Hindawi
Journal of Advanced Transportation
Volume 2023, Article ID 6615879, 13 pages
https://doi.org/10.1155/2023/6615879

Research Article

Subway Platform Passenger Flow Counting Algorithm Based on Feature-Enhanced Pyramid and Mixed Attention

Jing Zuo, Guoyan Liu, and Zhao Yu

School of Automation and Electrical Engineering, Lanzhou Jiaotong University, Lanzhou 730070, China

Correspondence should be addressed to Jing Zuo; jdzuojing@163.com

Received 21 June 2023; Revised 16 October 2023; Accepted 30 October 2023; Published 7 November 2023

Academic Editor: Peter J. Jin

Accurate access to real-time passenger flows on subway platforms helps to refine management in the era of networked operations. The narrow subway platforms suffer from significant crowd scale discrepancies and complex backgrounds when counting passenger flow. In the proposed passenger flow counting algorithm, the feature-enhanced pyramid structure is used to retain the channel information of deep features and eliminate the aliasing effect caused by fusion to enhance the feature representation of the original image and effectively solve the scale problem. The mixed attention mechanism suppresses background interference by capturing the global context relationship and focusing on the target area. On the ShanghaiTech Part_A dataset, the mean absolute error (MAE) and mean square error (MSE) of the proposed algorithm are 2.3% and 1.4% higher than those of the comparison algorithm, respectively. The MAE and MSE on the self-built platform dataset reach 3.1 and 5.7, respectively. The experimental results show that the accuracy of the proposed algorithm is improved and can meet the counting requirements of the subway platform scene.

1. Introduction

Crowd counting aims to estimate the number and density distribution of people in images or videos and is used in fields such as crowd behavior analysis and public safety management. The surge of metro passenger flow on the metro has posed a huge challenge to the organization of traffic and safe operation, such as the difficulty of transportation organization during peak periods and the lack of operability in emergency management. Real-time access to station passenger flow through crowd counting algorithms can provide scientific data support for organizational management and safety alerts. For example, the departure interval can be optimized according to the passenger flow of the subway platform obtained in real time, and the turn-back station can be accurately obtained [1]. The distribution of passenger density on the platform is displayed in combination with the Passenger Information System (PIS) and the Public Address System (PA), so as to induce passenger travel behavior [2] and reduce operational pressure during peak hours. At the same time, it can also implement control strategies [3] such as closing stations and overtaking according to the platform passenger flow, so as to reduce the potential safety hazards caused by congestion.

Traditional crowd-counting algorithms fall into three categories, detection-based methods take the whole human body or body parts as the object of detection and calculate the number of people [4]; regression-based methods treat the crowd as a whole and complete the counting by establishing a mapping relationship between the extracted features and the number of people, such as ridge regression [5] and Bayesian regression [6]; and density estimation-based methods count by learning linear mapping [7] or nonlinear mapping [8] relationships between features and density maps. Traditional methods rely on manual feature extraction, which is less accurate and only applicable to sparse scenes. At present, convolutional neural networks are widely used in crowd counting due to their excellent feature extraction and learning capabilities. According to the structure of the neural network model, it is generally divided into two categories: single-branch structure and multibranch structure. The early crowd-counting algorithms are all single-branch structures. Wang et al. [9] applied CNN to crowd counting for the first time and the model uses the regression method to count. Due to the limitation of network width and depth, the counting accuracy in dense scenes needs to be improved and cannot meet the requirements of cross-scene counting. To solve the cross-scene problem, Zhang et al. [10] proposed the cross-scene counting model (Crowd CNN), and the algorithm fine-tunes the counting model according to the characteristics of the input scene so that it can accomplish cross-scene counting. The different distances between the crowd in the image and the camera lead to different crowd scales. To solve the multiscale problem, various multibranch networks have been proposed. The multicolumn convolutional neural network (MCNN) proposed by Zhang et al. [11] has three branches, which employ convolutional kernels of different sizes for feature extraction of targets at different scales to solve the scale problem. Sam et al. [12] proposed a multicolumn selection network (Switch-CNN), where the input images are first cut, and then parts of the images with different density levels are fed into the corresponding branches separately, and counting is done separately using different regression networks. The quality of the density map determines the counting accuracy. To obtain high-quality density maps, Sindagi and Patel [13] proposed the contextual pyramid model (CP-CNN), which applies the global and local contextual information extracted from different branches to density map generation. Although multibranch networks achieve better counting results, they are accompanied by the problems of large number of parameters, training difficulties, and model redundancy. To solve these problems, dilated convolutions [14], deformable convolutions [15], and generative adversarial networks [16] have been introduced in the field of crowd counting to reduce the complexity of the models and improve the counting accuracy. For passenger flow counting in the subway scene, Sheng et al. [17] proposed a counting method with the head and shoulder of passengers as the detection object. This method performs well when the passengers are sparse, but the counting accuracy decreases due to severe occlusion during peak hours. Zhang et al. [18] used a multiscale feature extraction module and transposed convolutional upsampling to enhance multiscale features but did not consider the effect of background interference on the counting task. Xiao et al. [19] conducted crowd counting in the target area of the subway based on the background difference method, but the background difference method is mostly aimed at moving objects and is not suitable for platform scenes where passengers are mostly stationary or moving slowly. Hu et al. [20] used a hybrid Gaussian background modeling method to compensate for the deficiencies in background differencing, but the regression-based approach makes the correlation between the features learned by the network and the number of people weak, and the accuracy needs to be improved. The double-region learning algorithm proposed by He et al. [21] divides the subway surveillance image into near region and far region and adopts different strategies for counting the two subregions to solve the impact of perspective distortion. However, the method can only divide the image into two fixed regions without considering the variability of the scene. The MPCNet proposed by Zhang et al. [22] uses multicolumn dilated convolution to aggregate multiscale context information in crowded scenes, but the multicolumn structure inference speed is slow and cannot meet the requirements of real-time detection. Tiny MetroNet proposed by Guo et al. [23] adopts a micro-passenger feature extraction network as the backbone network to achieve a balance between counting accuracy and detection speed. In the MDP algorithm proposed by Liu et al. [24], the MetroNext based on the multiscale convolutional attention module can quickly obtain the location information of the train and passengers, and the optical flow algorithm is used to predict the direction of passenger movement. The combination of the two completes the detection of passengers on and off the train. Yang et al. [25] introduced CBAM into YOLOv4 to solve the problem of inhomogeneous illumination in the station to improve the accuracy and robustness of the network. The MPDNet proposed by Yang et al. [26] uses the pyramid vision transformer to extract features and then uses an adaptive spatial feature fusion algorithm to compensate for the loss of spatial information in feature extraction, achieving higher accuracy while meeting real-time requirements.

Most of the current research is aimed at outdoor open scenes, which is quite different from the subway platform scene. The existing passenger flow counting algorithms in the platform scene still need to be improved. For the subway platform, the narrow and long platform leads to more obvious differences in passenger scales in different areas of the monitoring image, and there may be a problem of missing detection of small-scale heads away from the camera side. The variety of building facilities in the station leads to complex background and difficult crowd feature extraction. In addition, most of the existing public datasets are images of open scenes, and there is no public dataset suitable for subway platform scenes. Based on the above analysis, this paper first constructs a metro platform dataset by capturing images from Lanzhou metro platform surveillance video and then proposes a subway platform passenger flow counting algorithm based on feature-enhanced pyramid and mixed attention. The pyramid structure effectively fuses the semantic information and spatial information of deep and shallow features to solve the problem of different crowd scales. A mixed attention module is constructed to aggregate global context information, and the problem of complex background is solved by paying more attention to the target area.

