Received: 21 July 2022 | Revised: 5 October 2022 | Accepted: 3 November 2022

DOI: 10.1049/ipr2.12686

IET Image Processing

The Institution of Engineering and Technology

WILEY

ORIGINAL RESEARCH

Fast intensive crowd counting model of Internet of Things based on multi-scale attention mechanism

Dong Liu | Zhiyong Wang | Xiangjia Meng

School of Information Engineering, Shandong Youth University of Political Science, Jinan, China

Correspondence
Xiangjia Meng, School of Information Engineering, Shandong Youth University of Political Science, Jinan 250103, China.
Email: Mengxiangjia888@163.com

Funding information
Youth Innovation Science and Technology Support Program of Universities in Shandong Province, Grant/Award Number: 2021KJ082; Shandong statistical scientific research project, Grant/Award Number: KT037; Project of Shandong Youth University of Political Science, Grant/Award Number: SJQNXM202211; Shandong Provincial Department of housing and urban rural development, Grant/Award Number: 2020rkx02-01; Project of aging research base of Shandong Youth Political College of China Aging Association, Grant/Award Number: SDYULZ10

Abstract

Object detection based on deep learning plays an important role in the application of the Internet of Things (IoT). Traditional methods consume a lot of computing resources and cannot be well deployed in the IoT environment. A lightweight object detection method based on attention mechanism is proposed and applied to crowd counting. In view of the low accuracy and poor real-time performance of multi-scale crowd detection, we design a crowd counting model based on YOLO v5, and apply it to the IoT environment. It is proposed to insert the transformer into the YOLO v5 backbone network. Based on the multi-head attention mechanism in the transformer encoder, the global dependency is modelled to make full use of the context information. The CNN is used to realize the fusion of multi-scale feature maps, and the feature enhancement modules concerned by the attention network are further counted. Experiments show that it can not only detect multi-scale targets, but also achieve real-time performance in video surveillance scenes.

1 | INTRODUCTION

The statistical method of human flow is of great significance in the field of public security. In various scenes, estimating human flow and population density through images or video frames has become a key area of computer vision research [1–4]. Especially in the existing environment of the Internet of Things (IoT) [5, 6], various sensors and camera equipment jointly collect a large amount of people flow information. How to effectively use these multi-modal data to achieve a more accurate estimation of the number of people in images or video frames, while also tracking in time, is a challenge of current intelligent counting [7]. The following five levels of crowd density are usually used to evaluate the degree of congestion: very high density, high density, medium density, low density and very low density [8]. Because crowd counting methods have important applications in many research fields, many researchers have studied these methods. In recent years, significant progress has been made in the research of people flow statistics based on deep learning, especially the object detection method using convolutional neural network (CNN) has achieved amazing results in pedestrian detection [9–11]. However, due to the challenging scale changes and complex scenes in the crowd, it is necessary to recognize the correlation information between frames while detecting the category and location information of the target, so as to ensure that the same person in the video will not be recognized and counted many times. The traditional convolution network cannot handle this task well [12–15].

Especially in recent years, the integrated development of IoT+ AI technology has provided a broad market and a wide range of use scenarios for the application of deep learning algorithms in the IoT environment [16, 17]. The IoT and artificial intelligence are complementary to a great extent [18]. With the help of its sensors, IoT devices collect a large amount of data during operation. Artificial intelligence can use these sensors to improve the performance of the device. Relevant statistics show that IOT devices generate 2.5 trillion bytes of data every day. Through artificial intelligence optimization, we can make full

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2022 The Authors. *IET Image Processing* published by John Wiley & Sons Ltd on behalf of The Institution of Engineering and Technology.

use of deep learning algorithm to realize scheduling, analysis, prediction and processing of these big data. Thereby providing valuable decision support [19–21].

From the above analysis, we can see that there are two problems in the task of pedestrian flow statistics: One is the problem of pedestrian object detection, that is, the accurate positioning and real-time tracking of pedestrian targets; the second is the overlapping occlusion of pedestrians in complex crowded scenes [22, 23].

Aiming at these problems, the major contributions of this paper are:

We propose an object detection method of YOLO V5 framework, and focuses on the statistical model of YOLO V5 traffic based on transformer. The model adopts transformer structure, which learns adaptive feature representation through multi-scale sampling location and dynamic attention weight. Then, the convolution neural network is used to realize multi-level feature fusion, and the feature enhancement module concerned by the attention network is further counted [24]. Through the CNN, the receptive field of feature learning is increased to enhance the significant feature representation in the global image background [25]. The learned representations can focus on the future and adapt to different sizes of people. The feasibility of the scheme is verified by comparative experiments. After the training of YOLO V5, the head recognition feature map generated by convolution neural network model has good classification effect.

2 | RELATED WORK

Object detection is a basic problem in computer vision, and it is also an important part of many applications such as autonomous driving, visual search and target tracking. In view of its large-scale and real-time applications, scalable training and fast reasoning are crucial. Although the traditional method using deep neural network (DNN) is very powerful in visual recognition, the computational cost may be high. In addition, in dealing with complex scenes, deep neural networks have shortcomings such as lack of scale invariance and inaccurate prediction, which may affect detection. This paper focuses on the deep learning method for the object detection of human head, and introduces the attention mechanism, focusing on the feature information of human head, so as to avoid the inaccurate counting caused by overlap. The intelligent human traffic statistics based on deep learning method studied here mainly includes two key technologies: feature learning and object detection.

