J. Meas. Sci. Instrum., 2025, 16(4): 515-525. http://jmsi.publish.founderss.cn/
DOI: 10.62756/jmsi.1674-8042.2025050 jmsi@nuc.edu.cn

Dynamic regional population counting and localization method based on high resolution fusion

ZHANG Jiaojiao[1], CHEN Yong[1,2*]

1. School of Electronic and Information Engineering, Lanzhou Jiaotong University, Lanzhou 730070, China;
2. Gansu Provincial Engineering Research Center for Artificial Intelligence and Graphics & Image Processing, Lanzhou Jiaotong University, Lanzhou 730070, China

***Corresponding author:** CHEN Yong (chenyong@mail.lzjtu.cn)

Received: May 6, 2024 **Revised:** May 30, 2024 **Accepted:** July 6, 2024

Abstract: Aiming at the problem of inaccurate crowd counting and location in dense scenes, a dynamic region-sensing crowd counting and location method based on high-resolution fusion was proposed. Firstly, U-HRNet was used as the main backbone to extract high-resolution features of the population and enhance the ability of feature extraction with different resolutions. Then, the dynamic regional awareness attention module was designed to make full use of the global and local feature information, refine the differentiated learning of target feature and background feature, reduce the interference of background feature, and improve the positioning performance of the model. Finally, the predicted threshold map and confidence map were input into the binarization module to output the prediction and counting results of the crowd independent individual target. Experimental results showed that the proposed method achieved good performance of counting and positioning in different scenarios.

Key words: deep learning; dense crowd count; high-resolution fusion; dynamic regional awareness; crowd location

0 Introduction

Crowd counting refers to the analysis and monitoring of high-density crowds by estimating the number and density of crowds in different scenarios, which has important application value in public security, monitoring system, traffic control and other fields. In recent years, with the rapid development of deep learning[1] technology, the counting accuracy of crowd counting algorithm has been improved to some extent. However, in practical applications, due to the limitation of shooting angle, a large number of irrelevant information interference, and scale inequality factors, it is still a huge challenge to accurately count dense crowds[2].

With the development of deep learning, crowd counting is derived from the positioning task, which is to locate each head to a more accurate position, becoming a new computer vision task[3]. Crowd location is a difficult problem in crowd analysis. Different from traditional counting methods, crowd positioning is to provide specific spatial location information and get the number of crowds by judging whether there are individuals in the target area. On this basis, crowd location is a target recognition algorithm based on the instance level, which has important application value in abnormal event detection[4], behavior detection[5], security monitoring and so on.

Traditionally, crowd counting and location are considered two separate tasks. The most advanced crowd counting methods are based on regression networks, which focus on regression counting and ignore the location of object instances. These methods are designed to predict density maps, where counts are calculated by integrals on the map[6]. Li et al.[7] proposed the void convolution detection network (CSRNet), which used void convolution to obtain a larger receptive field. But it is not suitable for population counting with large scale variation in dense scenes. Zhang et al. [8] proposed a multi-column convolutional neural network (MCNN) method to capture the feature information of crowds at different scales. But its fixed three-column convolution cannot meet the scale variability of crowd images. Ning et al.[9] proposed a deformable convolution network (ADCNet) with attention mechanism, which improved the quality of crowd density map through attention mechanism. But this method is greatly affected by background interference in dense scenes. Gao et al.[10]

proposed a network framework for spatial-channel attention regression (SCANet). But there are problems such as loss of individual feature information.

Recently, crowd targeting methods have been developed for dense crowds that can locate crowds and obtain comparable counting results. The positioning process is to generate key points on each head. Song et al.[11] proposed a purely point-based framework (P2PNet) for crowd location and counting. But there were counting errors in crowded crowd scenarios with different densities. Liu et al.[12] proposed that by conducting dynamic queries, the sparse and dense areas of the population could be delineated, and this approach was utilized for the statistics and positioning of the population size. The quadtree principle is used to locate the crowd, but the background information is ignored in the feature extraction stage, resulting in inaccurate counting and location. Zand et al. [13]proposed that MPS network used three VGG16 networks to extract multi-scale features of crowd images, and used confidence density map to achieve crowd localization. However, it is easy to lose useful population feature information during feature fusion.

To sum up, irregular scale transformations and complex background disturbances will inevitably occur in crowded scenes, which will have a certain impact on crowd counting and location algorithms, resulting in large deviations in the final statistical results. Aiming at the problem of inaccurate crowd counting and location in dense scenes, we proposed a high-resolution fusion crowd counting and location method based on dynamic area perception.

1 Proposed method

1.1 Global network structure

Aiming at the problem of inaccurate crowd counting and location in dense scenes, a dynamic region-sensing crowd counting and location method based on high-resolution fusion was proposed. The proposed model consisted of three parts: high-resolution feature extraction network, dynamic region awareness attention, and binarization module.

Firstly, U-HRNet was used as a high-resolution feature extraction network to extract features with different resolutions and improve the feature semantic representation ability. Then, the dynamic regional awareness attention was designed to make full use of the global and local feature information to enhance the differential learning of target features and background features, reduce the interference of background features, and enhance the performance of model counting and positioning. Finally, a binary graph for predicting individual target location was generated by using the predicted threshold graph and confidence graph, and the prediction of individual target is output. Its structure is shown in Fig.1, where the stage represents sub-networks with different resolutions, and the subsequent stages gradually increase the sub-networks with higher scores to lower scores, and these multi-resolution sub-networks are connected in parallel. At the same time, repeated multi-scale fusion is carried out, and each high-to-low score expression shares information from other parallel high-to-low score expressions, resulting in richer resolution expression.