2. Literature Review

The main difficulties of crowd counting in the platform scene are the large difference in head scale and the complex background of the platform. In this section, two types of networks related to the algorithm in this paper, i.e., multiscale feature fusion network and attention network, are reviewed.

2.1. Multiscale Feature Fusion Network. The different distances between the person and the camera in the image lead to the inconsistency of the head scale to be detected. The scale problem is one of the common problems in crowd counting, and multiscale feature fusion is an effective means to solve the scale problem. In the traditional method, the resolution of the input image is gradually reduced to construct the image pyramid in order to obtain the target of the corresponding scale in the image of each level. The effect of this method is significant, but the feature extraction of multiple inputs brings huge memory and time consumption. The feature pyramid [27] uses different layer feature maps as input and adds horizontal links and upsampling to fuse deep and shallow features, and the computational complexity of the model is reduced. The MARNet [28] proposed by Xie et al. improves the feature pyramid structure by introducing dilated convolutions with different dilation rates to enhance multiscale features to obtain richer context information. The STNet [29] proposed by Wang et al. uses a tree structure to hierarchically analyze the head scale, which enriches the scale level and solves the problem of large-scale changes in the head scale. SASNet [30] proposed by Song et al. can learn the correspondence between scale and feature level and obtain the final density map after weighting the confidence maps of different feature levels. The MZNet [31] proposed by Ma et al. enlarges or reduces the initial features to the corresponding level in each zooming path for aggregation and then propagates and utilizes multilevel context information in multiple zooming paths. MSIANet [32] proposed by Zhang et al. uses four branches of different receptive fields for feature extraction and then interacts the features of different branches to deal with continuous scale changes.

The above research studies use different methods to solve the scaling problem in image processing, which have achieved certain results but still have some problems, such as higher complexity of the model and feature loss. The feature-enhanced pyramid structure proposed in this paper uses a channel conversion module to highly preserve the channel features and a semantic consistency learning module to simplify the model while solving the aliasing effect.

2.2. Attention Network. The main idea of the attention mechanism is to allocate limited information processing resources to the parts of the input that are useful for task execution, and the widely used ones in crowd counting algorithms are channel attention, spatial attention, and pixel attention. The FANet [33] proposed by Niu et al. sets the weight of the background area to zero and weights the target area according to the area where the crowd is located and the density to exclude background interference. The MS-SPCANet [34] proposed by Wang et al. assigns different channel weights to different spatial positions of the channel feature map, in order to highlight useful information and suppress useless information to the greatest extent. MGANet [35] proposed by Li et al. uses spatial attention to focus on the human head region to solve the problem of foreground and background confusion and uses channel attention to enhance the dependence between features and improve semantic expression. In the coordinated attention module CA [36] proposed by Hou et al., the channel attention is decomposed into two one-dimensional feature coding processes, and the features are aggregated along two spatial directions. In this way, long-range dependencies can be captured in one spatial direction, while precise position information can be preserved in the other spatial direction. In CAFNet [37] proposed by Wang et al., pixel attention and channel attention are used to integrate low-level features into high-level features, and then density maps are generated by combining each layer of features that adaptively aggregate local context.

The existing research on attention mechanism is relatively rich, but there are still some limitations. Some studies only consider channel attention or spatial attention, which is not comprehensive enough, while the research considering both ignores the global relationship of feature maps. The mixed attention mechanism proposed in this paper uses the idea of nonlocal operation to obtain the long-distance dependence of spatial and channel feature maps to make full use of context information to obtain high-quality density maps.

3. Algorithm

The main difficulty in counting passenger flow in the subway platform scene comes from the high density of crowds during peak hours. The camera angle on the platform is low, and the head scale tends to increase from far to near and the scale difference is large, which needs to be taken into consideration in the algorithm design. In addition, since there are many escalators and other building facilities on the platform, the complex background brings difficulties to crowd feature extraction, and the interference brought by the complex background needs to be minimized when designing the algorithm.

Figure 1 shows the network framework of the algorithm in this paper, consisting of a VGG-16 network with the fully connected layer removed, a feature enhancement pyramid structure, and a mixed attention module. Taking the platform monitoring image as input, the first 13 layers of VGG-16 are used to extract the image features. The original features are sent into the feature-enhanced pyramid structure, and the problems of different crowd scales and small target missed detection are solved by aggregating features of different scales. Then, the fused features are sent to the mixed attention mechanism, which can effectively focus on the global information by capturing the long-distance dependencies of any two positions in the space or any two channels, which is helpful to solve the problem of background interference and occlusion. Finally, the attention feature map is upsampled to the size of the input image, and the predicted density map is obtained. After the integral sum, the number of passenger flows in the image can be obtained.

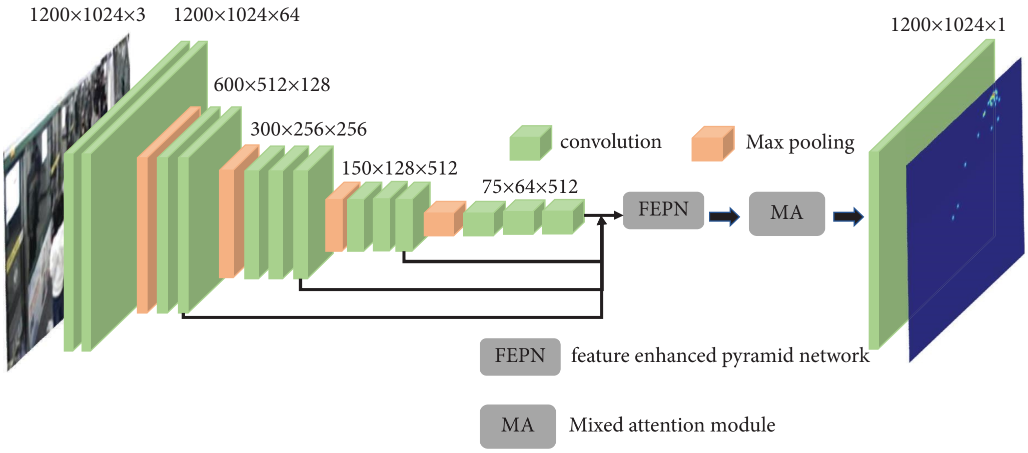


Figure 1: Subway platform passenger flow counting algorithm framework.

3.1. Feature-Enhanced Pyramid Structure. Targets of different scales in subway platform surveillance images will have a semantic generation gap after the same proportion of downsampling, which is manifested by the loss of small targets after multilayer convolution. The feature pyramid captures targets of different scales by fusing deep and shallow feature maps and solves the problem of missed detection of small targets. However, the traditional feature pyramid has the following disadvantages [38]. Firstly, the lateral link uses 1 × 1 convolution to reduce the number of channels of deep features so that the deep and shallow features can be fused, but this operation causes a large loss of channel information of deep features. Secondly, 3 × 3 convolution is used to eliminate the aliasing effect after feature fusion, which introduces redundant calculation. Therefore, this paper proposes an improved feature-enhanced pyramid structure, using a channel conversion module (CCM) and externally introduced semantic consistency learning module (SCLM) to solve the above two problems. The specific feature-enhanced pyramid structure is shown in Figure 2.