2.1 | Feature learning

Using depth information to solve the lack of traditional computer vision feature extraction and learning ability has become a hot topic. Since 2006, Bengio, Hinton and others have pushed the DNN to a new height. Lenet [26] network is the first convolutional neural network model proposed. Its structure is simple, but it includes all the basic modules of convolutional network, namely, convolution layer, pooling layer, and full connection layer. It is the basis of other deep learning models. Because it was proposed for a long time, the performance of computer hardware was low at that time, and there was a lack of large-scale training data, the effect of Lenet network in dealing with complex problems was not ideal. In 2012, Hinton et al. Built Alexnet [27] network, applied the basic principle of CNN to a very wide and deep network, and won the championship in Imagenet image recognition competition; In 2015, He et al. built residualnetwork (resnet [28]), which solved the problems of gradient disappearance becoming more obvious and computational complexity rising rapidly when Alexnet network depth gradually increased, so that the depth of the network was greatly expanded, and even reached 1000 layers. From this depth, the research of CNN has achieved rapid development.

2.2 | Object detection

Crowd counting involves pedestrian detection and pedestrian tracking, so object detection method is the core step of the task. In order to effectively detect the head of pedestrians in a complex and crowded environment, the accuracy of object detection is very important. Traditional object detection methods generally include three stages: Region selection, feature extraction and classification.

2.2.1 | Region selection

Locate the target location, set different scales and aspect ratios, and then use the sliding window strategy to traverse the whole image. The disadvantages of this exhaustive method are obvious, that is, it produces a large number of redundant windows, and the time complexity is very high. At the same time, it greatly affects the speed and performance of subsequent feature extraction and classification prediction.

The strategy of manual extraction is usually adopted, that is, manually designing feature extraction algorithm to recognize the target. Due to the variability of scene and the diversity of target posture, even the changes of image colour and saturation caused by illumination changes, it is difficult for the artificially designed feature extraction algorithm to have strong versatility, and the design of this algorithm also requires rich engineering experience. The commonly used features at this stage include sift [29], hog [30] etc.

2.2.2 | Classification

Classify the target according to the features extracted. The classifiers mainly include SVM, Adaboost etc. The quality of classification results directly depends on the accuracy and confidence of feature extraction in the second step.

From this point of view, the traditional detection algorithm sliding window strategy has high time complexity, and the robustness and versatility of the feature extractor are not strong

17519667, 0, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/ipr2.12686 by National Institutes Of Health Malaysia, Wiley Online Library on [07/06/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

enough, so it cannot meet the object detection requirements of traffic statistics under complex scenes.

The above object detection methods have the following problems: (1) The feature extraction networks are mostly designed for image classification tasks and are not optimized for object detection task pairs, resulting in the lack of robustness of the algorithms. (2) Extending the network depth by stacking convolutional modules, although good detection results are obtained, it is difficult to achieve real-time performance. (3) Insufficient performance for the human head small-scale object detection task.

With the development of artificial intelligence technology, researchers began to study candidate region based neural network object detection. the regional convolutional neural network (R-CNN) proposed by Girshick et al. first extracts candidate regions from images, and then adjusts and classifies these regions by standard CNN. However, there are a lot of repeated computations in the recognition process and the speed of the object detection network is not ideal. In 2015, Ren and Girshick proposed a faster region convolutional neural network (faster R-CNN), which uses a deep learning-based region proposal network (RPN) to extract target candidate regions, which greatly improves the computational speed of the model [31]. However, it still cannot meet the requirements of real-time detection. In 2016, Redmon et al [32] proposed the YOLO (you only look once) network, which takes the whole image to be detected as input and CNN as regressor to return the position information of the target in the image to be detected to achieve end-to-end object detection and recognition. In May 2020, the Ultralytics proposed a composite scale model YOLO V5 network based on YOLO V3 and YOLO V4. It has significantly improved the detection speed and model size, and meets the real-time requirements based on easy deployment. The core of YOLO algorithm is to skip the classifier to complete the object detection, and directly use the neural network to predict the envelope box and category by global information to achieve end-to-end object detection. On this basis, by drawing on the design idea of Transformer algorithm, we propose a method to improve YOLO v5 based on Transformer to improve crowd counting.

3 | THE PROPOSED METHOD

In order to obtain higher statistical accuracy for dense human flow with mutual occlusion, this paper adopts the object detection method of YOLO v5, which is ideal for both tracking speed and detection accuracy. The traditional feature extraction is based on the manual design of feature extractor, which is generally based on some statistical laws and the designer's own a priori knowledge, and cannot fully extract the information of the original image, and thus has a weak generalization performance and robustness. Transformer automatically extracts the deeper structure and features of the image by the learning method, and adds it to the YOLO backbone network, increasing the local perceptual field, sparse weights and the concept of parameter sharing, making it somewhat translation and scale invariant and more suitable for the learning of data with 2D structure like images.

Deep learning-based object detection can be divided into Two-Stage object detection and One-Stage object detection. The inference speed of Two-Stage strategy is low because the intermediate layer is used to propose possible target regions. The region suggestion layer extracts the target regions in the first stage. In the second stage, these proposed regions are used for classification and bounding box regression. On the other hand, the One-Stage strategy can predict all the bounding boxes and class probabilities in one inference with a higher inference speed. This makes One-Stage object detection more suitable for real-time applications. In this paper, we focus on YOLO-based object detection based on its application to foot traffic statistics.

3.1 | YOLO v5

YOLO v5 is a single-stage object detection model, which usually contains three main components, that is, backbone, neck and head, where the backbone is a CNN that can receive images of different sizes and form the overall features of the image, the neck represents a series of network layers that can fuse the image features extracted from the backbone according to certain laws to make the feature semantic information richer and use the processed features as the input output of the prediction layer, and finally, the head predicts the input features and the classifier obtains the object class and generates the final coordinates of the bounding box.