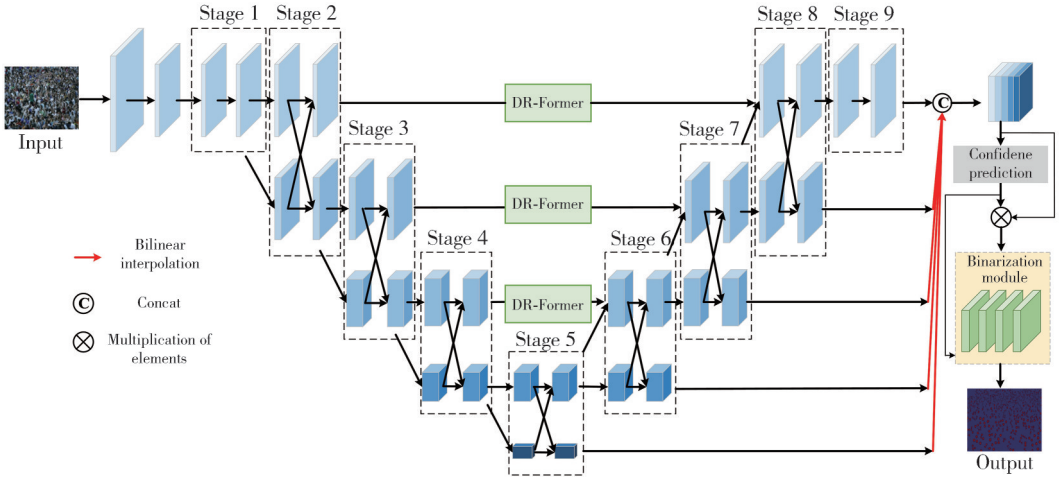

Fig. 1 Overall framework of network

1.2 High resolution feature extraction network

1.2.1 U-HRNet network

In order to realize the effective extraction of individual feature information from the input dense crowd images, the high-resolution U-HRNet network was used to extract the feature information when the crowd images were input into the model. Among them, HRNet[14] was first proposed for

human pose estimation, and then some researchers further proved that HRNet had good performance in many other tasks such as object detection and semantic segmentation[15]. It follows that HRNet is powerful not only in terms of high-level semantic representation, but also in terms of low-level spatial details. As shown in Fig.2, the 1/4 resolution is consistent from the beginning of the network to the end, and as the depth of the network increases, more low resolutions are added for semantic representation learning, thereby improving the high-resolution representation through multi-resolution fusion.

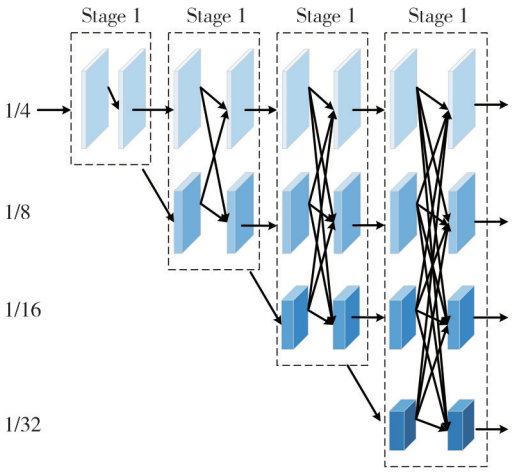

Fig. 2 HRNet network structure diagram

However, HRNet does not work well for some intensive forecasting tasks. For example, when predicting dense crowds, it is critical to introduce high-level global information to help pixels predict their semantic categories. It can be seen that HRNet has the disadvantage, namely the last feature of the 1/32 resolution branch with the strongest semantic representation is output directly and not fully utilized. The computation allocation between high and low resolution branches is not optimized, and more attention should be paid to the low resolution branches with strong semantic representation.

To address the above shortcomings, the method in this chapter adopts the high-resolution network U-HRNet. By reconstructing the macro-layout of U-Net, as shown in Fig. 3, the network uses convolution to completely replace the fully connected layer, so that the input size of the model is not limited.

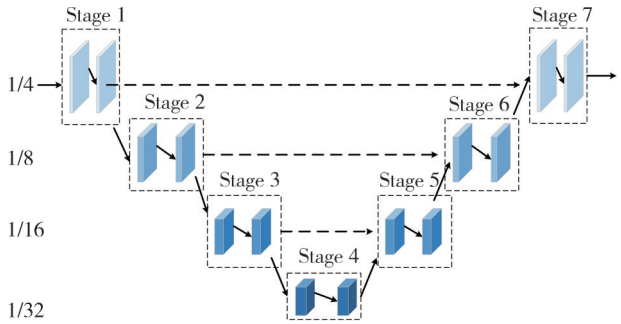

Fig. 3 U-Net network structure

To enhance the semantic representation effect of high-resolution output, after HRNet, the image will be input into a module whose resolution is reduced to 1/4 of the original, and this module will output feature maps with the same 1/4 resolution. Fig.4 shows the main structure of U-HRNet. This is consistent with the U-Net layout shown in Fig.3.

As shown in Fig. 4, its main body appears as a U-shaped network on a macro level, and consists of several HR modules on the network details. Each HR module consists of two branches.

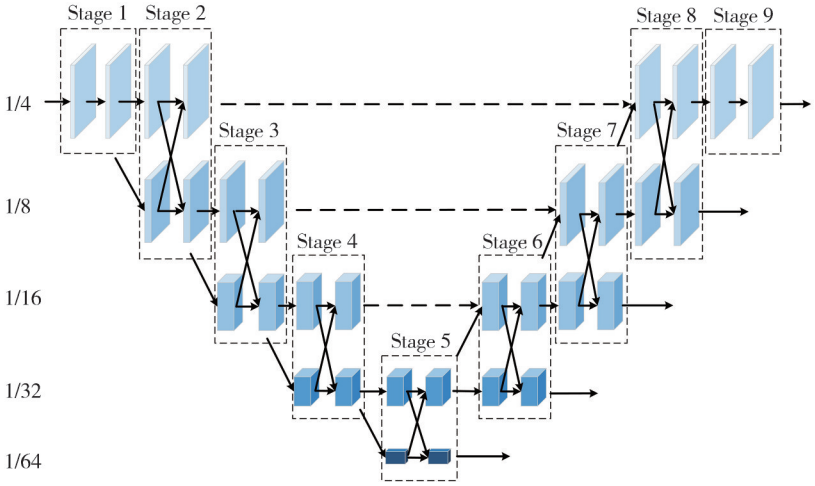

Fig. 4 U-HRNet network structure

First, the high-resolution branches of the last two stages of HRNet are removed, namely the 1/4 resolution branches of stage 3 and stage 4 and the 1/8 resolution branches of stage 4 in Fig.2, to release a large amount of computation.

