Hindawi
Mathematical Problems in Engineering
Volume 2020, Article ID 6692257, 11 pages
https://doi.org/10.1155/2020/6692257

Research Article

Crowd Counting and Abnormal Behavior Detection via Multiscale GAN Network Combined with Deep Optical Flow

Beibei Song[1] **and Rui Sheng**[2]

[1]*Jiuzhou Polytechnic, Xuzhou 221116, China*
[2]*Southwest China Institute of Electronic Technology, Chengdu 610036, China*

Correspondence should be addressed to Beibei Song; songbeibei@jzp.edu.cn

Received 5 November 2020; Revised 2 December 2020; Accepted 7 December 2020; Published 16 December 2020

Academic Editor: Yi-Zhang Jiang

Aiming at the problem of low performance of crowd abnormal behavior detection caused by complex backgrounds and occlusions, this paper proposes a single-image crowd counting and abnormal behavior detection via multiscale GAN network. The proposed method firstly designed an embedded GAN module with a multibranch generator and a regional discriminator to initially generate crowd-density maps; and then our proposed multiscale GAN module is added to further strengthen the generalization ability of the model, which can effectively improve the accuracy and robustness of the prediction detection and counting. On the basis of single-image crowd counting, synthetic optical-flow feature descriptor is adopted to obtain the crowd motion trajectory, and the classification of abnormal behavior is finally implemented. The simulation results show that the proposed algorithm can significantly improve the accuracy and robustness of crowd counting and abnormal behavior detection in real complex scenarios compared with the existing mainstream algorithms, which is suitable for engineering applications.

1. Introduction

With the expansion of urban scale and the increase of crowd, the probability of traffic accidents, congestion, stampede, and other emergencies in public place also increases [1, 2]. In order to properly deal with emergencies and ensure the safety of public places, the development of intelligent surveillance technology is becoming more and more important [3]. However, while surveillance systems are cheap and common, the cost of hiring the right people to observe and analyze recorded videos is still very high [4]. Therefore, real-time analysis of abnormal behavior in public places is particularly important for intelligent surveillance system because abnormal behavior can be prevented and stopped only when the surveillance system has the ability of understanding human behavior [2, 5].

Abnormal behavior detection is an important and challenging research direction in the field of computer vision, which is the foundation of scene understanding, visual object tracking, and other applications [6]. For a given video or surveillance image sequences, crowd abnormal behavior detection is to extract and classify the specific information representing the abnormal behavior of the crowd in the video sequences, such as crowd density and group behavior characteristics [7]. Abnormal behavior detection needs to use the visual cues of the object, extract the relevant features of the object from the videos, and analyze their behavior state. It is a difficult problem that requires high hardware resources and a lot of original innovation to solve. Recently, it is still a great challenge for people to solve the problems in complex environment by using intelligence algorithm [7].

Crowd abnormal behavior detection mainly includes feature extraction, feature fusion, and behavior classification. In feature extraction, most of the existing models select appropriate low-level features and use deep learning to extract high-level features based on low-level features and then fuse the extracted features to form a relatively representational spatiotemporal feature, which better represents the behavior of pedestrians' object [8]. Finally, according to the extracted features, a suitable classifier is designed to