The backbone network extracts features from the bottom up and takes the feature map after the four-layer convolution of Conv2_2, Conv3_3, Conv4_3, and Conv5_3 as input, recorded as C2-C5. The input feature map is sent to the channel conversion module to convert the reduced channel information into pixel information, that is, the channel information is retained by expanding the width and height of the feature map. As shown in Figure 3, first the channel conversion operation can reshape the low-resolution feature map $H \times W \times \alpha^2 C$ into the high-resolution feature map $\alpha H \times \alpha W \times C$ by upsampling. Since the backbone network uses 2 times downsampling, α is taken as 2 in the algorithm for the subsequent fusion of adjacent feature maps. At this time, the width and height of the feature map increase by 2 times, and the number of channels decreases to 1/4. Because the number of channels in each layer needs to be consistent with the feature map C2, 1 × 1 convolution is used to enrich the channel information. Finally, 3 × 3 convolution is used to downsample the feature map to the original size, which can aggregate the original channel information at the pixel level. The deep feature map after CCM processing retains rich channel information for subsequent fusion stages.

Due to the inconsistent distribution of features, the direct fusion of deep feature maps with shallow feature maps after sampling will lead to aliasing effects, and the continuity of features cannot be guaranteed. Therefore, before the fusion after CCM and upsampling operation, the semantic consistency learning module is used to standardize the distribution of features. As shown in Figure 2, the SCLM module consists of a 3 × 3 convolution and two 1 × 1 convolutions, and then the consistency features are output through the activation layer. The channel information of the original feature map after CCM and SCLM is preserved and the aliasing effect brought by the fusion process is eliminated, and thus the features are enhanced. The fused feature maps P3-P5 are upsampled to the size of P2 and then spliced in the channel dimension to obtain the feature map F, which preserves more feature information.

3.2. Mixed Attention Mechanism. In the convolution process, the receptive field is limited to a certain range leading to differences in the feature representation between pixels of the same category [39], which then leads to a decrease in counting accuracy. The idea of the nonlocal operation [40] is that when calculating the weight of a certain position, all other positions need to be weighted so that the global contextual information can be fully utilized. Inspired by this, a mixed attention module is built to solve the problem of complex background of station monitoring images from two dimensions. The spatial attention mechanism can capture global dependencies and suppress background interference by focusing on target regions with high similarity. The channel attention weights each channel to highlight the channels useful for the counting task and suppress the useless channels.

Figure 4 shows the specific structure of the mixed attention mechanism, with the left-hand branch being the spatial attention mechanism and the right-hand branch being the channel attention mechanism. The idea of the spatial and channel attention mechanisms is similar, except

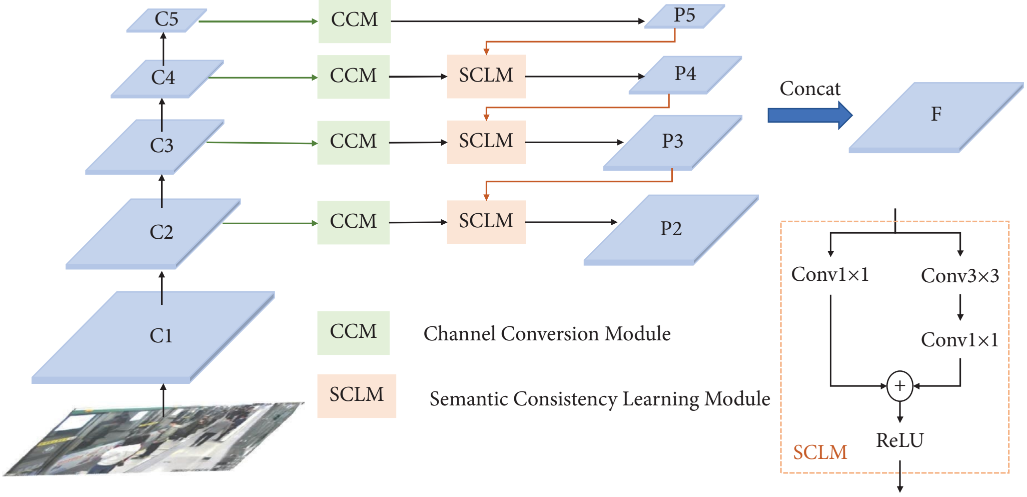


Figure 2: Feature-enhanced pyramid network diagram.

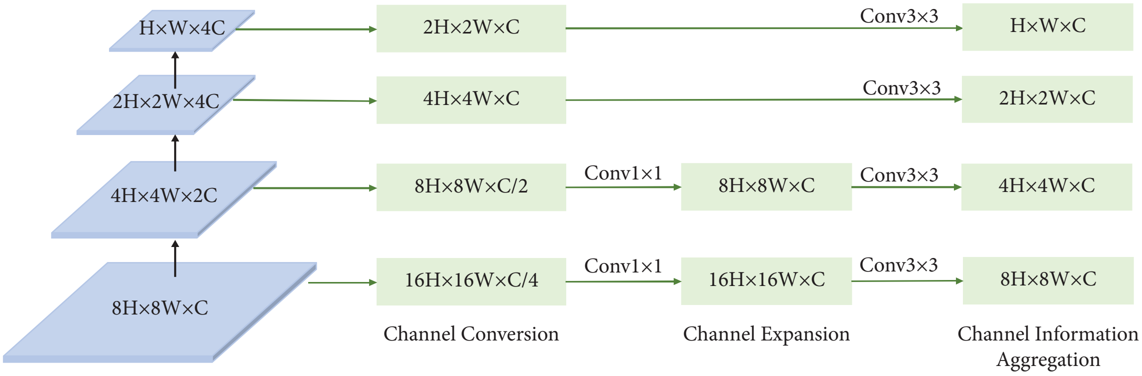


Figure 3: Structure diagram of channel conversion module.

that the spatial attention mechanism performs a 1×1 convolution operation to reduce the dimensionality before reshaping and transposing the feature map. The input feature map of the mixed attention mechanism is $F\in R^{C\times H\times W}$, where C, H, and W represent the channel, height, and width, respectively. After convolution, reshaping, and transposition, the feature maps $\{s_1, c_2\}\in R^{HW\times C}$ and $\{s_2, c_1\}\in R^{C\times HW}$ are obtained; then the matrix multiplication operation is performed and normalized by Softmax to obtain the spatial and channel attention maps s and c. The formulas are

$$s^{ij}=\frac{\exp\left(s_1^i\cdot s_2^j\right)}{\sum_{j=1}^{HW}\exp\left(s_1^i\cdot s_2^j\right)},$$
$$c^{ij}=\frac{\exp\left(c_1^i\cdot c_2^j\right)}{\sum_{j=1}^{C}\exp\left(c_1^i\cdot c_2^j\right)},\quad (1)$$

where s^{ij} represents the spatial weight of the i-th spatial position weighted by all positions j, c^{ij} represents the channel weight of the i-th channel weighted by all channels j, s_1^i and s_2^j represent the i-th and j-th positions of spatial feature maps s_1 and s_2, respectively, and c_1^i and c_2^j represent the i-th and j-th channels of channel feature maps c_1 and c_2, respectively. The output of the spatial and channel attention module is represented as

$$F_S=\lambda_1\sum_{j=1}^{HW}\left(s^{ij}\cdot s_3^j\right)+F^i,$$
$$F_C=\lambda_2\sum_{j=1}^{C}\left(c^{ij}\cdot c_3^j\right)+F^i,\quad (2)$$

where F_S and F_C denote the spatial and channel attention feature maps, respectively. s_3^j and c_3^j denote the j-th position or channel of the spatial feature map s_3 and channel feature map c_3, respectively, and matrix multiplication is used to reshape the feature maps into $R^{C\times H\times W}$. The coefficients λ_1 and λ_2 are learnable parameters that are initially set to zero and are adaptively assigned weights to local features through network training. F^i is the i-th position or channel of the input feature map. F_S and F_C are fused to obtain a mixed attentional feature map F_a with the same dimensions as F.

