Hindawi
Wireless Communications and Mobile Computing
Volume 2022, Article ID 2608798, 8 pages
https://doi.org/10.1155/2022/2608798

Research Article

Image Real-Time Detection Using LSE-Yolo Neural Network in Artificial Intelligence-Based Internet of Things for Smart Cities and Smart Homes

Zheng Zhi-Xian[1] and Fuquan Zhang[2,3]

[1]*Fujian Chuanzheng Communications College, Fuzhou Fujian, 350007, China*
[2]*College of Computer and Control Engineering, Minjiang University, Fuzhou, China 350108*
[3]*Digital Media Art, Key Laboratory of Sichuan Province, Sichuan Conservatory of Music, Chengdu, China 610021*

Correspondence should be addressed to Fuquan Zhang; zfq@mju.edu.cn

Received 6 January 2022; Accepted 22 February 2022; Published 9 March 2022

Academic Editor: Chao-Yang Lee

In this paper, a novel visual image real-time detection LSE-Yolo neural network is presented, which is in artificial intelligence-based Internet of Things for smart cites and smart homes. Despite the great achievements that have been acquired in image detection, the issue of visual image real-time detection combined with privacy data protection to serve for smart cities and smart homes has been overlooked. The technique we applied in our study is referred to as visual object detection, which can contribute to more healthy and comfortable life. When several studies have been carried out to test the validity, it is suggested that our proposed LSE-Yolo neural network has better performance in image real-time detection based on AIoT for smart cities and smart homes. And it is similar to state-of-the-art. The fruitful work has made great contributions to our present understanding of the visual image detection serving for smart cities and smart homes.

1. Introduction

Recently, with the improvement of residents' living standards and consumption power, the traditional life has been changed to the life with science and technology, which can make it more healthy, fast, convenient, and comfortable. To meet the needs of modern smart cities and homes, visual system in computer vision applied in smart cities and smart homes has become a hot topic in artificial intelligence and Internet of Things (AIoT). Therefore, image detection in AIoT has attracted much attention from the academia, which can serve for smart cities and smart homes. There have been several studies highlighting object detection for smart cities and smart homes [1–8] in recent years. And object detection plays an important role in the visual system application in AI-based for smart cities and homes. The common object detection algorithms can be classified into two categories: one is traditional detectors and the other is deep learning-based detectors. One of the typical traditional detectors is HOG detector, which can be can be regarded as an important improvement over scale-invariant feature transform and shape context, and it was firstly proposed by Dalal et al. in 2005. And DPM proposed by Felzenszwalb et al. [9] in 2008 was the culmination of traditional target detection methods. In deep learning-based detectors, there are two major classes: one-stage detectors [10–12] and two-stage detectors [13, 14], respectively. In one-stage detectors, features from the network to predict object classification and location directly. While in two-stage detectors, a proposal is required; that is, a preselection box containing the objects to be detected, and then, fine-grained object detection is carried out. In recent years, it has become a trend that object detection algorithms served for smart cities and homes. In literature, Khan et al. [3] in 2017 presented the detection of people through computer vision in the Internet of Things scenarios to improve the security in smart cities, smart towns, and smart homes. In 2019, Garcia et al. [1] proposed to perform object detection mechanism based