judge the crowd behavior represented by the features and the correct behavior detection results are output [9]. In recent years, in order to improve the performance of crowd abnormal behavior detection caused by explosions, terrorist attacks, and other emergencies, many crowd abnormal behavior detection algorithms based on video sequences have been proposed. These algorithms can be roughly divided into two categories: visual feature extraction method and physical feature analysis method [10–19]. The former uses vision and image processing technology to extract crowd features and then detect anomalies. Literature [11] uses Granger's causality and dynamic temporal planning to describe the group relevance characteristics and then uses SVM to detect the anomaly behavior. This method can intuitively reflect the shape of the crowd, but because of its single information and incomplete features, it has problems such as poor accuracy, low training efficiency, and limited data processing capabilities. In order to improve the behavior characteristics and the detection accuracy, literature [12] proposed a crowd abnormal behavior detection based on Bayesian model (BM). The concept of potential object and divergence center was introduced to represent crowd movement tendency so as to improve the integrity of behavior characteristics. However, crowd with high-density are easily affected by interference factors such as crowd occlusion and illumination changes, which lead to the decrease of abnormal behavior detection rate. The latter is based on physical characteristics analysis method, which constructs physical model to simulate and detect crowd behavior. Literature [13] proposed a social attribute perception based on social force model (SFM), which uses social barriers and congestion attributes to describe the interaction between behaviors in the crowd so as to further describe the behavior of pedestrians. However, the model has many parameters and poor real-time performance. To solve this problem, literature [14] combined the demographic results with the crowd entropy on the basis of the energy model (EM) and set the threshold value of the crowd distribution index to detect the crowd aggregation state, so as to detect the abnormal behavior of the crowd. The proposed method is robust to training data, but it needs specific preprocessing to estimate the threshold, which leads to high complexity. In order to reduce the computational complexity, literature [15] designed a group descriptor with low computational complexity on the basis of topological structure of quantized crowd manifold to detect abnormal behavior. The clustering model (CM) has strong representation ability for high-density crowd, but as the number of pedestrians in the group decreases, the accuracy of the behavioral consistency estimation decreases, which leads to the significant decrease of the representation ability. To solve this problem, according to the consistent characteristics of particle behavior, literature [16] proposed a global direction descriptor to extract the overall motion of the group, and then the local and global descriptors were fused to establish a direction-cluster model, which enhances the model's ability to characterize crowd characteristics. However, due to its excessive clustering of the direction of movement, the performance is significantly reduced when detecting crowd behaviors with confusing directions. In order to solve the above problems, some scholars [18] proposed a crowd abnormal behavior detection method based on synthetic optical-flow feature descriptor and trajectory in single image. The proposed method obtains the direction, speed, acceleration, and energy of crowd movement according to the optical-flow field changes of crowd movement in the video and then weights the obtained features to improve the representation ability of instantaneous characteristics of crowd behavior; in addition, the crowd motion trajectory is extracted to represent the continuous characteristics of crowd behavior. Finally, the instantaneous and continuous features of abnormal crowd movement are input into the two-stream convolution neural network, and a fusion layer is added after the convolution layer to form a complete spatiotemporal representation, so as to improve the performance of crowd abnormal behavior detection [19].

As the performance of the deep learning model improves, the accuracy of pedestrian detection in the crowd is also greatly improved. Therefore, we can consider improving the abnormal behavior detection performance based on pedestrian detection and counting. The proposed method firstly designed an embedded GAN module with a multi-branch generator and a regional discriminator to initially generate crowd density maps; and then our proposed multiscale module is added to further strengthen the generalization ability of the model; finally, synthetic optical-flow feature descriptor is adopted to obtain the crowd motion trajectory, and the classification of abnormal behavior is implemented.

2. Multiscale GAN Network

Crowd abnormal behavior detection is the use of CCTV cameras installed in public places to capture and detect abnormal events and then issue timely warnings. It has a wide range of applications in the field of intelligent surveillance. In recent years, the crowd abnormal behavior detection method has achieved good development, and many excellent detection algorithms have been proposed. However, many challenging problems have not been effectively solved, such as changes in crowd occlusion and complex backgrounds, leading to greatly reduced accuracy and robustness of crowd abnormal behavior detection algorithms. Therefore, detecting abnormal behavior of the crowd is still a difficult task. In order to improve the accuracy and robustness of crowd abnormal behavior detection algorithms, this paper proposes a crowd counting model based on multiscale network. The model can be regarded as an embedded GAN structure. The embedded GAN module learns crowd features and optimizes the local correlation of images. The scale module further extracts local multiscale features and generates the final crowd density image. The structure of the proposed model is shown in Figure 1, which consists of three parts: generation network, discrimination network, and scale module. The generation network and discrimination network are embedded in the whole model to construct an embedded GAN module, where the generation network is composed of partial structure of VGG-16