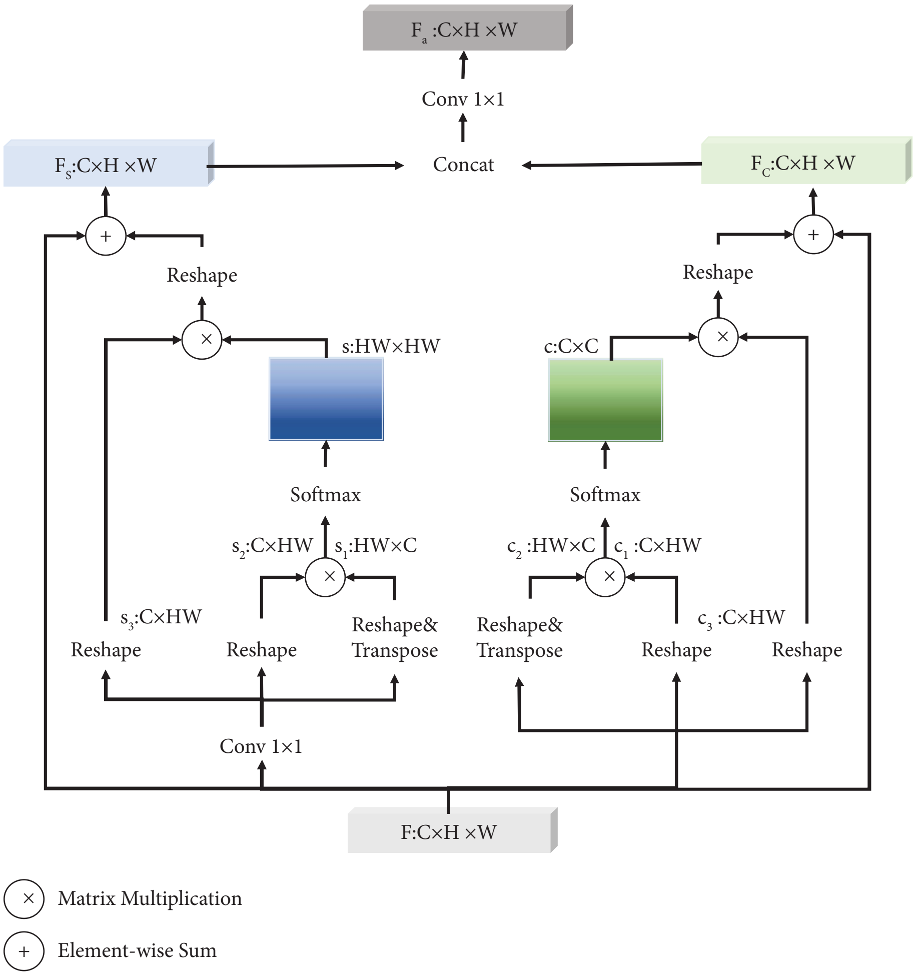


Figure 4: Structure diagram of the mixed attention mechanism.

3.3. Loss Function. The loss function is made up of two parts. The Euclidean distance loss function L_E is the pixel-level difference between the predicted density map and the true density map. The formula is

$$L_E = \frac{1}{N}\sum_{i=0}^{N}\left\|F(X_i;\theta) - F_i^{GT}\right\|_2^2, \tag{3}$$

where N is the number of images, X_i is the i-th input image, and θ is the learnable network parameter. $F(X_i;\theta)$ and F_i^{GT} are the predicted and true density maps for the i-th image. The Euclidean distance loss function is based on the premise that pixels are independent of each other, ignoring the correlation between them. Averaging all pixels without attention to structured information leads to blurred density maps and unclear details. To compensate for the shortcomings of the Euclidean distance loss function, the model introduces a structural similarity loss function L_S, which uses three local statistics of mean, variance, and covariance to calculate the similarity between the predicted density map and the true density map. The formula is

$$L_S = 1 - \frac{1}{M}\sum_{p\in P}\mathrm{SSIM}(p), \tag{4}$$

where M is the number of pixels in the density map and P is the image block corresponding to the same pixels p in the predicted and true density maps. SSIM is the structural similarity index and is calculated as

$$\mathrm{SSIM} = \frac{\left(2\mu_F\mu_{F^{GT}} + C_1\right)\left(2\sigma_{FF^{GT}} + C_2\right)}{\left(\mu_F^2 + \mu_{F^{GT}}^2 + C_1\right)\left(\sigma_F^2 + \sigma_{F^{GT}}^2 + C_2\right)}, \tag{5}$$

where μ_F, $\mu_{F^{GT}}$, σ_F^2, and $\sigma_{F^{GT}}^2$ denote the mean and variance of the predicted and true density maps, respectively, and $\sigma_{FF^{GT}}$ denotes the covariance between the predicted and true density maps. C_1 and C_2 are small constants set to prevent zeros in the denominator. SSIM $\in [-1, 1]$ and the image similarity is proportional to the value of SSIM.

The final loss function is obtained by weighting L_E and L_S:

$$L = L_E + \alpha L_S, \tag{6}$$

where α is the weighting coefficient used to balance pixel-level loss with structural loss and α is set to 0.001 through experiments.

4. Experimental Results and Analysis

The experiment was divided into two stages and the first was the training stage. Taking the training set images as input, the predicted value obtained by forward propagation was compared with the true value to obtain the loss value, and the parameters were updated in the process of backward propagation to make the loss value smaller and smaller until it reached the ideal value, completing the network training. The test set images were then fed into the trained network to obtain the predicted values, where the accuracy and robustness of the network were evaluated by the MAE and MSE.

4.1. Environment and Parameter Settings. All comparison experiments in this paper were completed on the Windows 11 system equipped with an NVIDIA GeForce RTX 3050 graphics card. The environment configuration was CUDA 11.6 + Anaconda 4.13 + Python 3.7 + PyTorch 1.10. The Gaussian distribution was used to initialize the convolutional layer parameters randomly, and the Adam algorithm was used to optimize the parameters. To balance the training speed and the loss, the initial learning rate was set to 1×10^{-5} and the learning rate decay parameter was set to 0.995. The training batch size was set to 16 and the number of iterations was set to 200. To better compare the performance of the algorithms, the experimental parameters of all the compared methods were set in the same way as the methods in this paper.

4.2. Evaluation Indicators. In this paper, mean absolute error (MAE) and mean square error (MSE) are used to evaluate the performance of the algorithm. MAE represents the error between the predicted and true values, reflecting accuracy, while MSE represents the degree of difference between the predicted and true values, reflecting robustness.

$$\begin{aligned} \mathrm{MAE} &= \frac{1}{N}\sum_{i}^{N}\left|C_i^P - C_i^{GT}\right|, \\ \mathrm{MSE} &= \sqrt{\frac{1}{N}\sum_{i}^{N}\left(C_i^P - C_i^{GT}\right)^2}, \end{aligned} \tag{7}$$

where N is the number of images and C_i^P and C_i^{GT} are the predicted and true number of people for the i-th image, respectively.

4.3. Dataset Description. To verify the performance of the proposed algorithm, experiments were conducted on ShanghaiTech and UCF_CC_50 public datasets and self-built station dataset, respectively.