Then, in order to improve the semantic representation of the high-resolution output, several stages are added after the lowest resolution phase. This stage gradually samples the feature map and merges it with the features of previous

stages. Features with the strongest semantic representation output from the lowest resolution stage can be merged earlier with lower-level high-resolution features, so that spatial details can be inferred more accurately through adequate extraction of the strongest representations. Finally, HR modules are rearranged in different stages. Adding modules in the low resolution phase and reducing them in the high resolution phase improves the semantic representation to a higher degree.

In addition, stages with 1/32 and 1/64 resolution branches are added to produce more feature information. Similar to U-Net, it connects stage 2 and stage 8, stage 3 and stage 7, stage 4 and stage 6, respectively. This allows the network to make full use of both high-level and low-level features, while propagating the gradient directly to the previous stage.

Three fusion modules are set before stage 8, stage 7, and stage 6 to fuse low-level features from the high-resolution branch output of stage 2, stage 3, and stage 4 with upsampled features from the high-resolution branch output of stage 7, stage 6, and stage 5, respectively. In order to enhance the connectivity of the network, the two input features in the channel dimension are first pooled with a kernel size of 2, and then fused effectively through Concat operation.

1.2.2 Dynamic region-aware attention

Because the crowd image scene is complex, the target size is different, and the crowd information and background information are interdependent, the target feature information is weak. In order to better highlight the feature information of individuals in the image and reduce the interference of noise and irrelevant background in the image, dynamic region-aware attention (DR-Former) is designed, as shown in Fig.5.

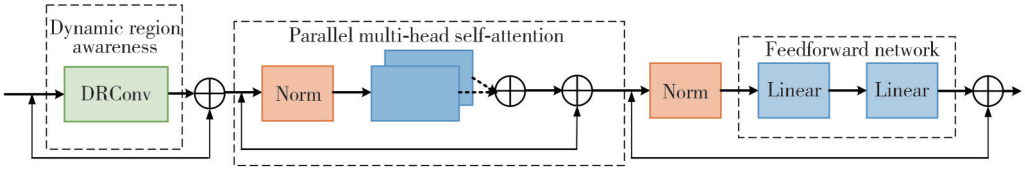

Fig. 5 Dynamic region-aware attention map

The module consists of three key modules: dynamic region awareness, parallel multi-head self-attention, and feedforward networking. Firstly, dynamic region-aware convolution is introduced to enhance local effective crowd location information, and then the global context information is extracted by the proposed parallel multi-head self-attention. Finally, a feedforward network with two linear layers is added for point-by-point enhancement of image individual features.

1.2.3 Dynamic region awareness

Dynamic region aware convolution (DRConv) transfers the channel filter to the spatial dimension with learnable guidance, effectively improving the feature extraction capability of the convolution. A learnable guide mask module is designed in this convolution, and then the spatial dimension of each mask is divided into several intervals by a region sharing model, and each interval shares a filter. By using many different filtering algorithms, the filtering algorithm can make better use of the characteristics of the crowd information itself.

Dynamic region-aware convolution uses standard convolution to generate orientation from an input. The spatial dimension is partitioned. In a pilot mask, pixels of the same color are attached to the same region. The filter generating module is used to generate the filter for convolution operation. The structure is shown in Fig.6.

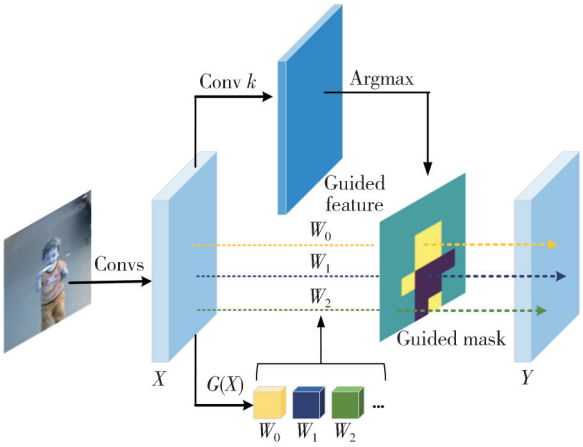

Fig. 6 Convolutional structure of dynamic region awareness

First, the input crowd feature map is globally pooled, and the length and width of the feature map are adjusted to 1. Then, the number of channels is reduced through the first fully connected layer, and the size of the feature map is adjusted to K through the second fully connected layer. Finally, the attention value corresponding to each convolution is obtained through the normalization operation $\pi_k(h)$.

$$0 \leqslant \pi_k(h) \leqslant 1, \sum_{k=1}^{K} \pi_k(h) = 1, \tag{1}$$

where K aggregated convolution and deviation values are used to manipulate the population feature map. With dynamic perceptual convolution, because different

convolution operations process smaller crowd image features, the computational amount does not increase much compared with ordinary convolution[16].

1.2.4 Parallel multi-head self-attention

Common CNN operation focuses on the local information of image features, which leads to the limitation of image global dependence. However, transformer network can effectively establish context relationships between features at different resolutions[17]. Traditional transformers stack multiple identical transformer "blocks" one after the other in order. However, the structural calculation of each block is complex and consists of many different layers that need to be combined in a specific way to achieve good performance. The method in this chapter optimizes traditional transformer blocks to reduce the depth of sequential subblocks in a parallel manner.

The original transformer structure is serial and computs MLP and attention subblocks in parallel to improve efficiency and minimize performance loss[18]. In the improved parallel block, both the MLP and attention subblocks take the same normalized input, combining the weights together by adding their outputs, combining individual jump connections. This parallelization can improve the counting efficiency of the model, and the training speed of the parallel block is increased by 15% compared with the standard "sequential" structure[19]. Its structure is shown in the Fig.7.