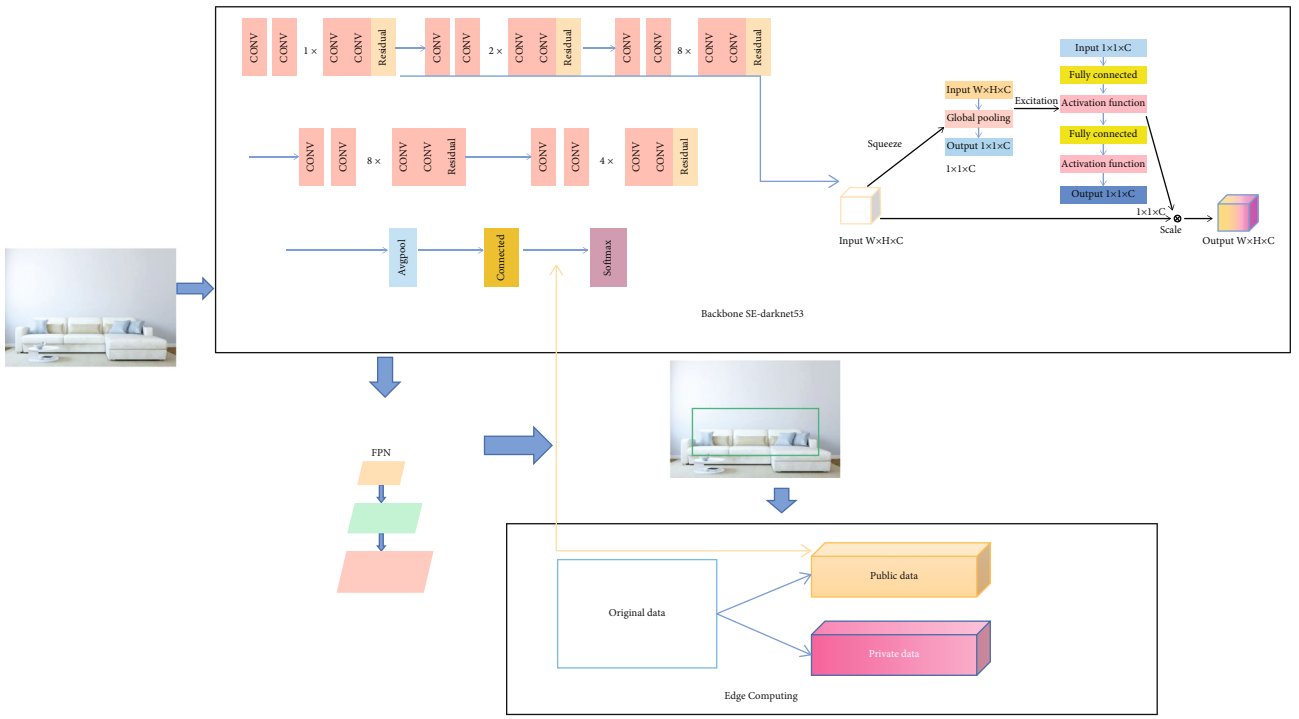

Figure 1: The architecture of our LSE-Yolo visual image real-time detection for smart cities and smart home.

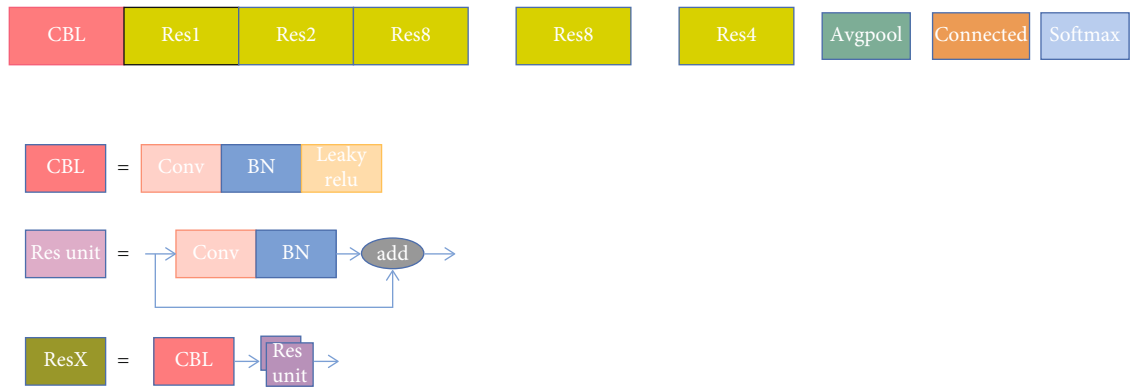

Figure 2: The architecture of darknet53.

on deep learning algorithm using embedded IoT devices for smart home appliance control in CoT. Mehmood et al. [2] in 2021 proposed deep learning-based hybrid approach for the development of an IoT-based intelligent home security and appliance control system in the smart cities. Park J et al. [4] in 2018 presented to use deep neural networks for activity recognition with multisensor data in a smart home. Meanwhile, Hu and Ni [7] proposed IoT-driven automated object detection algorithm for urban surveillance systems in smart cities. In 2019, Nayak et al. [5] proposed to use deep learning video-based real-time intrusion detection system for smart city applications. Then in 2020, Wang et al. [6] presented the algorithm for target detection in smart city combined with depth learning and feature extraction. Xu et al. in 2021 [8] presented feature-enhanced occlusion perception object detection for smart cities. Wang et al. [15] in 2019 proposed an end-to-end three-dimensional (3D) object detection method based on sparse convolution neural network and feature fusion for autonomous driving in smart cities. In 2020, Mettupally and Menon [16] presented a smart ecosystem for parking detection using deep learning and big data analytics for smart city. Great progress has been made in the object detection algorithm applied in smart cities and homes; some limitations are the following: (1) lack of sufficient data; it is hard to obtain a sound object detection model served for smart cities and smart homes. (2) Due to the limitation in common deep neural networks, it is difficult to be deployed in the mobile device and real-time detection carried out to contribute to smart cities and smart homes. (3) Neglecting the privacy data protection of the users, it is tough to realize the smart city- and smart home-based AIoT. In order to overcome the problems and limitations stated

Table 1: The structure of darknet53.