The ShanghaiTech dataset contains 1198 images, with a total of 330,165 individuals tagged. The dataset is divided into two parts. The images in Part_A are randomly obtained from the Internet while the images in Part_B are obtained from street surveillance in Shanghai. Part_A is characterized by a high density of crowds and variable scenes, while Part_B is characterized by a low density of crowds but suffers from the problem of large differences in crowd scales. This dataset is a challenging dataset across different scene types and densities.

The UCF_CC_50 dataset images cover a wide range of scenes such as marathons, stadiums, and concerts. The average number of people in the images is as high as 1280, while the number of people in the single image ranges from 94 to 5453, with a large gap in density levels between images, making the dataset challenging. The disadvantage of this dataset is the insufficient number of images, only 50, and thus a five-fold cross-validation method was used to conduct experiments in this paper. The 50 images were randomly and equally divided into five, one of which was used in turn as the test set and the other four were combined as the training set, and the results of the five experiments were averaged as the final result.

For deep learning crowd counting, the quality of the dataset will to a certain extent affect the counting effectiveness of the model. The existing public datasets are mostly images of open scenes, while the long and narrow subway platforms and numerous construction facilities pose the problem of cluttered backgrounds. Due to the height limitation of the platform, the height of its surveillance cameras also differs from the public dataset. In order to better evaluate the performance of the model in this paper, platform images were collected from the Lanzhou Metro to build the dataset. Five stations in Lanzhou Metro Line 1 with high passenger flow, including Xizhanshizi, Xiguan, Dongfanghong Square, Wulipu, and Lanzhou University, were selected to capture images from the surveillance video at one end of the platform waiting area during the morning peak (e.g., 7:00–9:00), evening peak (e.g., 17:30–20:00), and flat peak periods (e.g., 10:00–16:00) of weekdays and weekends. The dataset is labelled with a total of 2000 images, of which 1500 are used as the training set and 500 as the test set. The size of each image is 1200 × 1024.

Typical images for each dataset are shown in Figure 5.

4.4. Experimental Result Analysis. Table 1 shows the experimental results of the proposed algorithm and five other classical or advanced comparison algorithms on the ShanghaiTech dataset. The comparison between the experimental results of the two-part datasets shows that the counting results of sparse scenes are better than those of dense scenes, indicating that dense scenes are still the key direction for future research on crowd counting. The proposed algorithm achieves the best results on this dataset compared to the comparison algorithm. Compared with the best MIA [43] model, the MAE and MSE of Part_A improved by 2.3% and 1.4%, respectively. The MAE and MSE of Part_B improved by 0.9% and 1.6%, respectively, indicating the effectiveness of the feature-enhanced pyramid structure and the mixed attention mechanism, which can perform the counting task well in the case of higher crowd density and different scales.

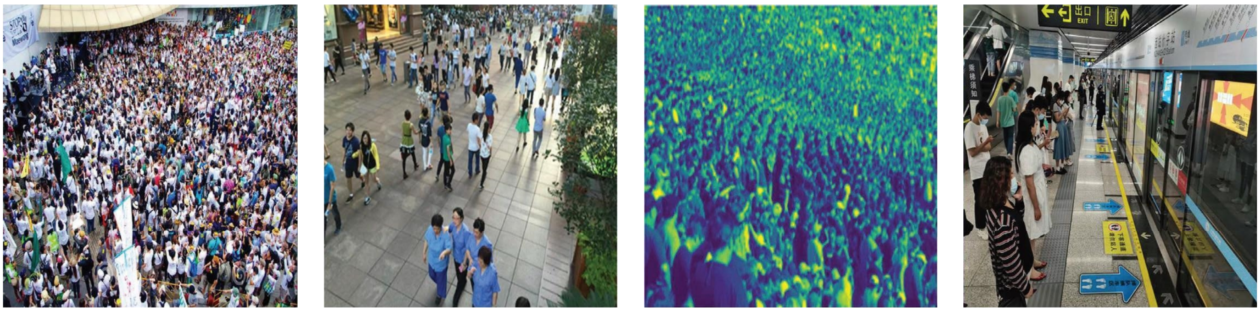

(a) (b) (c) (d)

Figure 5: Typical images for each dataset. (a) ShanghaiTech Part_A. (b) ShanghaiTech Part_B. (c) UCF_CC_50. (d) Self-built station dataset.

Table 1: Experimental results of the ShanghaiTech dataset.

Algorithm	Part_A		Part_B	
	MAE	MSE	MAE	MSE
MCNN [11]	110.2	173.2	26.4	41.3
CSRNet [14]	68.2	115.0	10.6	16.0
CAN [41]	62.3	100.0	7.9	12.9
MSEN [42]	63.5	106.2	8.2	12.3
MIA [43]	59.4	96.2	7.7	11.9
Proposed algorithm	57.1	94.8	6.8	10.3

Table 2: Experimental results of UCF_CC_50 dataset.

Algorithm	MAE	MSE	Params (MB)	Inference time (s)
MCNN [11]	377.6	509.1	23.62	1.28
CSRNet [14]	266.1	397.5	21.95	0.92
CAN [41]	212.2	243.7	24.94	1.51
MSEN [42]	226.7	310.6	23.81	1.33
MIA [43]	224.9	318.4	22.77	1.07
Proposed algorithm	213.1	254.7	20.48	0.84

Table 3: Experimental results of the self-built station dataset.

Algorithm	MAE	MSE
MCNN [11]	7.6	12.1
CSRNet [14]	4.6	9.2
CAN [41]	4.9	9.3
MSEN [42]	6.2	10.5
MIA [43]	5.3	9.8
Proposed algorithm	3.1	5.7

Table 2 shows the experimental results on the UCF_CC_50 dataset. It can be seen that only the context-aware model (CAN) [41] is superior to the proposed algorithm in the comparison algorithm, and the accuracy and robustness of other algorithms are lower than the proposed algorithm. The CAN network, which uses spatial pyramid pooling to compute scale-aware features, is a multicolumn network that adaptively encodes contextual information. The multiscale enhanced network (MSEN) [42] and the multivariate information aggregation (MIA) [43], which also employed multicolumn structures, have also achieved good results, indicating that the multicolumn structured model works better on this dataset. The algorithm in this paper is a single-column structure, which has less parameters and simpler calculation while achieving competitive results, and can also meet the counting requirements of various dense scenes. The last two columns are the number of parameters and the inference time of each algorithm; the model in this paper is a single-column structure; therefore, the number of parameters is less and the inference time is shorter.

The experimental results of the self-built platform dataset are shown in Table 3. The algorithm in this paper has achieved the best results because the algorithm has been improved on the traditional pyramid. The application of CCM and SCLM makes the channel information of the original feature map retained and eliminates the aliasing effect caused by the fusion process, enhances the feature representation, and helps to solve the scale problem. In addition, the mixed attention mechanism in the algorithm utilizes the idea of nonlocal image processing. By focusing on the relationship between local features, the global context information is fully aggregated to generate a high-quality prediction density map.

To further verify the effectiveness of the algorithm in this paper, the platform of Xizhanshizi Station of Lanzhou Metro Line 1 on April 26, 2023 (Wednesday), was selected, and the passenger flow on the platform was counted every 10 minutes during the period of 6:30–9:00, and a total of 16 groups of predicted passenger flow and the real passenger flow on the platform and the relative error were obtained, as shown in Figure 6. It can be seen from the figure that the number of passengers on the platform increases gradually with time, and the number of passengers on the platform increases significantly after 7:30 and remains at a high level, which is consistent with the trend of passenger flow in the morning peak of weekdays. The relative errors of the 16 groups of data are all within 4.5%, and the average absolute percentage error is 2.71%, which proves the effectiveness and accuracy of the passenger counting algorithm in this paper.