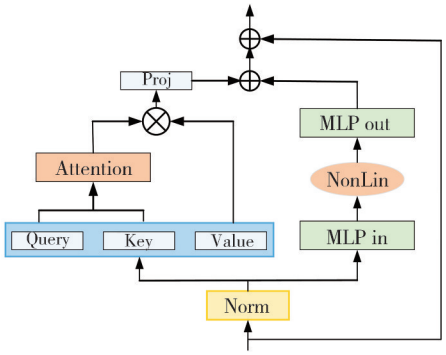

Fig. 7 Parallel multi-head self-attention structure

1.3 Binarization module

In order to obtain the threshold values of each region according to the image content, a binarization module is proposed, which is composed of a threshold encoder and a binarization layer. The former encodes the feature map of the image and generates a single value or map. The latter uses this value or map to binarize the confidence map and output the instance map. The feature graphs obtained by the dynamic region awareness convolution and the improved transformer are input into the binarization module and divided into binary graphs with learnable thresholds. Fig.8 shows the binary module structure.

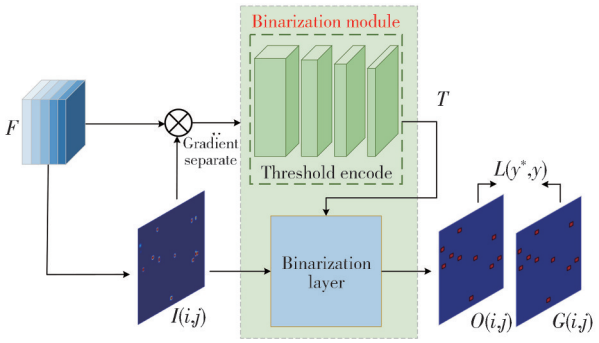

Fig. 8 Binary module structure diagram

In this module, the output threshold of confidence graph I is predicted according to feature graph F. With multiple convolutional layers as threshold encoders, feature F can map threshold encoders to threshold T by parameter θ.

$$T=\Phi(F;\theta). \tag{2}$$

The threshold encoder learns a specific value for the input crowd scenario. However, the profile image confidence distribution of different scales is very different. The main reason is that there are data biases in the data set, that is, a few small scale and large scale heads, resulting in a generally lower confidence than the medium-scale heads, and it is difficult for the network to learn some discriminant features between real samples and the background. Therefore, a pixel-level threshold structure is proposed, which consists of four convolution layers with PReLU activation function, two convolution nuclei with step size 1, and an average pooling layer. The specific structure is shown in the Fig.9.

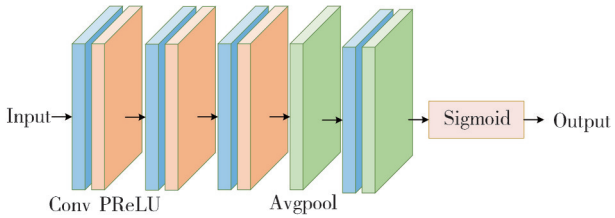

Fig. 9 Threshold encoder structure diagram

The configuration is as follows: Conv: 3×3, PReLU; Conv: 3×3, PReLU; Conv: 3×3, PReLU; Avgpool; Conv: 1×1, Avgpool; Sigmoid. In order to cover a large spatial acceptance field and save memory, the size of the input feature is adjusted to 1/8 of the original size. In addition, after the last two convolutional layers, average pooling of 15×15 is used with a step length of 1. In experiments, some very low (less than 0.1) or high (close to 1) values may be generated in the threshold plot, which makes the network prone to fluctuations. In addition, the high threshold results in many voids in the

head area. So, the Sigmoid function is compressed.

The goal of the confidence encoder is to have high confidence in the target region and low confidence in the background region. The threshold learner causes the target region to have a low threshold and the background region to have a high threshold. In this way, background noise can be filtered out as much as possible, and low confidence prospects such as small scale and large scale head can be retained for positioning. Since the two tasks are opposed, a gradient separation operation is added between the trunk and the threshold learner, as shown in Fig. 9. For threshold encoders, the derivative of T at each pixel is -1. This means that when it flows through the threshold T to the threshold encoder, the gradient will be reversed. Therefore, the parameters θ in the threshold learner can be optimized by

$$\theta = \theta + \beta \nabla_\theta L(y^*, y), \tag{3}$$

where β is the learning rate of the threshold encoder. In addition to optimizing threshold encoders, hard loss can be calculated by binary prediction and labeling and backpropagated from input I to a confidence mapped prediction network. Unlike threshold encoders, the gradient of a confidence predictor has the same symbol as the loss $L(y^*, y)$. Suppose the input I is output by a confidence predictor of parameter θ, and it is updated with the learning rate γ as

$$\theta = \theta - \gamma \nabla_\theta L(y^*, y). \tag{4}$$

After the threshold graph is obtained, it is sent to the binarization layer together with the confidence graph, and the binarization layer generates the segmentation graph. Finally, the frame and center of the independent instance are obtained by detecting the connected component. Its structure is shown in Fig.10.

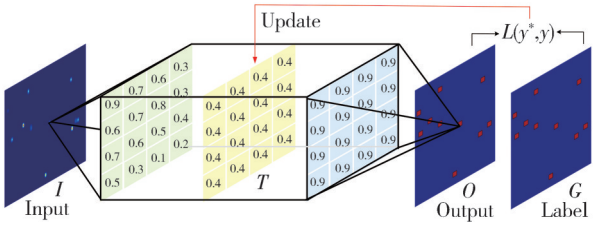

Fig. 10 Structure diagram of binary layer

The binarization layer has two inputs: the confidence graph I and the threshold T. For input I, the value range of each pixel is [0, 1]. The goal of the binarization layer is to learn a threshold T to segment the input image so that its output image O is as close as possible to the target image G. Given an input sequence and a target sequence, the optimal segmentation of all data is achieved by customizing thresholds for different images.