YOLO v5 currently has significant advantages in terms of speed and accuracy, which includes important modules such as Focus, Cross Stage Partial (CSP) Bottleneck, Spatial Pyramid pooling (SPP) and Path Aggregation Network (PANet). The backbone is mainly composed of CSP Darknet53, which is mainly composed of five layers of residual network resblock_body, whose input image pixels are 608 × 608, where resblock_body has a special convolution operation to reduce the resolution, and each layer of resblock_body gradually reduces the pixels by twice, and its main function is to extract feature information of the image data. YOLO v5 uses PANet as the neck of the model, the input is the feature mapping of the backbone output, and feature fusion is performed on it to obtain features with richer semantic information, which are sent to the head for detection. In terms of feature extraction, the Feature Pyramid Network (FPN)-based PANet structure is improved to convey not only semantic information but also location information. A bottom-up pyramid is added to the FPN structure to pass strong localization features from the bottom layer to the top layer to complement the FPN feature fusion.

3.2 | Transformer

Transformer contains the basic encoding and decoding process. The Transformer is fused into the feature extraction network for feature extraction in object detection to achieve accurate

17519667, 0, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/ipr2.12686 by National Institutes Of Health Malaysia, Wiley Online Library on [07/06/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

prediction of target location and its class. Here, only the encoder part of Transformer is used:

1. Add the corresponding position encoding to each embedding vector of the input, which can better express the relationship between the feature vectors at different positions in the later calculation.
2. The feature matrix encoded by the position is sent to the multi-head attention layer to compute the attention values in parallel, and the one-dimensional feature matrix X is obtained after the full extraction of features.
3. The feature matrix X obtained from the multi-head attention layer is fed into Add&Norm to prevent degradation of the neural network during training and to improve the stability of training by normalization through residual connectivity.
4. The feature matrix from the Add&Norm layer is fed into the fully connected layer to map the features to a higher dimensional space, which is then filtered by the ReLU activation function, and then the feature matrix is changed back to its original dimension after the screening. Finally, the output of Encoder is obtained after the Add&Norm layer.

The decoding process is similar to the encoding process, with the difference that the decoding adds a layer of masked multi-head attention layer on top of the encoding, which is used to mask the values for which the information is currently unavailable.

3.3 | The proposed model

Many methods have been proposed to mimic the human reasoning ability in object detection. On the other hand, most of these methods are complex and use the Two-Stage detection architecture. Therefore, they are not applicable to real-time applications. The current One-Stage object detection processes each image region individually. When considering the image size, the smaller perceptual field causes them to be unaware of the different image regions. They rely entirely on high-quality local convolutional features to detect targets. However, this is not the way the human visual system works. Humans have a "Reasoning" ability to perform visual tasks with the help of acquired knowledge.

Improvement ideas for the backbone network, as shown in Figure 1.

1. Here a new approach of incorporating visual REASONING into One-Stage object detection is used to integrate the Multi-Head Attention-based REASONING layer into the backbone's Through this approach, more meaningful, fine-grained and enhanced feature mapping can be used to extract information about the relationships between different image regions by using In this way, reasoning information about the relationship between different image regions can be extracted by using more meaningful, finer-grained and enhanced feature mapping.
2. Transformer extracts global image features, CNN extracts subtle human head local variation features, and global and local features are fused. First, the Transformer's multi-head attention mechanism models global dependencies to make full use of contextual information, and then, CNN is used to implement multi-scale feature map fusion, and further statistics of the feature enhancement module of attention network attention is used to increase the perceptual field of feature learning by CNN to enhance the salient feature representation in the global image context and obtain more detailed features.

The specific improvement methods:

1. Add Transformer encoder to its backbone network part, and each Transformer encoder block contains two sub-layers. The first sublayer is a multi-head attention layer and the second sublayer Multilayer Perceptron (MLP) is a fully connected layer. The Transformer encoder block adds the ability to capture contextual information of different global features. This makes the recognition of human head and face more accurate in complex scenes, especially when facing overlapping situations.
2. Increase feature fusion strategy. As the depth of the network deepens, the feature information extracted by the shallow network and the deep network differs greatly. The shallow network tends to extract features about texture, colour, and edges, with more detailed representation of object contours; the features extracted by the deep network are more abstract, with high-level semantic information, but lacking shallow detailed feature information. The deep semantic information extracted by the Transformer is fused with the shallow features extracted by the convolutional neural network of the YOLO prediction layer, and finally the three different features resulting from the fusion are classified and output: Both category and location information is included.

3.4 | Definition of loss function

The loss functions of the IoU class are all based on the IoU (intersection over union) between the prediction frame and the labelled frame. IoU is used to evaluate the importance of the prediction frame. in order to facilitate the definition of the loss function in this paper, IoU is introduced here first. the prediction frame is judged to be valid according to the ratio of the area of the intersection merge set of the labelled frame and the prediction frame as Equation (1).

$$IoU = \frac{area(G \cap P)}{area(G \cap P)} \tag{1}$$

where G (Ground box) denotes the label box, P (Prediction box) denotes the prediction box, and area denotes the area of the box. In the experiments of this chapter, IoU takes 0.5 as the threshold value, when IOU is greater than or equal to 0.5, it means that the

17519667, 0, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/ipr2.12686 by National Institutes Of Health Malaysia, Wiley Online Library on [07/06/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