Figure 7 shows partial density maps obtained from the proposed model on different datasets, with every two rows of experimental result maps coming from the same dataset, arranged in the order of the ShanghaiTech, UCF_CC_50, and self-built station datasets. The experimental results on the first four rows of the public datasets show that the counting error is greater for dense scenes than that for

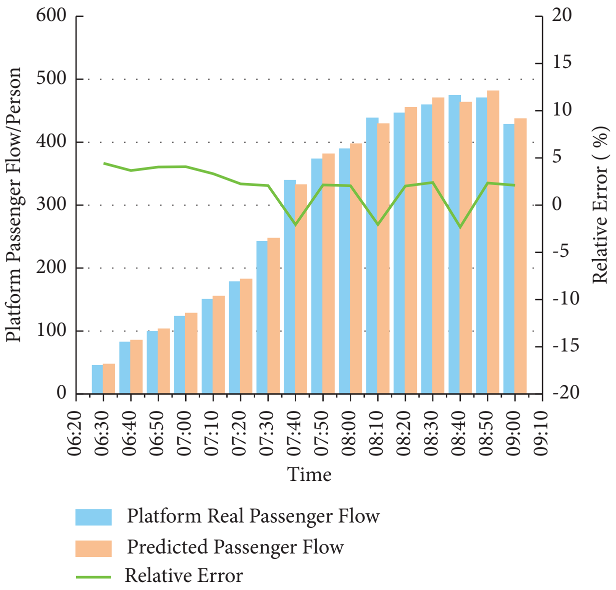


Figure 6: Platform real passenger flow and predicted passenger flow.

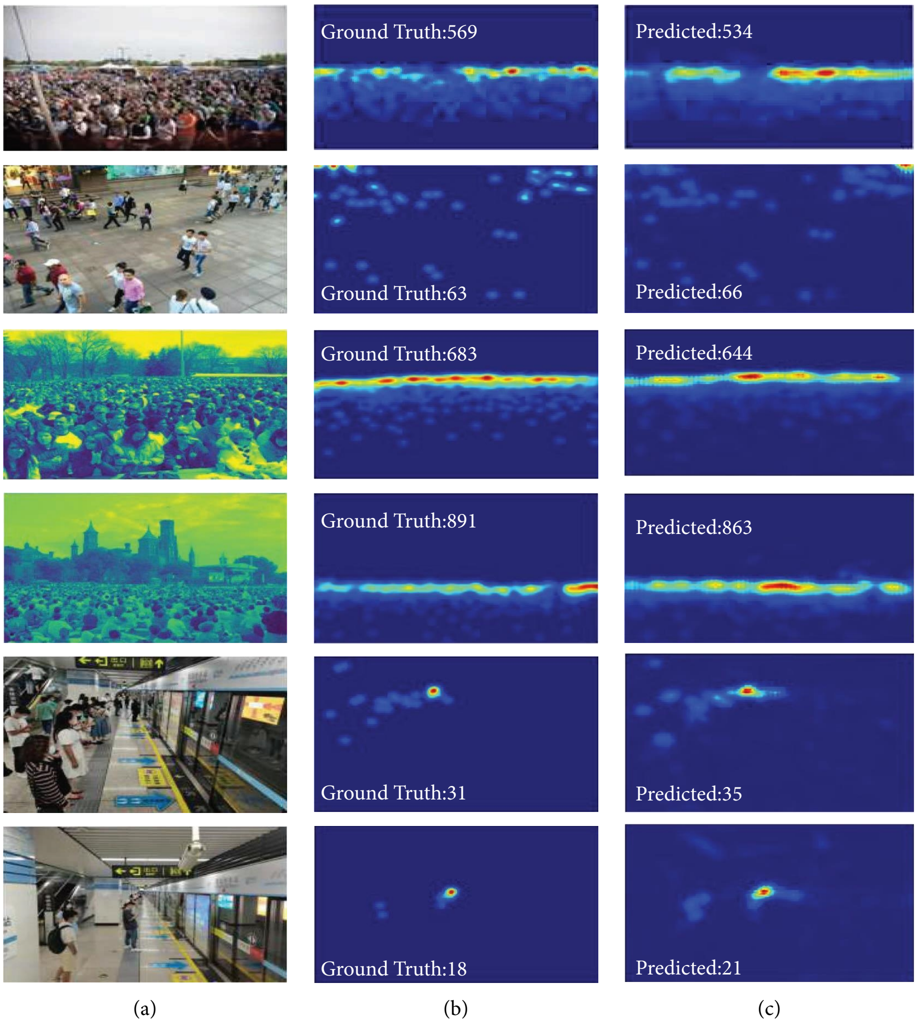
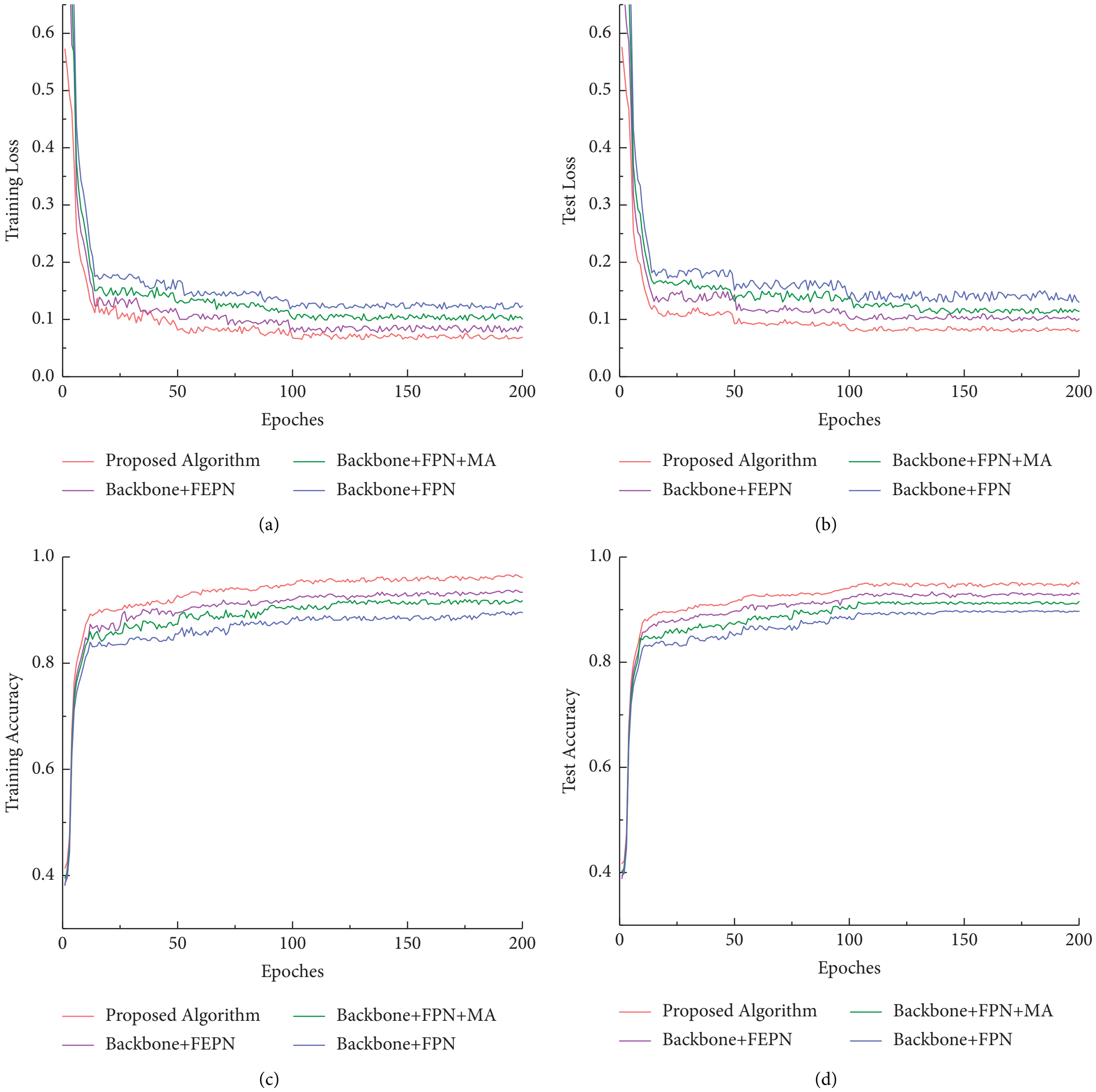