	Type	Filters	Size	Output
	Convolutional	32	3 × 3	512 × 512
	Convolutional	64	3 × 3/2	128 × 128
	Convolutional	32	1 × 1	
1×	Convolutional	64	3 × 3	
	Residual			128 × 128
	Convolutional	128	3 × 3/2	64 × 64
	Convolutional	64	1 × 1	
2×	Convolutional	128	3 × 3	
	Residual			64 × 64
	Convolutional	256	3 × 3/2	32 × 32
	Convolutional	128	1 × 1	
8×	Convolutional	256	3 × 3	
	Residual			32 × 32
	Convolutional	512	3 × 3/2	16 × 16
	Convolutional	256	1 × 1	
8×	Convolutional	512	3 × 3	
	Residual			16 × 16
	Convolutional	1024	3 × 3/2	8 × 8
	Convolutional	512	1 × 1	
4×	Convolutional	1024	3 × 3	
	Residual			8 × 8
	Avgpool		Global	
	Connected		1000	
	Softmax			

above, in our work, a new neural network for visual image system-based AIoT is presented to serve for smart cities and smart homes to perform real-time detection. To solve the problem of the insufficient data, supervised data augmentation is adopted in our network including single-sample data augmentation and multiple data augmentation, in which the data can be enlarged on the basis of existing data. And more lightweight neural network is proposed in our study, which can be installed easily and no dependencies. So it can be deployed on the mobile device to carry out real-time detection-based AIoT for the smart cities and smart homes. In addition, to protect the privacy data, edge computing data are collected and calculated locally or on edge nodes without being uploaded to the cloud, and important and sensitive information does not need to be transmitted through the network, thus effectively avoiding the problem of privacy leakage. Therefore, the proposed LSE-Yolo neural network can realize the smart city- and smart home-based AIoT. The contributions in our study are the following:

(1) A new neural network for visual image systems-based AIoT is presented to serve for smart cities and smart homes to perform real-time detection

(2) Supervised data augmentation is adopted in our network to address the problem of the insufficient data

(3) To protect the privacy data, edge computing is used to classify the data into public data and private data to avoid the privacy leakage

This paper is divided into 4 sections as follows: Section 2 introduces the materials and methods, that is, the architecture of our LSE-Yolo neural network. The results and discussion of our work are shown in Section 3. Finally, Section 4 summed up this paper.

2. Materials and Methods

Figure 1 shows the architecture of the novel LSE-Yolo neural network for visual image systems to perform detection-based AIoT to serve for smart cities and smart homes. In the image input, it would be conveyed to the model to perform the operation. The input image is the sofa in a home which is sampled by the mobile device randomly. Firstly, it would go through the backbone network, that is, the darknet53, which contains 53 convolutional layers to extract the features and the information. And the Squeeze-and-Excitation module is added after each residual layer to perform the Squeeze part and Excitation operation, which can solve the loss problem caused by different importance of different channels of feature map in the process of convolutional pooling. Then in the FPN module, it would be carried out by upsampling by high-level features and top-down linking by low-level features, and predictions are made for each layer. Finally, the output of the image which be detected and taken as the sofa will proceed to the database to perform edge computing to be classified the public data or the private data for the smart cities and smart homes. And the classified public data can be transferred to the model to improve the performance of the model, while important and sensitive private data does not need to be transmitted through the network, thus effectively avoiding the problem of privacy leakage. Therefore, it is seen that our LSE-Yolo visual image system for smart cities and smart home is considerable, which can make people's life more convenient and comfortable. To sum up, the proposed neural network can perform real-time detection along with privacy data protection and sufficient data.