2 Results and discussion

2.1 Experimental data and evaluation index

Due to the large scale of the dataset and the varying sizes of the input population images, the original images were first scaled by a ratio of 0.8 to 1.2, then randomly cropped, and finally 512×1 024-sized images were obtained. The network was optimized using the Adam optimizer with an initial learning rate of 1×10^{-6}. This model was configured in the PyTorch deep learning framework, trained and tested on the NVIDIA GeForce RTX 2080 Ti GPU, and comparative algorithm experiments were conducted under the same environment.

How to evaluate the counting accuracy and generalization ability of the model is an important problem in the task of population counting research. Commonly used quantitative indicators are: mean absolute error (MAE) and mean square error (MSE). The MAE is the mean of the deviation between the predicted number of people and the actual number of people, and the mean square error is the mean of the square of the deviation between the predicted number and the true number of people. When evaluating the counting algorithm, the small MAE and MSE values often reflect the good performance of a counting model. MAE and MSE are calculated by

$$MAE = \frac{1}{N}\sum_{i=1}^{N}\left|G_i - G_i^{GT}\right|, \tag{5}$$

$$MSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(G_i - G_i^{GT}\right)^2}, \tag{6}$$

where N represents the number of test images in the test set; G_i and G_i^{GT} represent the total number of people predicted by the network and the total number of people in the real population density map, respectively.

2.2 Experimental results of ShanghaiTech dataset

Firstly, a performance comparison experiment was performed on ShanghaiTech dataset. The proposed method was compared with P2PNet[11], PET[12], MPS[13], IIM[20] and other methods for population counting. The mean square error and mean absolute error were used to compare the counting performance of the model. The smaller the error value, the higher the counting accuracy. The accuracy rate and recall rate were used to quantitatively evaluate and compare the positioning performance. The higher the value, the higher the positioning accuracy. The results are shown in

Tables 1 and 2.

Table 1 ShanghaiTech PartA dataset count evaluation index comparison

Method	MAE	MSE	Precision/%	Recall/%
PET	60.4	104.1	79.5	78.0
MPS	71.4	110.7	77.3	80.6
IIM	83.2	96.8	76.3	70.5
P2PNet	58.7	98.6	78.9	82.7
Ours	**55.6**	**95.4**	**80.7**	**81.4**

Table 2 ShanghaiTech PartB dataset count evaluation index comparison

Method	MAE	MSE	Precision/%	Recall/%
PET	10.8	19.4	78.9	83.6
MPS	9.6	15.0	87.8	83.3
IIM	15.6	20.4	89.8	78.6
P2PNet	10.6	17.0	86.2	90.7
Ours	**9.1**	**14.8**	**91.7**	**93.9**

It could be seen that the MAE and MSE values of the proposed algorithm are lower than those of the comparison algorithm. The smaller the evaluation indexed of MAE and MSE, the smaller the counting error. Although the size of the network model in this chapter is not the smallest, its counting performance is generally better than other algorithms and its counting accuracy is higher.

To enhance the visualization effect, Fig.11 presents a visualization comparison between the method proposed in this chapter and three comparative methods. Randomly selected from the ShanghaiTech PartA dataset are three groups of dense crowd images. In the field of deep learning, ground truth (GT) refers to accurate labels or data used for model training and evaluation. It serves as a reference standard for machine learning algorithms to measure models and judge their prediction accuracy. In deep learning tasks, models are typically trained to predict specific attributes or labels of input data. These labels are usually provided alongside datasets manually annotated by experts (with high accuracy), offering precise reference values that are defined as GT. GT plays three core roles in deep learning: training models, evaluating model performance, and comparing differences between different algorithms. It should be noted that GT represents the actual number of people, while Per denotes the number of people predicted by the model. The counting results of the proposed method are more accurate and closer to the true values.

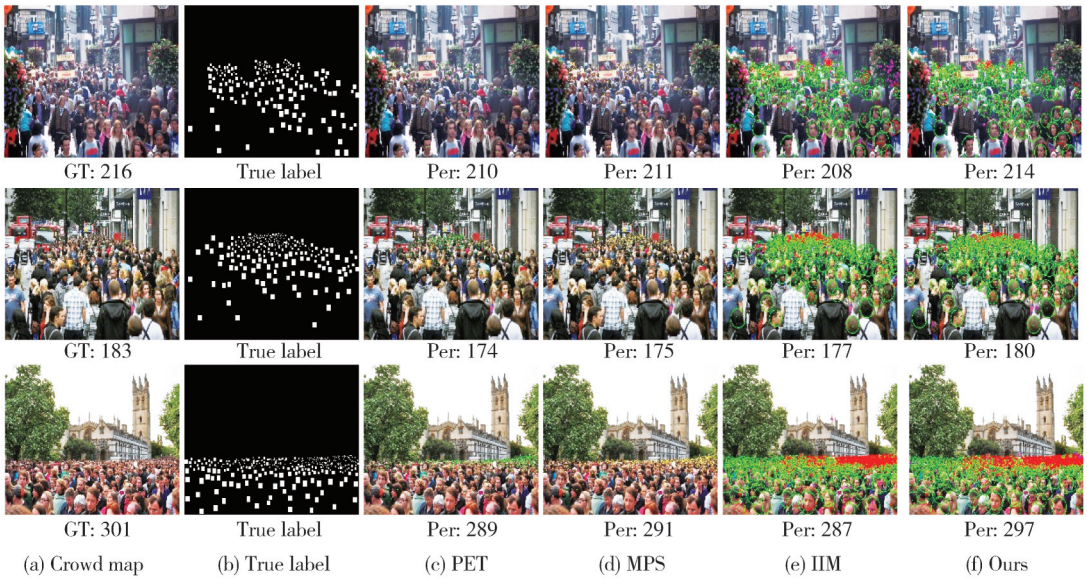

Fig. 11 Results of counting and positioning of ShanghaiTech PartA dataset

2.3 Experimental results of NUWP dataset

On the NWPU-crowd data set, in order to further compare and illustrate the effectiveness of the proposed method, the results are further compared with P2Pnet, PET, MPS, IIM and other methods, as shown in Table 3. On the NWPU-crowd data set, for crowd counting under different density distributions and severe background interference scenarios, compared with other comparison algorithms, the network proposed in this paper achieves better counting results.