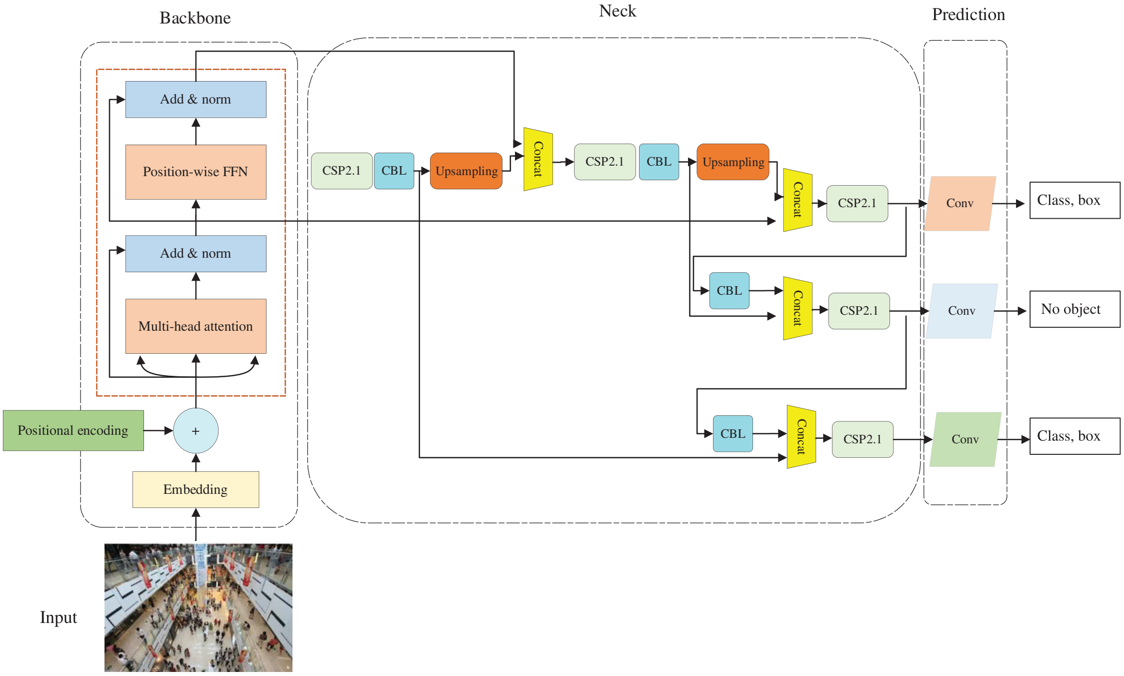

FIGURE 1 The overall architecture of the proposed model

prediction box is valid, and when *IoU* is less than 0.5, it means that the prediction box is invalid. From the formula, the value of *IoU* is between [0,1], and the closer the value of *IoU* is to 1, the better, the closer it is to 1, the more the prediction box overlaps with the label, and the more accurate the positioning is; conversely, the less the prediction box and the label box overlap, the worse the positioning is.

In a large-scale crowd counting scenario, the centres of the boxes should be as close as possible in order to better detect and count the people in overlapping situations, so the loss function *L* will have the following definition.

$$L = 1 - IoU + R(P, G) = 1 - IoU + \frac{\rho^2(p, g)}{c^2} \qquad (2)$$

where ρ represents the distance between the central end of the prediction box and the dimension box, and *P* and *G* are the central points of the two boxes. *C* represents the diagonal length of the minimum bounding rectangle of two boxes. When the distance between the two frames is infinite, the distance between the centre point and the diagonal length of the circumscribed rectangular frame approach infinitely, $R \rightarrow 1$.

4 | EXPERIMENT AND ANALYSIS

4.1 | Experimental environment

The algorithms were trained and tested on an intel Core i7-12700k CPU@3.7 GHz, 32G RAM, GeForce GTX 2080 Ti 8G platform for algorithm performance. The operating system was Ubuntu 16.04, and the supporting environment was: CUDA10.0, CUDNN9.2, Python3.7.2. keras = 2.1.3, pytorch-gpu = 1.4.0, opencv-python = 4.4.0.

TABLE 1 Classification of object detection

Ground truth		
Prediction	**Positive**	**Negative**
Positive	TP (True Positive)	FN (False Negative)
Negative	FP (False Positive)	TN (True Negative)

4.2 | Evaluation

Here, the improved algorithm is applied to a dense footfall counting scenario for pedestrian object detection and counting. Recall, Precision and PR curve are important evaluation metrics in the field of object detection. Before introducing the evaluation indicators, first introduce the classification of object detection. Generally, the confusion matrix is used to represent the classification of target detection, as shown in Table 1.

True Positive (TP): When the Machine Learning model correctly predicts the condition, it is said to have a True Positive value.

True Negative (TN): When the Machine Learning model correctly predicts the negative condition or class, then it is said to have a True Negative value.

False Positive (FP): When the Machine Learning model incorrectly predicts a negative class or condition, then it is said to have a False Positive value.

False Negative (FN): When the Machine Learning model incorrectly predicts a positive class or condition, then it is said

17519667, 0, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/ipr2.12686 by National Institutes Of Health Malaysia, Wiley Online Library on [07/06/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

to have a False Negative value.

$$\mathrm{Re}call = \frac{TP}{TP+FN} \tag{3}$$

$$\mathrm{Pr}\,ecision = \frac{TP}{TP+FP} \tag{4}$$

Recall is the proportion of samples with positive predictions to the number of samples with positive predictions. Accuracy represents the proportion of samples with positive predictions to the number of samples with positive predictions, and is specific to the prediction results.