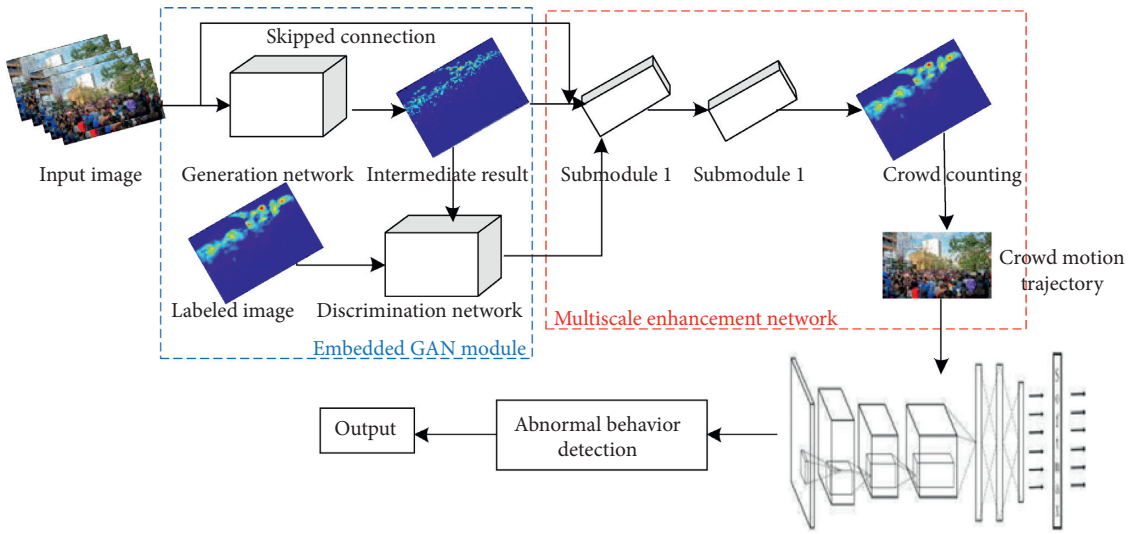

Figure 1: The proposed framework for single-image crowd counting and abnormal behavior detection.

backbone model and multibranch dilation convolution structure, and the discrimination network only supervises the generation of intermediate results. In addition, the model adopts skipped connection to keep the structure and context information of the input image.

2.1. Generation Network. Inspired by literature [10], this paper constructed the backbone of the generation network based on VGG-16 model. The model has strong feature extraction ability and transfer learning ability, which is conducive to feature extraction of complex crowd. The original VGG-16 model contains 13 convolution layers and 5 pooling layers, so the size of the deep feature map of the network is very small, which is not conducive to the modeling of small-scale object. In order to avoid the information loss of small-scale objects caused by oversampling, this paper first removes the fully connected layer of the original VGG-16 model and then uses its first 10 convolution layers and three pooling layers to build the network backbone. In addition, in order to aggregate more abundant multiscale information, a multibranch structure is designed to build the back end of the generation network. The multibranch structure is designed based on the dilation convolution and can expand the perception range of the network without increasing the amount of parameters, which is conducive to coping with changes in the size and scale of the crowd between images. The back-end network is composed of three branches; each branch contains the dilation convolution with different expansion factors, and the expansion factors are 1, 2, and 4. The branch with expansion factor of 1 is used to capture the features of small-scale objects, while the other branches expand the perception range to capture the features of large-scale objects. As mentioned in literature [17], it is difficult for independent branches to learn the characteristics of different patterns, which leads to parameter redundancy. Therefore, in this paper, the feature maps of each branch network are concatenated in each layer, and 1×1 convolution is used for cross-channel feature aggregation to strengthen the information interaction between each branch, so as to make full use of the complementarity of each branch extraction feature to make the output feature map has more expressive power and scale diversity. The specific structure of the generation network is shown in Figure 2. The parameters in the box in Figure 2 are represented as "convolution layer-convolution kernel size-channel number-expansion factor."