Table 3 NWPU-Crowd dataset count evaluation index comparison

Method	MAE	MSE	Precision/%	Recall/%
PET	74.4	328.5	68.4	66.6
MPS	90.8	318.2	47.7	68.7
IIM	88.9	412.5	82.9	70.2
P2PNet	71.1	354.5	75.3	65.1
Ours	**70.9**	**299.1**	**85.4**	**76.8**

To visually evaluate performance, three groups of dense crowd images were randomly selected from this

dataset, with the visualization comparison between the proposed method and three comparative methods presented in Fig.12. It can be observed that the counting error of the algorithm in this chapter is smaller than that of the other three methods, demonstrating its superior counting performance for crowd scenes with varying density distributions and severe background interference. This result verifies the effectiveness of the improved resolution feature extraction module and dynamic region-aware attention module introduced in this study.

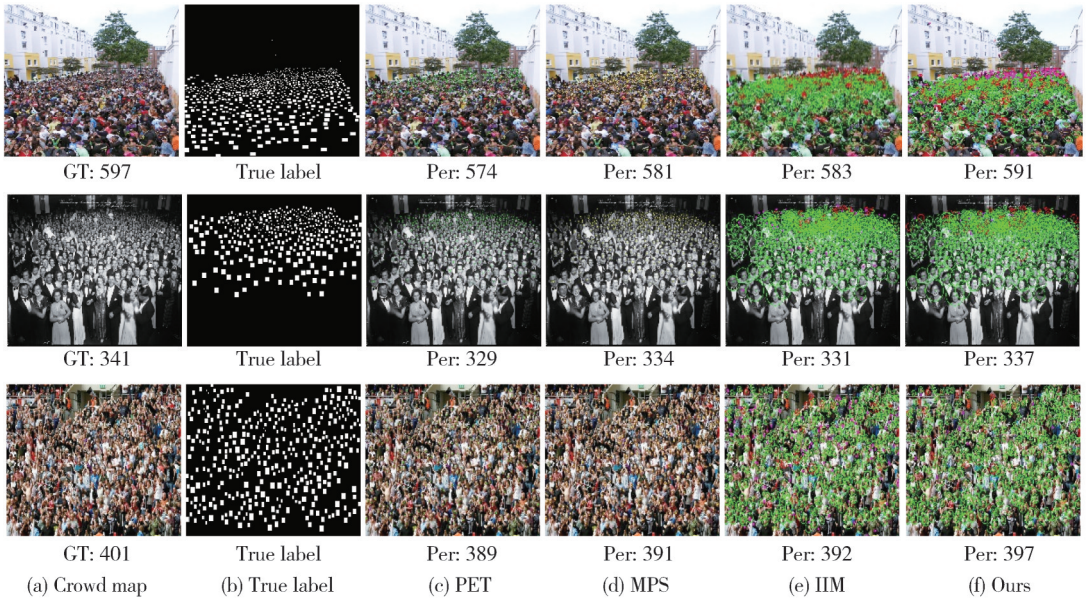

Fig. 12 Results of counting and positioning of NWPU-Crowd dataset

2.4 Experimental results of FDST dataset

Then the FDST dataset is compared by counting experiments. The method in this paper is compared with the counting of P2PNet, PET, MPS, and IIM methods, and the results are shown in Table 4. It can be seen that the method presented in this paper still has the smallest counting error compared with other methods for scenes with large crowd size differences, which further verifies its effectiveness.

Table 4 FDST dataset count evaluation index comparison

Method	MAE	MSE	Precision/%	Recall/%
PET	5.1	6.1	95.9	95.2
MPS	3.5	4.8	96.1	96.7
IIM	2.6	2.9	95.4	95.3
P2PNet	2.3	3.1	96.2	97.8
Ours	**2.1**	**2.2**	**97.6**	**98.5**

Three groups of images are randomly selected on the FDST data set as visualization experiment results, and the comparison is shown in Fig.13. It could be seen that there is certain background interference in the FDST data set, which is very challenging. From the comparison results, it could be found that the method in this paper still achieves better results.

To highlight the positioning effectiveness of the method proposed in this chapter, the positioning results of the last two columns in Fig. 13 have been locally magnified, as shown in Fig. 14. It can be observed that the IIM algorithm used for comparison neglects the detailed information of low-resolution features, thereby failing to effectively identify small-scale target information.

In order to verify the occlusion detection of the method in this paper, the vehicle occlusion detection in the experiment is locally enlarged, as shown in Fig. 15. Through local amplification, the target individual in the white car couldn't be located, while the right side of the method in this chapter could be effectively detected. the method in this chapter could effectively solve the background interference problem and improve the positioning accuracy.

The above experimental visualization results show true positive, false negative, and false positive. In order to distinguish these test results more clearly, they are marked with green circles, red circles, and magenta dotted, respectively. Fig.16 is a magnification of several examples of false positives in this method to facilitate observation of experimental results. Fig. 16(a) incorrectly identifies the human-shaped model sign, while Fig. 16(b) wrongly detects the pedestrian in the occlusion situation.