In order to better assess the effectiveness of the detection and evaluate the performance and robustness of the proposed model, this paper uses the mean average precision (MAP), the most commonly used evaluation metric for each category, to assess the effectiveness of the model's detection results. The average precision (AP) can be calculated from the area enclosed by the curve between precision and recall, and then the average of all APs is calculated to obtain mAP, which is defined by (5) and (6), where C represents the number of categories.

$$AP_i = \int_0^1 p_i(r)dr \tag{5}$$

$$mAP = \frac{1}{C}\sum_{i=1}^{C} AP_i \tag{6}$$

4.3 | Datasets

4.3.1 | The MS COCO dataset [33]

Deep learning-based detection algorithms require a larger sample base, and a trained network weighting model will be more accurate if the sample base is large and wide-ranging. For deep learning detection tasks, the YOLO series provides pre-trained networks on the MS COCO dataset, a Microsoft-built dataset containing detection, segmentation, keypoints etc. MSCOCO is primarily designed to address the problem of detecting non-iconic views of objects (which corresponds to common views). The MSCOCO dataset is designed to solve the problems of detecting non-iconic views of objects, contextual reasoning between objects and the precise 2D localization of objects (which is often referred to as the segmentation problem).

4.3.2 | VOC [34]

The VOC dataset is one of the most commonly used standard datasets in the field of target detection and can be divided into four main categories and 20 sub-categories, containing three folders: JPEGImages, ImageSets and Annotation. The Annotation folder holds the annotation information for each image and is stored in xml format.

4.3.3 | ImageNet dataset [35]

This dataset is a computer vision dataset created by Professor Feifei Li of Stanford University. The dataset consists of 14,197,122 images and 21,841 Synset indexes. Synset is a node in the WordNet hierarchy, which is in turn a collection of synonyms. Similar to the classification data, the localisation task has 1000 categories. Accuracy is calculated based on the top five detections. All images have at least one edge in them. The detection problem for 200 targets has 470,000 images, with an average of 1.1 targets per image.

Here, during the training process, the YOLO core framework is tested in real application scenarios, pre-training is performed through the ImageNet dataset to obtain the classification network, and the training process model is trained on the MSCOCO dataset or VOC dataset to obtain it.

4.4 | Comparative analysis

The crowd counting task requires the detection of the class and location information of the target while identifying the correlation information between frames to ensure that the same person in the video is not identified and counted multiple times. Different training and validation methods were performed for different camera angles (flat or top angle) and for the sparsity of people.

For scenes where people are relatively sparse: Full body detection and tracking of pedestrians in the scene. As shown in Figure 2, the model identifies the pedestrians detected in the scene and displays the number of pedestrians in the scene in the top left corner, enabling footfall statistics.

For relatively dense scenarios: As shown in Figure 3, in relatively dense scenarios, the problem of occlusion between people can be very serious, and this can lead to a higher rate of missed detection if the overall detection of pedestrians is chosen. Therefore, the head-tracking method is used in this scenario to detect and track the heads of the pedestrians in the scene, and the statistics of the pedestrian flow are counted based on the detected heads. Experiments and analysis of head counting in a dense state will follow.

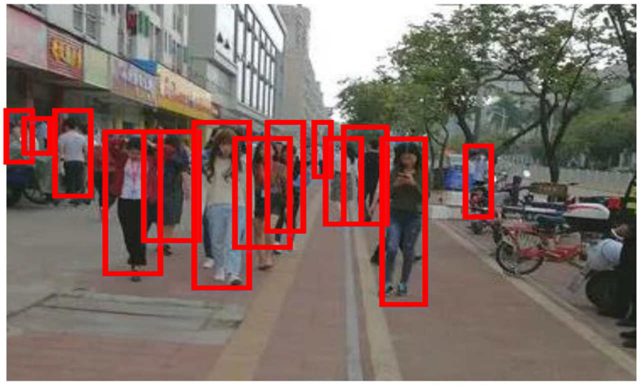

FIGURE 2 Pedestrian detection in sparse scenes

17519667, 0, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/ipr2.12686 by National Institutes Of Health Malaysia, Wiley Online Library on [07/06/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

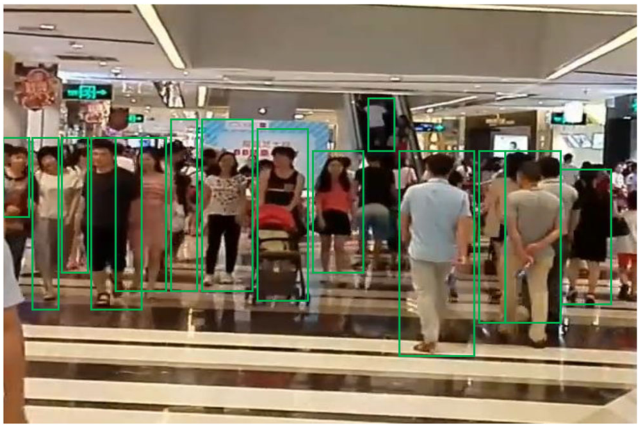

FIGURE 3 Pedestrian detection in dense scenes

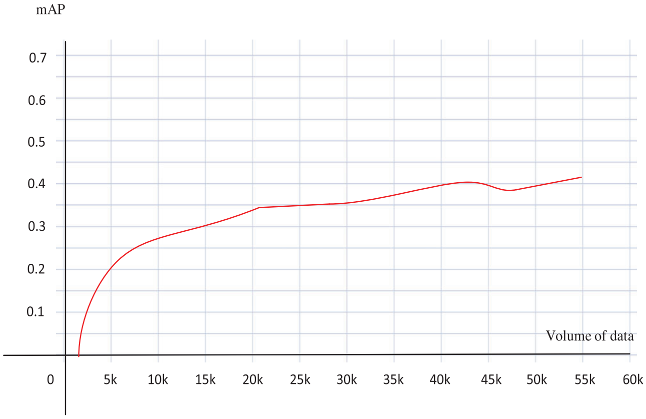

FIGURE 4 mAP curve of the baseline system

4.4.1 | Performance under overlapping heads

In this experiment, we use the YOLO v5 baseline system with the improved model for performance comparison experiments. The two methods were trained on three GPUs respectively, and a total of 200 epochs were trained in each experiment. The initial learning rate was set to 0.001, and the learning rate was tuned to 0.01 at 50 epochs and 0.1 at 150 epochs respectively.