2.1. Network Architecture. In our LSE-Yolo neural network, the darknet53 is used as our backbone network as shown in Figure 2, which is the basic unit of our network. There are several advantages of darknet53. Firstly, it can be installed easily and just take a few minutes to finish installing. Secondly, it has no dependencies and does not rely on any libraries. Thirdly, its structure is clear, and the source code of the backbone network is easy to view and modify. Finally, it has python interface and is easy to transplant. The architecture of the darknet53 is shown in Table 1, which is mainly composed of a series of 1 × 1 and 3 × 3 convolution layers. It is seen that 53 convolution layers are contained in the darknet53, where the fully connected layer is regraded as convolution layer but the residual layer is not taken as

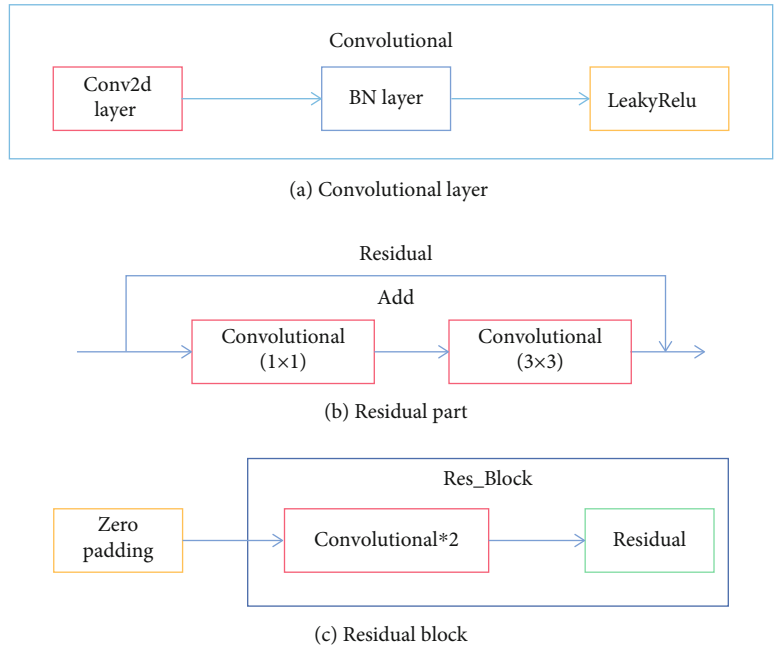

Figure 3: The structure of residual block.

convolution layer. It is found that there is a residual layer after two convolution carried out, in which the original information is retained and the extracted features are integrated by adding residuals.

As shown in Figure 3(a), each convolutional layer is followed by a normalization layer and an activation layer. In Figure 3(b), it is demonstrated that in the residual part, $1 \times 1 + 3 \times 3$ is used to deepen the network depth to enhance feature sampling. Therefore, in Figure 3(c), it can be regarded as residual block when two convolutional module and a residual module are included. Moreover, each residual block is preceded by a zero padding.

If we train a normal network using a standard optimization algorithm with no residuals, no shortcuts or jump connections, say gradient descent, or some other popular optimization algorithm, it is found that with the deepening of network depth, training errors will first decrease and then increase. In theory, as the network deepens, the performance of the network should be more sound. While in fact, for a normal network with no residual modules, as the network deepens, there are more and more training errors. Therefore, darknet53 has benefited from the residual module; it can help solve the problem of gradient vanishing and gradient explosion, allowing to train deeper networks while maintaining good performance. In our network, to address the loss problem, the Squeeze-and-Excitation module is added after each residual layer. Also, the FPN module is added to fuse the high-level features and low-level features. Meanwhile, supervised data augmentation is adopted in our network to address the problem of the insufficient data and edge computing to protect the privacy data.

2.2. Attention Mechanism SE Module. Visual attention mechanism is a special brain signal processing mechanism of human vision, which can quickly scan the global image to obtain the target area that needs to be focused on and greatly improves the efficiency and accuracy of visual information processing. And the attentional mechanism in deep learning is essentially similar to the human-selective visual attention mechanism. Its goal is also to select from the information that is more critical to the current task and goal. In our work, the Squeeze-and-Excitation (SE) module is adopted in our LSE-YOLO neural network, to solve the loss problem caused by different importance of different channels of feature map in the process of convolutional pooling. In Figure 4, the SE module is explained, where Squeeze part and Excitation part are contained. In the module, W and H represent the width and height of the feature map and C represents the number of channels. Therefore, the size of the input feature map is $W \times H \times C$. In the Squeeze part, the squeeze operation can be regarded as a global average pooling, where the feature map can be compressed into $1 \times 1 \times C$ vector. Then in the Excitation part, it consisted of two fully connected layers to perform Excitation operation, where SERatio is a scaling parameter that is intended to reduce the number of channels and thus the computation effort. In the first fully connected layer, there are $C * \text{SERatio}$ of neurons, where the input is $1 \times 1 \times C$ and the output is $1 \times 1 \times \text{C} \times \text{SERadio}$. In the second fully connected layer, the input is $1 \times 1 \times C \times \text{SERadio}$, and the output is $1 \times 1 \times C$. Finally, the scale operation was carried out to multiply the channel weights. The output of the parameters number and the calculated quantity is $2 \times \text{C} \times \text{C} \times \text{SERatio}$.