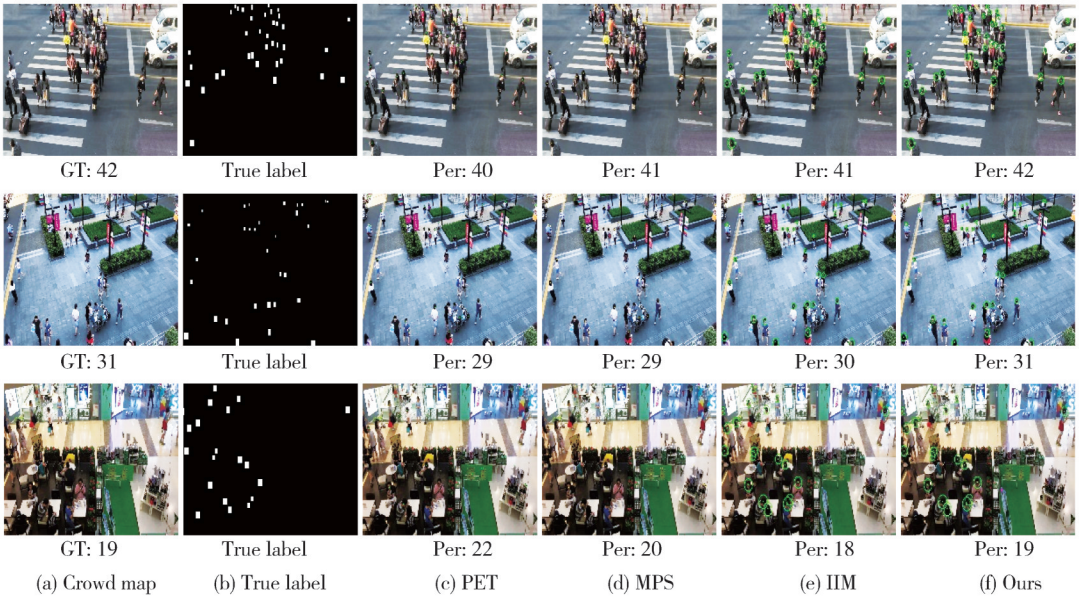

Fig. 13 Results of counting and positioning of FDST dataset

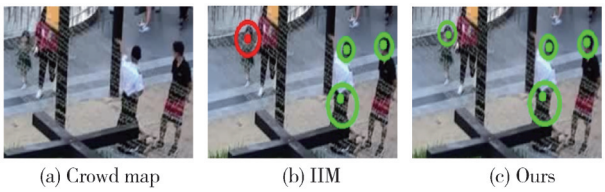

Fig. 14 Missing local magnification

Fig. 15 Occlusion effect enlarge image

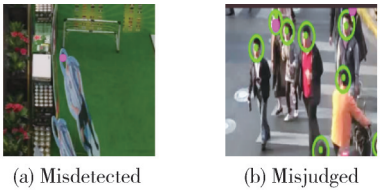

Fig. 16 False positive test results

2.5 Model complexity analysis

The time complexity and space complexity of the model were compared and analyzed. In deep learning neural network models, the computational complexity of the algorithm was generally measured by the number of operations (GFLOPs), and the spatial complexity of the algorithm was measured based on the required space size of the model. The proposed algorithm was compared with four classic deep learning methods for population counting and localization (including P2Pnet, PET, MPS, and IIM). The data regarding the time complexity and the size of the space model involved in the comparison results are presented in Table 5.

Table 5 Calculation volume and model size

Method	GFLOPs	Model size/MB
PET	**76.2**	**352**
MPS	129.4	628
IIM	77.6	572
P2Pnet	98.1	791
Ours	83.1	608

Due to the limited number of network layers, PET networks only use simple convolution operations for crowd feature extraction. Therefore, their computational complexity and model size are the smallest among the five methods. However, this method couldn't extract effective feature information and has poor counting and localization results. Compared to MPS and P2Pnet, the computational complexity of this method is significantly reduced, and the model size is also lower than the above two methods. Based on Tables 4 and 5, it can be seen that the algorithm proposed in this paper performs better in terms of overall indicators and algorithm performance.

2.6 Ablation experiment

Finally, in order to verify the effectiveness of the proposed network module, an ablation experiment was carried out. Based on the IIM method, the ablation experiments were divided into four groups for comparison: 1) IIM, 2) U-HRNet, 3) IIM+DR-former, 4) U-HRNet+DR-former. Performed on the NWPU-crowd dataset, the results are shown in the Table 6.

After using the U-HRNet high-resolution network in the IIM-based front-end network, the MAE and MSE values decrease. Because high-resolution fusion can ensure that the final prediction granularity is as close to

the pixel level as possible, and can obtain more accurate local population distinctions, such as clearer population edges. The extracted strong semantic information of the population ensures the accuracy of the overall prediction, especially in areas where it is difficult to distinguish or where the population size is large. High-resolution and advanced semantic representations are essential for dense population prediction.

Table 6 Results of NWPU-crowd ablation experiment

Method	MAE	MSE	Precision/%	Recall/%
IIM	52.74	85.06	75.2	73.2
U-HRNet	52.31	84.70	78.6	75.9
IIM+ DR-Former	51.96	82.14	79.4	79.5
U-HRNet+DR-Former	**50.45**	**79.36**	**81.9**	**83.7**

Generally, low-resolution feature maps can achieve stronger semantic representations, while high-resolution feature maps can better identify local features such as population edges, but have weaker semantic information. U-HRNet maintains the parallelism of low-resolution and high-resolution feature maps, adds more stages after the most semantic feature map, and repeatedly exchanges information between different resolutions. After adding the DR-former module, the recall rate and accuracy of the NWPU-Crowd dataset are improved, which proves that this module can effectively solve the problem of unbalanced population distribution scale. The addition of the dynamic region-aware attention module effectively reduces the error in population quantity statistics and reduces MAE and MSE to a certain extent. Through ablation experiments on public datasets, the effectiveness of each module is verified.

3 Conclusions

A dynamic region-aware crowd counting and positioning model based on high-resolution fusion was proposed, and a high-resolution feature extraction module was designed to enhance the crowd feature information. To fully utilize the context features and suppress irrelevant backgrounds, a dynamic region-aware attention mechanism was proposed to improve the accuracy of the crowd counting model. Finally, a binaryization module was used to output the prediction of individual crowd targets. Experimental results showed that, compared with the contrast algorithms, the proposed method had higher accuracy in crowd counting in different scenarios.