The highest mAP of the baseline system was 0.41, while the highest mAP of the improved method on the validation set was 0.63, indicating that the model showed better detection capability even in dense situations with overlapping heads. The performance test results are shown in Figures 4 and 5.

Under the constraint of the loss function, the improved method improves the performance of the head detection algorithm by training the pre-trained model on the ImageNet dataset, and then improving the model to fine-tune the pre-trained network model, making the convergence of the network much better and improving the detection results to a certain extent. The previous target detection experiments show that the improved method works better for pedestrian target detection, and this experiment effectively reduces the false detection and leakage of the baseline model detection when the pedestrian density is high through the head detection of the pre-trained model.

TABLE 2 Performance comparison at visual angle 0°

Model	Recall (%)	Error rate (%)	Computing time (min)
Faster R-CNN	88.1	7.3	75
YOLO	80.3	10.3	68
YOLO v4	81.5	11.2	67
YOLO v5	81.6	8.1	65
Proposed method	89.3	7.4	69
Proposed method+pre-train	**90.8**	**5.3**	62

TABLE 3 Performance comparison at visual angle 45°

Model	Recall (%)	Error rate (%)	Computing time (min)
Faster R-CNN	89.0	6.5	76
YOLO	81.6	9.7	70
YOLO v4	82.1	9.6	71
YOLO v5	84.8	7.3	69
Proposed method	91.0	6.1	68
Proposed method+pre-train	**92.1**	**3.7**	60

4.4.2 | Performance test under different visual angles

We collect monitoring equipment data from different visual angles (the angle between the monitoring equipment and the ground) to verify the recognition effect of the algorithm in different angles. The comparison indicators adopted are:

1. Error rate: As long as a target is detected on a picture without a target (regardless of the number of frames), the picture is considered as false detection. The proportion of the pictures that are erroneously detected in the batch of no target pictures is the error rate of the picture level.
2. Recall rate: As long as a target is detected on a picture with a target (regardless of the number of frames), the picture is considered to be recalled. The proportion of the recalled pictures in the batch target pictures is the picture level recall rate.

In the experiment, the algorithm proposed here is compared with Faster R-CNN, YOLO v5, YOLO v4 and YOLO base model. The results are shown in Table 2–4.

According to the above experimental data, on the premise of ensuring the computational efficiency, our scheme reduces the error rate and improves the recall rate. Yolo v5 model with backbone as transformer is adopted. By using image pre-training strategy, 92.1% recall rate and 3.7% error rate are achieved.

17519667, 0, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/ipr2.12686 by National Institutes Of Health Malaysia, Wiley Online Library on [07/06/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

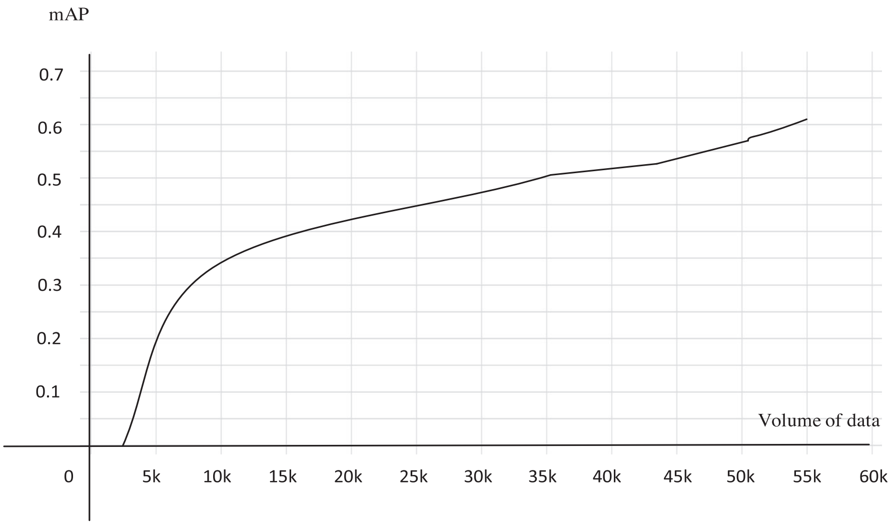

FIGURE 5 mAP curve of the proposed method

TABLE 4 Performance comparison at visual angle 90°

Model	Recall (%)	Error rate (%)	Computing time (min)
Faster R-CNN	**89.6**	6.1	73
YOLO	81.3	9.0	71
YOLO v4	83.1	8.9	71
YOLO v5	83.9	7.5	68
Proposed method	88.1	6.3	63
Proposed method+pre-train	88.9	**4.5**	61

5 | DEPLOYMENT

The prototype system is deployed in the IoT environment based on the edge cloud hybrid architecture. As shown in Figure 6, the user realizes video collection in public places through the front-end monitoring and collection equipment, stores the original multimedia files in the data centre, and the data interacts with the cloud platform through the Internet of things gateway. The cloud platform provides AI computing interface, which includes open-source API and customized SDK to facilitate access by the deep learning platform. The algorithm of the deep learning platform completes the task calculation, and finally outputs the traffic statistics results, traffic prediction results and abnormal situation warnings to the cloud computing platform according to the actual needs to provide decision support for users.