2.3. Neck Network FPN. In our work, the feature pyramid network (FPN) is added in the Neck part of LSE-Yolo to solve the multiscale problem in object detection, in which the performance of small object detection is greatly improved without increasing the calculation amount of the original model through simple network connection changes.

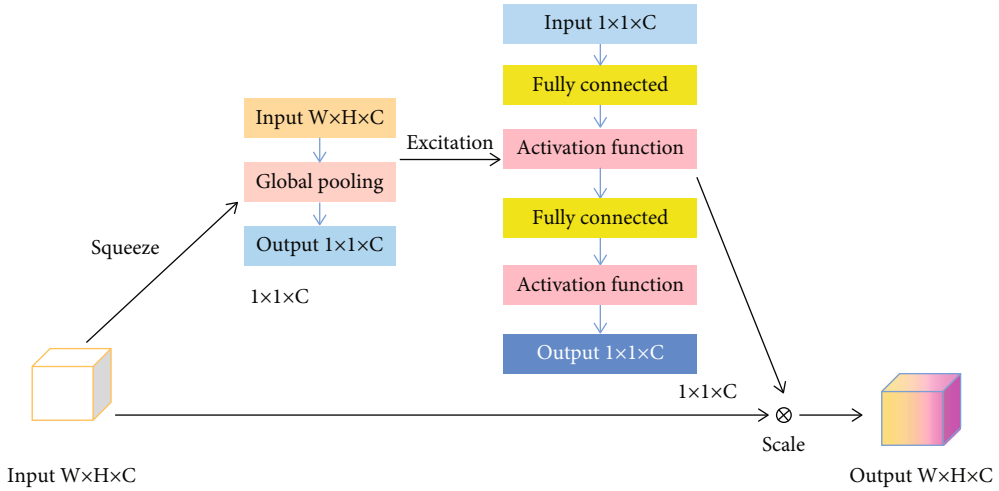

Figure 4: The architecture of SE module.

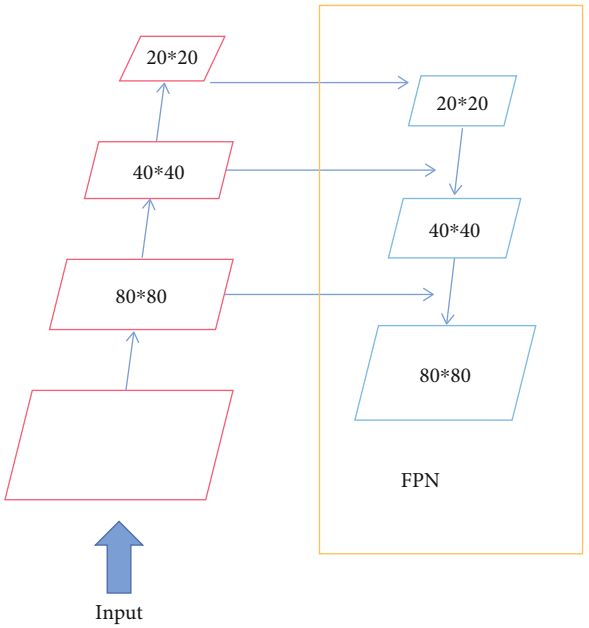

Figure 5: The architecture of FPN.

In Figure 5, the structure of FPN is demonstrated, upsampling by high-level features and top-down linking by low-level features, and predictions are made for each layer. In the bottom-up process, it is the common forward propagation process of neural network, and the feature map is usually smaller after the convolution kernel calculation. When in the top-down process, it is to carry out upsampling of more abstract and semantic high-level feature maps. And in horizontal connection, the result of upsampling is merged with the feature map of the same size generated from the bottom-up. The features of the two layers connected horizontally are the same in spatial dimensions to underlie position details. In addition, the horizontal connection can reduce the number of feature maps. In a word, the FPN added in the Neck part of our network can utilize both high resolution of low-level features and high semantic information of high-level features. And the prediction results can be achieved and performed separately at each fused feature layer through combining the features of these different layers.

2.4. Loss Function. In our work, the loss function in prediction box used is IOU loss and GIOU loss when BCE_loss is used in object loss and class loss.