Acknowledgement

This work was supported by MOE (Ministry of Education in China) Project of Humanities and Social Sciences (No. 19YJC760012), Key Research and Development Project of Lanzhou Jiaotong University (No.ZDYF2304).

Declaration of conflicting interests

The authors have no conflict of interests related to this publication.

References

[1] WANG Y F, A K X, LIU K W, et al. Research on fault diagnosis method of switch machine based on deep learning. Journal of Test and Measurement Technology, 2023, 37(2): 106-111.

[2] ZENG X, WANG H K, GUO Q, et al. Correlation-attention guided regression network for efficient crowd counting. Journal of Visual Communication and Image Representation, 2024, 99: 104078.

[3] WANG Q, GAO J Y, LIN W, et al. NWPU-crowd: a large-scale benchmark for crowd counting and localization. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 43(6): 2141-2149.

[4] SAM D B, PERI S V, SUNDARAMAN M N, KAMATH A, et al. Locate, size, and count: accurately resolving people in dense crowds via detection. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 43: 2739-2751.

[5] ZHENG L X, ZHANG J, WANG X Y, et al. Multimodal-based abnormal behavior detection method in virtualization environment. Computers & Security, 2024, 143: 103908.

[6] WANG Y, HOU J H, CHAU L P. Object counting in video surveillance using multi-scale density map regression//ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing, May 12-17, 2019, Brighton, UK. New York: IEEE, 2019: 2422-2426.

[7] LI Y H, ZHANG X F, CHEN D M. CSRNet: dilated convolutional neural networks for understanding the highly congested scenes//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 18-23, 2018, Salt Lake City, UT, USA. New York: IEEE, 2018: 1091-1100.

[8] ZHANG Y Y, ZHOU D S, CHEN S Q, et al. Single-image crowd counting *via* multi-column convolutional neural network//2016 IEEE Conference on Computer Vision and Pattern Recognition, June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE, 2016: 589-597.

[9] LIU N, LONG Y C, ZOU C Q, et al. ADCrowdNet: an attention-injective deformable convolutional network for crowd understanding//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 15-20, 2019, Long Beach, CA, USA. New York: IEEE, 2020: 3220-3229.

[10] GAO J Y, WANG Q, YUAN Y. SCAR: Spatial-/channel-wise attention regression networks for crowd

counting. Neurocomputing, 2019, 363: 1-8.
[11] SONG Q Y, WANG C G, JIANG Z K, et al. Rethinking counting and localization in crowds: a purely point-based framework//2021 IEEE/CVF International Conference on Computer Vision, October 10-17, 2021, Montreal, QC, Canada. New York: IEEE, 2022: 3345-3354.
[12] LIU C X, LU H, CAO Z, et al. Point-query quadtree for crowd counting, localization, and more//2023 IEEE/CVF International Conference on Computer Vision, June 18-22, 2023, Vancouver, Canada. New York: IEEE, 2023: 1676-1685.
[13] ZAND M, DAMIRCHI H, FARLEY A, et al. Multiscale crowd counting and localization by multitask point supervision//ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing, May 23-27, 2022, Singapore. New York: IEEE, 2022: 1820-1824.
[14] SUN K, XIAO B, LIU D, et al. Deep high-resolution representation learning for human pose estimation// 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 15-20, 2019, Long Beach, CA, USA. New York: IEEE, 2019: 5686-5696.
[15] ZHENG Z Z, HU Y H, GUO T F, et al. AGHRNet: an attention ghost-HRNet for confirmation of catch-and-shake locations in jujube fruits vibration harvesting. Computers and Electronics in Agriculture, 2023, 210: 107921.
[16] CHEN J, WANG X J, GUO Z C, et al. Dynamic region-aware convolution//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 20-25, 2021, Nashville, TN, USA. New York: IEEE, 2021: 8060-8069.
[17] WANG S H, ZHENG D K, LI Y C. LiDAR-SLAM loop closure detection based on multi-scale point cloud feature transformer. Measurement Science and Technology, 2024, 35(3): 036305.
[18] BOBBY H, THOMAS H. Simplifying transformer blocks//International Conference on Learning Representation, May 7-11, 2024, Vienna, Austria. New York: IEEE, 2024:1-29.
[19] CHOWDHERY A, NARANG S, DEVLIN J, et al. PaLM: scaling language modeling with pathways. Journal of Machine Learning Research, 2023, 24(1): 11324-11436.
[20] GAO J Y, HAN T, WANG Q, et al. Learning independent instance maps for crowd localization. 2020: arXiv: 2012. 04164. https://arxiv. org/abs/2012. 04164

基于高分辨率融合的动态区域人群计数定位方法

张娇娇[1], 陈　永[1,2*]

1. 兰州交通大学 电子与信息工程学院, 甘肃 兰州 730070;
2. 甘肃省人工智能与图形图像处理工程研究中心, 甘肃 兰州 730070

摘　要: 针对密集场景下现有人群计数方法存在计数与定位不准确的问题, 提出了一种基于高分辨率融合的动态区域感知人群计数与定位方法。首先, 采用U-HRNet为主干网络, 提取人群高分辨率特征, 增强不同分辨率特征提取能力。然后, 设计动态区域感知注意力模块, 充分利用全局与局部特征信息, 细化目标特征和背景特征的差异化学习, 降低背景对人群计数的干扰, 提高模型定位性能。最后, 将预测的阈值图和置信图输入到二值化模块中, 输出人群计数结果。实验结果表明, 所提方法在不同场景下的计数与定位都取得了更好的表现。

关键词: 深度学习; 密集人群计数; 高分辨率融合; 动态区域感知; 人群定位

引用格式: ZHANG Jiaojiao, CHEN Yong. Dynamic regional population counting and localization method based on high resolution fusion. Journal of Measurement Science and Instrumentation, 2025, 16(4): 515-525. DOI: 10.62756/jmsi. 1674-8042.2025050