2.2. Discrimination Network. The regional discrimination network was first applied to image conversion tasks. This paper uses PatchGAN to construct the discriminant network in the embedded GAN module, and its specific structure is expressed as follows: C(4,64,2)-C(4,128, 2)-C(4,256,2)-C(4,512, 1)-C(4,1,1), where C represents the convolution layer, and the parameters in parentheses are the size of the convolution kernel, the number of channels, and the convolution step-length. Except for the last layer, after each convolutional layer, Batch Normalization (BN) and LeakyReLU activation functions are added. Different from the conventional discrimination network, our adopted network is a fully convolutional network, and its output is an $N \times N$ matrix instead of a scalar value. Each element in the matrix is mapped to a partial image patch of the original image, reflecting the authenticity of the image block. In view of this matrix calculation error, the network can be more focused on the local area of the image, which is beneficial to guide the generation network to obtain a crowd density image with higher local correlation.

2.3. Multiscale Network. The embedded GAN module described above learns crowd characteristics and optimizes the local correlation of density images. On this basis, a scale module is designed to further extract local features of different scales from different regions, thereby enhancing the generalization ability of the model.

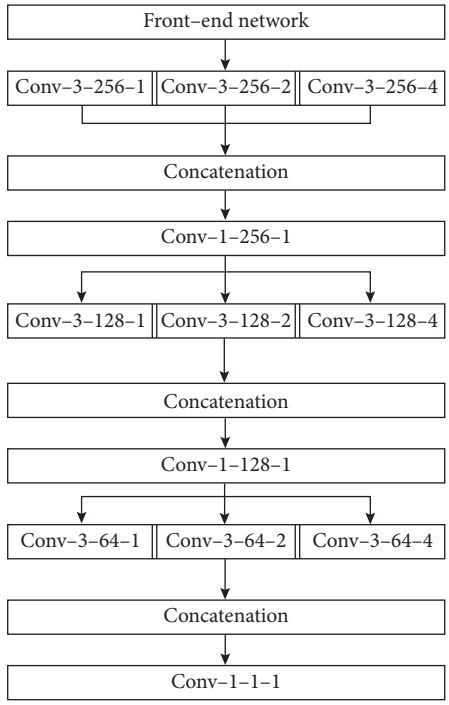

Figure 2: Specific structure of the generation network.

The scale module is composed of two submodules with the same structure in series, and the submodule is designed based on the pyramid pooling structure. As shown in Figure 3, for the input of the upper layer of the network, the submodule first performs feature extraction through two front-end convolutional layers with a size of 3×3 and then pools the output of the front-end convolution layer by four levels. Since the scene in the crowd image is very complex containing many objects, and the population size and scale show continuous changes, the global average pooling in the traditional pyramid pooling structure is not enough to reflect the respective scale characteristics of different objects, so the four levels of pool size are set as 2×2, 3×3, 6×6, and 8×8. The above operations divide the feature map into several subregions with different sizes according to the scale, and average pooling is used to reflect the local features of each subregion. Then, the size of the original feature map is sampled by bilinear interpolation operation and then concatenated with the original feature map. Finally, a 3×3 back-end convolution layer is used to aggregate the concatenated feature map across channels to generate the final output of the submodule.

In this paper, the original image is input into the first submodule after skipped connection, and the output of the first submodule is concatenated with the output of the embedded GAN module and then input into the second submodule. Through the above operations, the scale module can further extract local features of different scales from different regions to cope with the characteristics of continuous changes in the scale of the crowd and realize the generalization ability of the overall model.