6 | CONCLUSION

This method builds on YOLO v5, adapts and optimises the structure of YOLO v5, and combines the Transformer structure with a convolutional neural network to solve the problem of inaccurate pedestrian traffic counting in the dense pedestrian occlusion scenario described above. The Transformer module

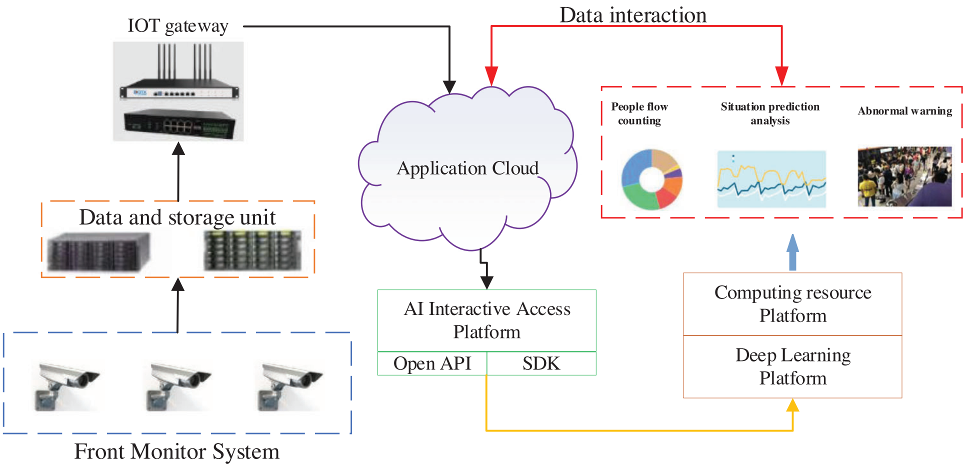

FIGURE 6 People flow counting and monitoring system based on IoT

17519667, 0, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/ipr2.12686 by National Institutes Of Health Malaysia, Wiley Online Library on [07/06/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

is inserted into the backbone network and neck detection layer to obtain target context information for solving the target The pre-trained model is fine-tuned to achieve detection of small targets, that is, accurate detection of overlapping head counts, thus achieving more accurate footfall statistics. In the future, we will try to deploy the trained models in lightweight devices using a migration learning approach.

AUTHOR CONTRIBUTION

D.L.: Conceptualization, Formal Analysis, Methodology, Funding Acquisition, Writing-Original Draft Preparation. Z.W.: Data Curation, Resources, Software, Writing-Review and Editing. X.M.: Data Curation, Funding Acquisition, Investigation, Project Administration.

ACKNOWLEDGEMENTS

This work is supported by the Youth Innovation Science and Technology Support Program of Universities in Shandong Province, China(Grant No.2021KJ082), and Shandong statistical scientific research project (Grant No. KT037), Project of Shandong Youth University of Political Science (Grant No. SJQNXM202211), and Soft science research project of Shandong Provincial Department of housing and urban rural development (Grant No. 2020rkx02-01), and Project of aging research base of Shandong Youth Political College of China Aging Association (Grant No. SDYULZ10).

CONFLICT OF INTEREST

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

DATA AVAILABILITY STATEMENT

The experimental data sets are publicly available and can be downloaded from the link:
MS COCO: http://cocodataset.org
VOC: http://host.robots.ox.ac.uk/pascal/VOC/
ImageNet: https://image-net.org/

ORCID

Dong Liu https://orcid.org/0000-0003-4875-0882
Zhiyong Wang https://orcid.org/0000-0001-5544-7385
Xiangjia Meng https://orcid.org/0000-0001-6742-3532

