)}{n}, \tag{3}$$

where n is the total frame number, $w(i)$ is the FCU's power (kW) at the ith frame. The power comparisons of different control algorithms are listed in Tab 4.

Table 4. Power expectation comparisons.

Method	Baseline	Scenario 1	Scenario 2
Power expectation (kW)	8.41	6.77	6.68
Relative power reduction (%)	-	19.50	20.58

The results indicate that our OCC algorithm can save about 20% HVAC energy compared to a non-OCC baseline strategy, which has shown a great energy-saving potential of our method.

CONCLUSION

In this work, we propose an occupancy estimation system using a multi-source fusion network and object detector. Experimental results demonstrate that our algorithm achieves SOTA performance and verifies proposed insights. Mathematical results find our algorithm has a significant potential to save about 20% of energy in buildings. As for future work, we will extend our algorithm and datasets to obtain occupancy distribution information for different rooms and scenes.

ACKNOWLEDGMENT

This work is supported by Key R&D Project of China under Grant No. 2017YFC0704100, 2016YFB0901900, National Natural Science Foundation of China under Grant No. 62192751 and 61425024, the 111 International Collaboration Program of China under Grant No. BP2018006, 2019 Major Science and Technology Program for the Strategic Emerging Industries of Fuzhou under Grant No. 2019-Z-1, and in part by the BNRist Program under Grant No. BNR2019TD01009, the National Innovation Center of High Speed Train R&D project (CX/KJ-2020-0006).

REFERENCES

ATIENZA MáRQUEZ, A., CEJUDO LóPEZ, J. M., FERNáNDEZ HERNáNDEZ, F., DOMíNGUEZ MUñOZ, F. & CARRILLO ANDRéS, A. 2017. A comparison of heating terminal units: Fan-coil versus radiant floor, and the combination of both. *Energy and Buildings,* 138, 621–629.

CHEN, K., PANG, J., WANG, J., XIONG, Y., LI, X., SUN, S., FENG, W., LIU, Z., SHI, J., OUYANG, W., LOY, C. C. & LIN, D. Hybrid Task Cascade for Instance Segmentation. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 15–20 June 2019 2019. 4969–4978.

CHEN, L., LIN, S., LU, X., CAO, D., WU, H., GUO, C., LIU, C. & WANG, F.-Y. 2021. Deep Neural Network Based Vehicle and Pedestrian Detection for Autonomous Driving: A Survey. *IEEE Transactions on Intelligent Transportation Systems,* 22, 3234–3246.

CHOI, H., UM, C. Y., KANG, K., KIM, H. & KIM, T. 2021. Review of vision-based occupant information sensing systems for occupant-centric control. *Building and Environment,* 203, 108064–108064.

CHU, X., ZHENG, A., ZHANG, X. & SUN, J. Detection in Crowded Scenes: One Proposal, Multiple Predictions. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 13–19 June 2020 2020. 12211-12220.

EL AHMAR, W. A., ERLIK NOWRUZI, F. & LAGANIERE, R. Fast Human Head and Shoulder Detection Using Convolutional Networks and RGBD Data. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2020. 479–487.

KE, R., ZHUANG, Y., PU, Z. & WANG, Y. 2021. A Smart, Efficient, and Reliable Parking Surveillance System With Edge Artificial Intelligence on IoT Devices. *IEEE Transactions on Intelligent Transportation Systems,* 22, 4962–4974.

LIANG, D. X. & WEIAND BAI, X. An End-to-End Transformer Model for Crowd Localization. *In:* AVIDAN, S. B., GABRIELAND CISSé, M. F. & GIOVANNI MARIAAND HASSNER, T., eds. Computer Vision – ECCV 2022, 2022. Springer Nature Switzerland, 38–54-38–54.

LU, Z., LU, W. Z., ZHANG, J. L. & SUN, D. X. 2009. Microorganisms and particles in AHU systems: Measurement and analysis. *Building and Environment,* 44, 694–698.

NAYLOR, S., GILLOTT, M. & LAU, T. 2018. A review of occupant-centric building control strategies to reduce building energy use. *Renewable and Sustainable Energy Reviews,* 96, 1–10.

SANDLER, M., HOWARD, A., ZHU, M., ZHMOGINOV, A. & CHEN, L.-C. Mobilenetv2: Inverted residuals and linear bottlenecks. Proceedings of the IEEE conference on computer vision and pattern recognition, 2018. 4510–4520-4510–4520.

SIMONA, D., HONG, T. & LANGEVIN, J. 2018. The human dimensions of energy use in buildings: A review. *Renewable and Sustainable Energy Reviews,* 81, 731–742.

SUN, K., MA, X., LIU, P. & ZHAO, Q. 2022a. MPSN: Motion-aware Pseudo-Siamese Network for indoor video head detection in buildings. *Building and Environment,* 222, 109354–109354.

SUN, K., ZHAO, Q., ZHANG, Z. & HU, X. 2022b. Indoor occupancy measurement by the fusion of motion detection and static estimation. *Energy and Buildings,* 254, 111593–111593.

TEED, Z. & DENG, J. Raft: Recurrent all-pairs field transforms for optical flow. European conference on computer vision, 2020. 402–419-402–419.

TRIVEDI, D. & BADARLA, V. 2020. Occupancy detection systems for indoor environments: A survey of approaches and methods. *Indoor and Built Environment,* 29, 1053–1069.

VORA, A. & CHILAKA, V. 2018. FCHD: Fast and accurate head detection in crowded scenes. *arXiv preprint arXiv:1809.08766.*

XIE, J., LI, H., LI, C., ZHANG, J. & LUO, M. 2020. Review on occupant-centric thermal comfort sensing, predicting, and controlling. *Energy and Buildings,* 226, 110392–110392.

YUAN, W., GU, X., DAI, Z., ZHU, S. & TAN, P. 2022. New crfs: Neural window fully-connected crfs for monocular depth estimation. *arXiv preprint arXiv:2203.01502.*

ZHAO, X., PANG, Y., YANG, J., ZHANG, L. & LU, H. Multi-Source Fusion and Automatic Predictor Selection for Zero-Shot Video Object Segmentation. Proceedings of the 29th ACM International Conference on Multimedia, 2021. Association for Computing Machinery, 2645–2653-2645–2653.

ZHENG, A., ZHANG, Y., ZHANG, X., QI, X. & SUN, J. Progressive End-to-End Object Detection in Crowded Scenes. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022. 857-866.

ZOU, J., ZHAO, Q., YANG, W. & WANG, F. 2017. Occupancy detection in the office by analyzing surveillance videos and its application to building energy conservation. *Energy and Buildings,* 152, 385–398.