Figure 7: Partial density plot of each dataset. (a) Original images. (b) True density maps. (c) Predictive density maps.

Table 4: Ablation experimental results.

Algorithm	MAE	MSE	Params (MB)
Backbone + FPN	71.5	110.8	19.47
Backbone + FPN + MA	65.7	107.6	23.85
Backbone + FEPN	63.3	103.2	20.27
Backbone + FEPN + MA	57.1	94.8	24.65



(a)

(b)

(c)

(d)

Figure 8: Loss and accuracy convergence curve. (a) Training loss convergence curve. (b) Test loss convergence curve. (c) Training accuracy convergence curve. (d) Test accuracy convergence curve.

sparser scenes, but in general, the enhanced feature fusion and the suppression of background interference by the attention mechanism allow the algorithm to achieve good counting results The predicted values in the last two rows of the experiment are greater than the true values, and observation of the density distribution shows that it is the reflection of passengers by the platform screen doors that causes the repeat counts to bring about the slightly larger predicted values. The experimental results show that the model performs well on both public and self-built station datasets and can make accurate predictions in scenes with very high crowd density, large variations in crowd size, and severe background interference.

4.5. Ablation Experiments. To verify the effectiveness of the modules in the network, ablation experiments were conducted in Part_A of the ShanghaiTech dataset. The backbone network is denoted as Backbone, the traditional feature pyramid structure is denoted as FPN, the feature-enhanced pyramid structure is denoted as FEPN, and the mixed attention mechanism is denoted as MA. The experimental results are shown in Table 4. The comparison between the first two rows and the last two rows shows that the embedding of mixed attention improves the counting accuracy and robustness of the network, indicating that fully utilizing global contextual information works well in crowd counting studies. The comparison between the first and third rows illustrates that the feature-enhanced pyramid structure with channel transformation and semantic consistency learning brings about an improvement in network performance compared to the traditional feature pyramid structure. The loss and accuracy convergence curves of the ablation experiment are shown in Figure 8; in order to ensure the simplicity and readability of the image, the training loss curve and the test loss curve are presented in two figures, and the training accuracy curve and the test accuracy curve are also presented in two figures.

The feature-enhanced pyramid structure proposed in this paper is improved on the traditional feature pyramid structure. While the model achieves excellent performance, it also needs to pay attention to whether this improvement brings redundant calculation. The number of model parameters reflects the calculation amount and running time of the model to a certain extent. Therefore, this paper analyzes the improvement of the feature pyramid structure based on the number of model parameters. As shown in the last column of Table 4, the comparison of the first and third rows shows that the improvement of the feature pyramid brings less than 1MB increase in parameters, which proves that the feature-enhanced pyramid network algorithm does not bring redundant calculation while improving the network counting accuracy.

Figure 8 shows the loss convergence curve and accuracy convergence curve of the model. In the early stage, the fluctuation of training loss is large, mainly because the parameter learning of the network is not yet completed and the model is disturbed by useless information. As the learning proceeds, the training loss curve tends to be stable and converges, indicating that the model has effectively completed the learning. The accuracy convergence curve indicates that the parameters of the model are well set and learned, and the counting performance of the model is good.

5. Conclusion

Based on the problems of large changes in crowd scale and strong background interference in subway platform passenger flow counting, the algorithm proposed in this paper uses a feature-enhanced pyramid structure to retain channel information and eliminate aliasing effects. The enhanced feature representation is more conducive to solving the scale problem. By embedding a mixed attention module in the algorithm, the idea of nonlocal image processing is used to capture the global context information, to obtain a high-quality prediction density map. The algorithm achieves good results on the two public datasets and the self-built station dataset, which proves the effectiveness of the algorithm in this paper. However, there are still some shortcomings in the study. For example, reflections of passengers from platform screen doors may lead to repeated counts and thus large predictions, and preprocessing of the images to cover or cut sections of screen doors with severe reflections will be considered in the future. For the problem that passengers are completely occluded by pillars or other passengers on the platform, resulting in missed detection and small prediction results, the idea of the target detection algorithm can be used for reference in the future to reduce the impact of occlusion on crowd detection from the loss function.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon reasonable request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This study was supported by the National Natural Science Foundation of China (no. 52262045) and Gansu Provincial Key R&D Program-Industrial Project Funding (23YFGA0045).

References

[1] H. Wang, *Optimization for Train Service Plan with Full-Length and Short-Turn Routes on a Metro Line*, Beijing Jiaotong University, Beijing, China, 2021.

[2] P. Xu, T. Liu, and B. Si, “Modeling and empirical study on travel preference of morning rail transit commuters using smart card data,” *Systems Engineering-Theory & Practice*, vol. 43, no. 5, pp. 1484–1498, 2023.

[3] G. Lu, Y. Lei, and H. Zhang, “Passenger flow control strategy of urban rail transit based on multi-agent simulation,” *Journal of Tongji University*, vol. 50, no. 08, pp. 1189–1197, 2022.

[4] P. Dollar, C. Wojek, B. Schiele, and P. Perona, “Pedestrian detection: an evaluation of the state of the art,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 4, pp. 743–761, 2012.

[5] Y. Zhang, G. Li, J. Lei, and J. He, “FF-CAM: Crowd counting based on frontend-backend fusion through channel-attention mechanism,” *Chinese Journal of Computers*, vol. 44, no. 2, pp. 304–317, 2021.

[6] A. Chan and N. Vasconcelos, “Counting people with low-level features and Bayesian regression,” *IEEE Transactions on Image Processing*, vol. 21, no. 4, pp. 2160–2177, 2012.

[7] Y. Wang and Y. Zou, "Fast visual object counting via example-based density estimation," in *Proceedings of the 2016 IEEE International Conference on Image Processing (ICIP)*, pp. 3653–3657, IEEE, Phoenix, AZ, USA, September 2016.

[8] V. Pham, T. Kozakaya, O. Yamaguchi, and R. Okada, "Count forest: Co-voting uncertain number of targets using random forest for crowd density estimation," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3253–3261, IEEE Computer Society, Santiago, Chile, December 2015.

[9] C. Wang, H. Zhang, L. Yang, S. Liu, and X. Cao, "Deep people counting in extremely dense crowds," in *Proceedings of the 23rd ACM international conference on Multimedia*, pp. 1299–1302, ACM, Brisbane, Australia, October 2015.

[10] C. Zhang, H. Li, X. Wang, and X. Yang, "Cross scene crowd counting via deep convolutional neural networks," in *Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 833–841, IEEE, Boston, MA, USA, June 2015.