As shown in Formula (1), it is explained that the IoU loss function and B^{gt} represent the target box when B is the prediction box. It can be used in regression tasks of bounding box due to the characteristic of nonnegative scale invariance, identity, symmetry, and triangle inequality.

$$L_{\mathrm{IoU}} = 1 - \frac{\left|B \cap B^{gt}\right|}{\left|B \cup B^{gt}\right|}. \tag{1}$$

The GIoU loss function is shown in Formula (2), where C represents the minimum bounding box for B and B^{gt}. It aims to solve the problem that the value of IoU loss is unified as 1 when there is no overlapping area between the detection box and the ground truth.

$$L_{\mathrm{GIoU}} = 1 - \mathrm{IoU} + \frac{\left|C - B \cap B^{gt}\right|}{|C|}. \tag{2}$$

When the calculation of mutual information is taken into account, BCE_loss which is suitable for classification is used in object loss and class loss as shown in Formula (3), which represents the loss corresponding to the N sample. ω is the super parameter and y is the actual label.

$$\mathrm{Loss}Xiyi = -\omega\mathrm{i}[yi \log xi + (1 - yi) \log (1 - xi)]. \tag{3}$$

2.5. Data Augmentation. In our work, data augmentation is used to reduce the overfitting of the network, and the training samples of MS COCO dataset are shown in Figure 6. By transforming the training images, a network with stronger generalization ability can be obtained, which can better

Figure 6: Samples of MS COCO dataset.

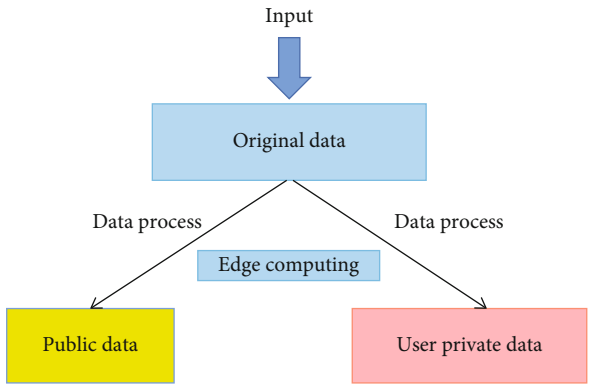

Figure 7: The architecture of data privacy protection.

Table 2: The comparison of AP and parameters (state-of-the-art) of different models for MSCOCO.

Methods	AP	Parameters
Yolov3+NMS	41.7	67.27 M
Yolov3 baseline	38.5	63.00 M
Yolov4 baseline	43	31 M
Yolov5 baseline	44.5	21.4 M
Ours	45.2	26.3 M

adapt to the application scenarios. It can be divided into supervised data augmentation and unsupervised data augmentation methods. In supervised data augmentation, it can be fall into single sample data augmentation and multiple data augmentation methods, while unsupervised data augmentation can be divided into two directions: generating new data and learning augmentation strategies. In our network, supervised data augmentation is adopted including single sample data augmentation and multiple data augmentation, in which the data can be enlarged on the basis of existing data. In single sample data augmentation, that is, when enhancing a sample, all operations are carried out around the sample itself, including geometric transformation classes and color transformation classes. Also, there are several common approaches to transform the image geometrically including flipping, rotation, clipping, deformation, scaling, and other operations. In addition, Mosaic and Mixup are also used as multiple data augmentation in our network, which contribute to the effect of the small object detection, the robustness, and the stability of the model.

2.6. Data Privacy Protection. In our work, the edge computing is added to protect the data privacy for smart cities and smart homes in Figure 7. It can be taken as an operation performed by using the edge strip near the data source, and the efficiency can be improved, namely, proximity computing. The cloud computing model that relies on a single centralized processing mode for the construction of smart cities or smart homes cannot cope with all the problems, and it needs the integration of multiple computing modes to solve the problems. There are several advantages of edge computing compared with other traditional cloud computing. Firstly, it can protect the localized data. The data in the Internet of Things is so vital to the users' lives that uploading it to a cloud computing center increases the risk of exposing users' private data. As shown in Figure 7, edge computing data are collected and calculated locally or on edge nodes without being uploaded to the cloud, and important and sensitive information does not need to be transmitted through the network, thus effectively avoiding the problem of privacy leakage. Secondly, it can reduce the cloud data to transfer and energy consumption. Thirdly, it can realize real-time computing.