REFERENCES

1. Uijlings, J.R.R., Van De Sande, K.E.A., Gevers, T., et al.: Selective search for object recognition. Int. J. Comput. Vision 104(2), 154–171 (2013)
2. Felzenszwalb, P.F., Girshick, R.B., McAllester, D., et al.: Object detection with discriminatively trained part-based models. IEEE Trans. Pattern Anal. Mach. Intell. 32(9), 1627–1645 (2009)
3. Redmon, J., Divvala, S., Girshick, R., et al.: You only look once: Unified, real-time object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, pp. 779–788 (2016)
4. Carion, N., Massa, F., Synnaeve, G., et al.: End-to-end object detection with transformers. In: European Conference on Computer Vision, Glasgow, pp. 213–229 (2020)
5. Li, Z., Gou, F., De, Q., Ding, L., Zhang, Y., Cai, Y.: RealNet: Combining optimized object detection with information fusion depth estimation co-design method on IoT [J/OL]. https://arxiv.org/abs/2204.11216 (2022)
6. Kuzmic, J., Brinkmann, P., Rudolph, G.: Real-time object detection with Intel NCS2 on hardware with limited resources for low-power IoT devices (2022)
7. Liu, H., Wu, C., Li, C., Zuo, Y.: Fast robust fuzzy clustering based on bipartite graphfor hyper-spectral image classification. IET Image Process. 16, 3634–3647 (2022)
8. Barron, J.T.: A general and adaptive robust loss function. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, pp. 4326–4334 (2019)
9. Chan, A.B., Vasconcelos, N.: Counting people with low-level features and Bayesian regression. IEEE Trans. Image Process. 21(4), 2160–2177 (2012)
10. Li, M., Zhang, Z., Huang, K., et al: Estimating the number of people in crowded scenes by mid based foreground segmentation and head-shoulder detection. In: IEEE Conference on Pattern Recognition (ICPR), Tampa, Florida, USA, pp. 1–4 (2008)
11. Felzenszwalb, P.F., Girshick, R.B., McAllester, D., et al.: Object detection with discriminatively trained part-based models. IEEE Trans. Pattern Anal. Mach. Intell. 32(9), 1627–1645 (2010)
12. Shitrit, H.B., Berclaz, J., Fleuret, F., et al.: Multi-commodity network flow for tracking multiple people. IEEE Trans. Pattern Anal. Mach. Intell. 36(8), 1614–1627 (2014)
13. Chen, C.-H., Chen, T.-Y., Wang, D.-J., et al.: A cost-effective people-counter for a crowd of moving people based on two-stage segmentation. J. Inf. Hiding Multimedia Signal Process. 3(1), 12–23 (2012)
14. Huang, S., Li, X., Zhang, Z., et al.: Body structure aware deep crowdcounting. IEEE Trans. Image Process. 27(3), 1049–1059 (2018)
15. Wang, Y., Zhang, W., Liu, Y., et al.: Multi-density map fusion network forcrowd counting. Neurocomputing 397, 31–38 (2020)
16. Hameed, A., Violos, J., Leivadeas, A.: A deep learning approach for IoT traffic multi-classification in a smart-city scenario. IEEE Access 10, 21193–21210 (2022)
17. Saeik, F., Avgeris, M., Spatharakis, D., Santi, N., Dechouniotis, D., Violos, J., et al.: Task offloading in edge and cloud computing: A survey on mathematical artificial intelligence and control theory solutions. Comput. Network 195, 108177 (2021)
18. Mukhopadhyay, S.C., Suryadevara, N.K.: Internet of Things: Challenges and Opportunities in Internet of Things. Springer, Berlin (2014)
19. Sivanathan, A., Gharakheili, H.H., Loi, F., Radford, A., Wijenayake, C., Vishwanath, A., et al.:Classifying IoT devices in smart environments using network traffic characteristics. IEEE Trans. Mobile Comput. 18, 1745–1759 (2019)
20. Liu, Q., Chen, L., Jiang, H., Wu, J., Wang, T., Peng, T., Wang, G.: A collaborative deep learning microservice for backdoor defenses in Industrial IoT networks. Ad Hoc Networks 124, 102727 (2022)
21. Deepa, S.N.: Intelligent ubiquitous computing model for energy optimization of cloud IOTs in sensor networks. Int. J. Pervasive Comput. Commun. 18(1), 18–42 (2022)
22. Chen, W.T., Fang, H.Y., Ding, J.J., Kuo, S.Y.: Pmhld: Patch map-based hybrid learning dehazenet for single image haze removal. IEEE Trans. Image Process. 29, 6773–6788 (2020)
23. Bochkovskiy, A., Wang, C.Y., Liao, H.: YOLOv4: Optimal speed and accuracy of object detection. arXiv:2004.10934 (2020)
24. Liu, W., Yuan, W., Chen, X., Lu, Y.: An enhanced CNN-enabled learning method for promoting ship detection in maritime surveillance system. Ocean Eng. 235, 109435 (2021)
25. Zhang, C., Li, H., Wang, X., et al.: Cross-scene crowd counting via deep-convolutional neural networks. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2015), Boston, MA, pp. 833–841 (2015)
26. Shen, Z., Xu, Y., Ni, B., et al.: Crowd counting via adversarial cross-scaleconsistency pursuit. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2018), Salt Lake City, UT, pp. 5245–5254 (2018)
27. Ryan, D., Denman, S., Fookes, C., et al.: Crowd counting using multiple local features. In: Proceedings of the 2009 Digital Image Computing: Techniques and Applications, Melbourne, Australia, pp. 81–88 (2009)
28. Mahdi, H., Gang, P., Min, Y.: Counting moving people in crowds using motion statistics of feature-points. Multimedia Tool. Appl. 72(1), 453–487 (2014)

17519667, 0, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/ipr2.12686 by National Institutes Of Health Malaysia, Wiley Online Library on [07/06/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

29. Arteta, C., Lempitsky, V., Noble, J.A., et al.: Interactive object counting. In: Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, pp. 504–518 (2014)
30. Ryan, D., Denman, S., Fookes, C., et al.: Scene invariant multi camera crowd counting. Pattern Recogn. Lett. 44(15), 98–112 (2014)
31. He, K., Gkioxari, G., Dollár, P., et al.: Mask R-CNN. In: IEEE International Conference on Computer Vision, Venice, Italy, pp. 2980–2988 (2017)
32. Redmon, J., Farhadi, A.: Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767 (2018)
33. Chen, X., et al.: Microsoft COCO captions: Data collection and evaluation server. arXiv:1504.00325 (2015). http://cocodataset.org
34. Everingham, M., Zisserman, A., Williams, C.K.I., Van Gool, L., Allan, M., Bishop, C.M., et al.: The 2005 Pascal visual object classes challenge, Proc. Mach. Learn. Challenges Workshop, pp. 117–176 (2005). http://host.robots.ox.ac.uk/pascal/VOC/
35. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A large-scale hierarchical image database, CVPR, Miami, 248–255 (2009). https://image-net.org/

How to cite this article: Liu, D., Wang, Z., Meng, X.: Fast intensive crowd counting model of Internet of Things based on multi-scale attention mechanism. IET Image Process. 1–10 (2022). https://doi.org/10.1049/ipr2.12686

17519667, 0, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/ipr2.12686 by National Institutes Of Health Malaysia, Wiley Online Library on [07/06/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License