[11] Y. Zhang, D. Zhou, S. Chen, S. Gao, and Y. Ma, "Single-image crowd counting via multi-column convolutional neural network," in *Proceedings of the 2016 IEEE conference on computer vision and pattern recognition*, pp. 589–597, IEEE, Las Vegas, NV, USA, June 2016.

[12] D. Sam, S. Surya, and R. Babu, "Switching convolutional neural network for crowd counting," in *Proceedings of the 2017 IEEE conference on computer vision and pattern recognition*, pp. 5744–5752, IEEE, Honolulu, HI, USA, July 2017.

[13] V. Sindagi and V. Patel, "Generating high-quality crowd density maps using contextual pyramid CNNs," in *Proceedings of the 2017 IEEE international conference on computer vision*, pp. 1879–1888, IEEE, Venice, Italy, October 2017.

[14] Y. Li, X. Zhang, and D. Chen, "CSRNet: dilated convolutional neural networks for understanding the highly congested scenes," in *Proceedings of the 2018 IEEE/CVF conference on computer vision and pattern recognition*, pp. 1091–1100, IEEE, Salt Lake City, UT, USA, June 2018.

[15] J. Dai, H. Qi, Y. Xiong et al., "Deformable convolutional networks," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 764–773, IEEE, Piscataway, NJ, USA, January 2017.

[16] I. Goodfellow, J. Pouget-Abadie, M. Mirza et al., "Generative adversarial nets," in *Proceedings of the 27th Int Conference on Neural Information Processing Systems*, pp. 2672–2680, MIT Press, Cambridge, MA, USA, October 2014.

[17] Z. Sheng, K. Tian, Q. Tian, and H. Qu, "A faster R-CNN based high-normalization sample calibration method for dense subway passenger flow detection," in *Proceedings of the 2018 11th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, pp. 1–5, IEEE, Beijing, China, October 2018.

[18] J. Zhang, J. Liu, and Z. Wang, "Convolutional neural network for crowd counting on metro platforms," *Symmetry*, vol. 13, no. 4, p. 703, 2021.

[19] Q. Xiao, Y. Xiao, and F. Chen, "The passenger flow counting research of subway video based on image processing," in *Proceedings of the 2017 29th Chinese Control And Decision Conference (CCDC)*, pp. 5195–5198, IEEE, Chongqing, China, May 2017.

[20] X. Hu, H. Zheng, W. Wang, and X. Li, "A novel approach for crowd video monitoring of subway platforms," *Optik*, vol. 124, no. 22, pp. 5301–5306, 2013.

[21] G. He, Q. Chen, D. Jiang, X. Lu, and Y. Yuan, "A double-region learning algorithm for counting the number of pedestrians in subway surveillance videos," *Engineering Applications of Artificial Intelligence*, vol. 64, pp. 302–314, 2017.

[22] J. Zhang, G. Zhu, and Z. Wang, "Multi-column atrous convolutional neural network for counting metro passengers," *Symmetry*, vol. 12, no. 4, p. 682, 2020.

[23] Q. Guo, Q. Liu, W. Wang, Y. Zhang, and Q. Kang, "A fast occluded passenger detector based on MetroNet and Tiny MetroNet," *Information Sciences*, vol. 534, pp. 16–26, 2020.

[24] Q. Liu, Q. Guo, W. Wang, Y. Zhang, and Q. Kang, "An automatic detection algorithm of metro passenger boarding and alighting based on deep learning and optical flow," *IEEE Transactions on Instrumentation and Measurement*, vol. 70, pp. 1–13, 2021.

[25] J. Yang, Y. Zheng, K. Yan et al., "SPDNet: a real-time passenger detection method based on attention mechanism in subway station scenes," *Wireless Communications and Mobile Computing*, vol. 2021, Article ID 7978644, 13 pages, 2021.

[26] J. Yang, M. Gong, X. Dong, J. Liang, and Y. Wang, "MPDNet: a Transformer-based real-time passenger detection network in metro stations," *Frontiers in Physics*, vol. 10, Article ID 1017951, 2022.

[27] T. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 936–944, IEEE, Honolulu, HI, USA, July 2017.

[28] J. Xie, C. Pang, Y. Zheng et al., "Multi-scale attention recalibration network for crowd counting," *Applied Soft Computing*, vol. 117, Article ID 108457, 2022.

[29] M. Wang, H. Cai, X. Han, J. Zhou, and M. Gong, "STNet: scale tree network with multi-level auxiliator for crowd counting," *IEEE Transactions on Multimedia*, vol. 25, pp. 2074–2084, 2023.

[30] Q. Song, C. Wang, Y. Wang et al., "To choose or to fuse? Scale selection for crowd counting," in *Proceedings of tThe Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI-21)*, pp. 2576–2583, AAAI Press, California, CA, USA, February 2021.

[31] J. Ma, Y. Dai, Z. Jia, F. Sun, Y. Tan, and J. Liu, "Crowd counting from single images using recursive multi-pathway zooming and foreground enhancement," *Pattern Recognition*, vol. 141, Article ID 109585, 2023.

[32] S. Zhang, W. Zhao, L. Wang, W. Wang, and Q. Li, "MSIANet: multi-scale interactive attention crowd counting network," *Journal of Electronics and Information Technology*, vol. 45, no. 06, pp. 2236–2245, 2023.

[33] J. Niu, G. Li, and Y. Yu, "Crowd counting method based on feature fusion and attention mechanism," in *Proceedings of 2021 IEEE International Conference on Artificial Intelligence and Industrial Design (AIID)*, pp. 673–677, IEEE, Guangzhou, China, May 2021.

[34] L. Wang, J. Li, S. Zhang, C. Qi, P. Wang, and F. Wang, "Multi-Scale and spatial position-based channel attention network for crowd counting," *Journal of Visual Communication and Image Representation*, vol. 90, Article ID 103718, 2023.

[35] P. Li, M. Zhang, J. Wan, and M. Jiang, "Multi-scale guided attention network for crowd counting," *Scientific Programming*, vol. 2021, Article ID 5596488, 13 pages, 2021.

[36] Q. Hou and D. F. J. Zhou, "Coordinate attention for efficient mobile network design," in *Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13708–13717, IEEE, Nashville, TN, USA, June 2021.

[37] T. Wang, T. Zhang, K. Zhang, H. Wang, M. Li, and J. Lu, "Context attention fusion network for crowd counting," *Knowledge-Based Systems*, vol. 271, Article ID 110541, 2023.

[38] S. Chen, J. Zhao, Y. Zhou et al., "Info-FPN: an Informative Feature Pyramid Network for object detection in remote sensing images," *Expert Systems with Applications*, vol. 214, Article ID 119132, 2023.

[39] S. Gu and Z. Lian, "A unified RGB-T crowd counting learning framework," *Image and Vision Computing*, vol. 131, Article ID 104631, 2023.

[40] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7794–7803, IEEE, Salt Lake City, UT, USA, June 2018.

[41] W. Liu, M. Salzmann, and P. Fua, "Context-aware crowd counting," in *Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5094–5103, IEEE, Long Beach, CA, USA, June 2019.

[42] T. Xu, Y. Duan, J. Du, and C. Liu, "Crowd counting method based on multi-scale enhanced network," *Journal of Electronics and Information Technology*, vol. 43, no. 6, pp. 1764–1771, 2021.

[43] G. Liu, Q. Wang, X. Chen, and Y. Meng, "A multivariate information aggregation method for crowd density estimation and counting," *Optics and Precision Engineering*, vol. 30, no. 10, pp. 1228–1239, 2022.