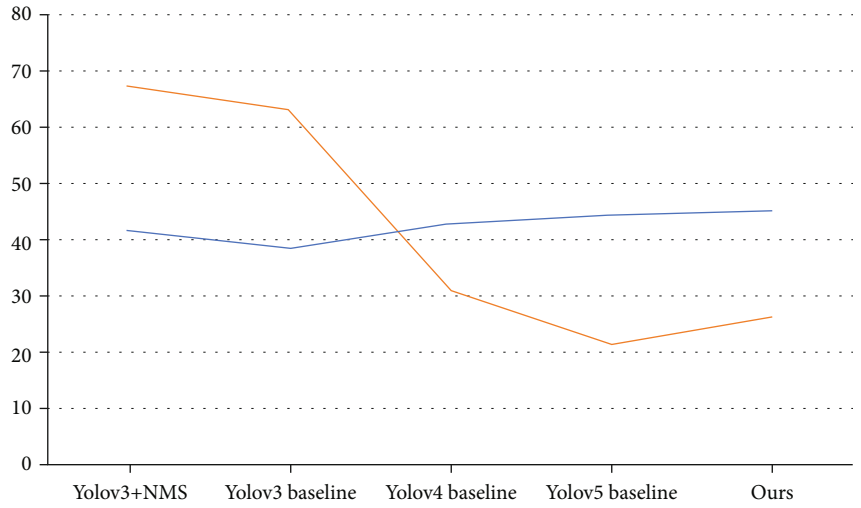

Figure 8: The line chart of the comparison of AP and parameters of different models.

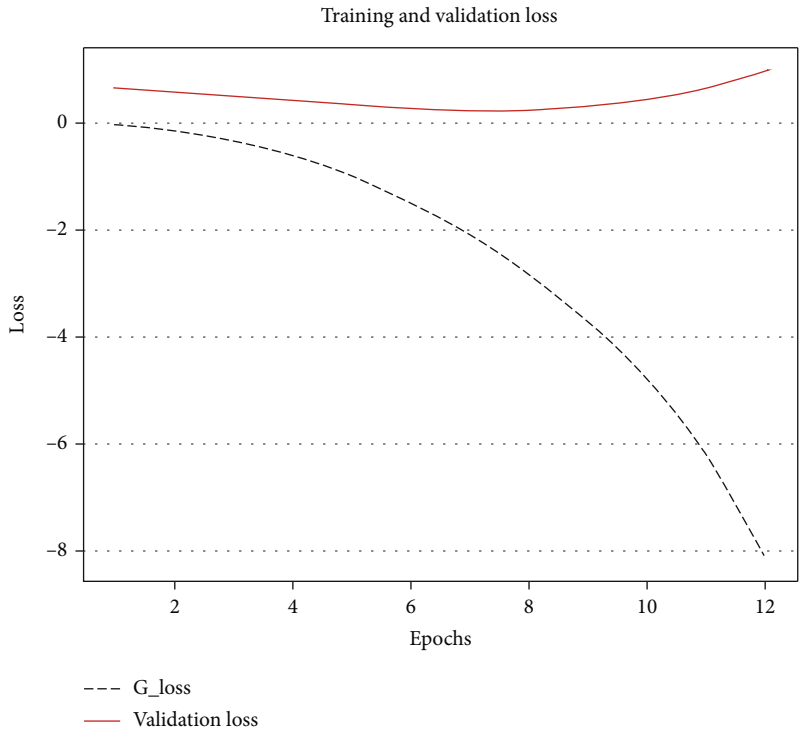

Figure 9: The training and validation loss.

3. Results and Discussion

In Table 2, the results obtained from the studies of the proposed neural network and other common neural network are shown. The experiment is based on Tesla p100, and the dataset is MSCOCO dataset. It is found that our proposed LSE-Yolo neural network in artificial intelligence-based Internet of Things for smart cities and smart homes has better performance than other common neural networks. In our work, the AP is 45.2, which is higher compared to the neural network, and the parameters is 26.3 M, which is lower than Yolov5. And the line chart can be shown in Figure 8. In Figure 9, the training and the validation loss are demonstrated. In a word, the comprehensive of the proposed LSE-Yolo is highly consistent with the prediction of the theoretical model, which is suitable for smart cities and smart homes. Despite the great advantages mentioned above, there are still some problems which have not been addressed in our proposed models. It counts that the tradeoff between the accuracy and the speed when applied in smart cities and homes is still a challenge. Therefore, it deserved to be further studied to solve the above problems.

4. Conclusions

In conclusion, it is stated that our proposed LSE-Yolo neural network in artificial intelligence-based internet of things for smart cities and smart homes is fruitful. It can not only bring us more healthy and comfortable life but also make great contributions to meeting the needs of the modern smart cities and smart homes. However, there are still some limitations in our study, which is how to realize the tradeoff between the accuracy and the speed of the model. In addition, it should be noted that the data for smart cities and smart homes is still a challenge. And 3D object deep detection method which should be developed and combined with augmented reality for smart cities and smart homes is still the problem to be solved. Therefore, the limitations above remain to be solved in the future study, which aims to make it more convenient and fast for the smart cities and smart homes.

Data Availability

The data we used is available and can be accessed to perform image detection system using neural network in artificial intelligence-based Internet of Things for smart cities and smart homes study. And part of them are available to you from the corresponding author upon request (zfq@mju.edu .cn).

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the study of this work and publication of this paper.

Acknowledgments

Our work is supported by the 2021 Fujian Middle-aged and Young Teacher Education Research Project (Science and Technology), Project Nos. JAT210719 and JAT210704; supported by Digital Media Art, Key Laboratory of Sichuan Province, Sichuan Conservatory of Music, Project No. 21DMAKL01; supported by the First Batch of Industry-University Cooperation Collaborative Education Project funded by the Ministry of Education of the People's Republic of China, 2021, Project No. 202101071001; supported by Minjiang College 2021 School-Level Scientific Research Project Funding, Project No. MYK21011.

References

[1] C. G. Garcia, D. Meana-Llorian, B. G-Bustelo, J. M. C. Lovelle, and N. Garcia-Fernandez, "Midgar: Detection of people through computer vision in the Internet of Things scenarios to improve the security in smart cities, smart towns, and smart homes," *Future Generation Computer Systems*, vol. 76, pp. 301–313, 2017.

[2] F. Mehmood, I. Ullah, S. Ahmad, and D. H. Kim, "Object detection mechanism based on deep learning algorithm using embedded IoT devices for smart home appliances control in CoT," *Journal of Ambient Intelligence and Humanized Computing*, pp. 1–17, 2019.

[3] S. Khan, S. Nazir, and H. U. Khan, "Smart object detection and home appliances control system in smart Cities," *Computers, Materials and Continua*, vol. 67, no. 1, pp. 895–915, 2021.

[4] J. Park, K. Jang, and S. B. Yang, "Deep neural networks for activity recognition with multi-sensor data in a smart home," in *2018 IEEE 4th World Forum on Internet of Things (WF-IoT)*, pp. 155–160, Singapore, 2018.

[5] R. Nayak, M. M. Behera, U. C. Pati, and S. K. Das, "Video-based real-time intrusion detection system using deep-learning for smart city applications," in *2019 IEEE International Conference on Advanced Networks and Telecommunications Systems (ANTS)*, pp. 1–6, Goa, India, 2019.

[6] F. Wang, Z. Xu, Z. Qiu, W. Ni, J. Li, and Y. L. Luo, "Algorithm for target detection in mart city combined with depth learning and feature extraction," *Wireless Communications and Mobile Computing*, vol. 2020, Article ID 8885670, 7 pages, 2020.

[7] L. Hu and Q. Ni, "IoT-driven automated object detection algorithm for urban surveillance systems in smart cities," *IEEE Internet of Things Journal*, vol. 5, no. 2, pp. 747–754, 2018.

[8] J. Xu, H. Wang, M. Xu, F. Yang, Y. Zhou, and X. Yang, "Feature-enhanced occlusion perception object detection for smart cities," *Wireless Communications and Mobile Computing*, vol. 2021, Article ID 5544194, 14 pages, 2021.

[9] P. Felzenszwalb, D. McAllester, and D. Ramanan, "A discriminatively trained, multiscale, deformable part model," in *2008 IEEE conference on computer vision and pattern recognition*, pp. 1–8, Anchorage, AK, USA, 2008.

[10] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: unified, real-time object detection," in *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 779–788, USA, 2016.

[11] J. Redmon and A. Farhadi, "YOLO9000: better, faster, stronger," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7263–7271, USA, 2017.

[12] J. Redmon and A. Farhadi, "YOLOv3: an incremental improvement," https://arxiv.org/abs/1804.02767.

[13] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE international conference on computer vision*, pp. 1440–1448, Chile, 2015 .

[14] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: towards real-time object detection 404 with region proposal networks," *Advances in Neural Information Processing Systems*, vol. 29, no. 6, 2015.

[15] L. Wang, X. Fan, J. Chen, J. Cheng, J. Tan, and X. Ma, "3D object detection based on sparse convolution neural network and feature fusion for autonomous driving in smart cities," *Sustainable Cities and Society*, vol. 54, article 102002, 2020.

[16] S. Mettupally and V. Menon, "A smart eco-system for parking detection using deep learning and big data analytics," in *2019 SoutheastCon*, pp. 1–4, Huntsville, AL, USA, 2019.