2.4. Loss Function. The Euclidean loss function commonly used in crowd detection and behavior analysis assumes that the pixels are independent of each other, ignoring the local correlation of the image. Therefore, this paper uses three loss functions to jointly optimize the model, namely, L_1 loss, adversarial loss, and Euclidean loss. L_1 loss and adversarial loss constrain the initial prediction image produced by the embedded GAN module and optimize its local correlation to obtain the final prediction image of the Euclidean loss constraint model. The L_1 loss is defined as

$$L_i = \frac{1}{n} \sum_{i=1}^{n} \left\| G(x_i) - y_i \right\|_1^1, \tag{1}$$

where n is the number of training samples, x_i is the input image, y_i is the corresponding label image, and $G(x_i)$ is the intermediate prediction result of the model generated by the generation network according to the input image. The adversarial loss is defined as

$$\min_G \max_D L_A(G, D) = E_{y \sim P_{\text{data}(y)}} [\log D(y)] + E_{x \sim p_{\text{data}(x)}} [\log (1 - D(G(x)))], \tag{2}$$

where x is the input image; y is the corresponding label image; G and D are generation network and discrimination network, respectively; and $G(x)$ is the intermediate prediction result of the model generated by the generation network according to the input image. The Euclidean loss function is defined as

$$L_E = \frac{1}{n} \sum_{i=1}^{n} \left\| m_i - y_i \right\|_2^2, \tag{3}$$

where n is the number of training samples; m_i is the density image finally predicted by the model; and y_i is the corresponding label image. The three loss functions are weighted and combined to form the final objective function of the model, defined as

$$L = aL_A + \beta L_1 + L_E, \tag{4}$$

where α and β are the weight coefficients to balance the three losses.

2.5. Training Steps. Since the multiscale network designed in this paper is an embedded GAN structure, the overall model cannot follow the training steps of the traditional GAN model. Therefore, this paper adopts a new alternative training step to optimize the model. In this training step, the generation network will perform two parameter updates. The specific steps are as follows:

Step 1. Load the training dataset and perform data preprocessing.

Step 2. Initialize model training parameters and input training data.

Step 3. Improve the gradient of equation (2) to update the parameters of the discriminant network.

Step 4. Reduce the gradient of equations (1) and (2) to update the parameters of the generated network.

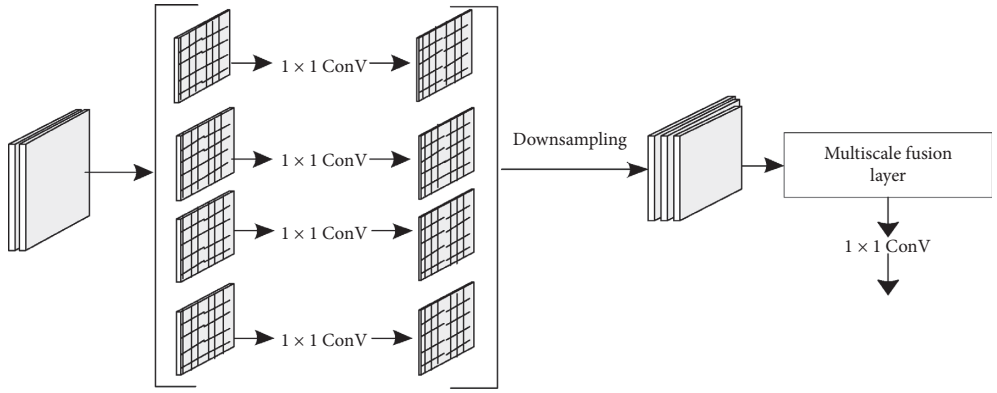

Figure 3: Multiscale enhancement module.

Step 5. Reduce the gradient of equation (3) to update the parameters of the generation network and the scale module, respectively.

Step 6. Repeat steps 3 to 5 until the end of training.

3. Deep Optical Flow for Abnormal Behavior Detection

Normally, the direction and speed of crowds are similar. However, when an abnormal event occurs, people will run away quickly due to fear to avoid potential danger. The abnormal behavior of the crowd has the characteristics of fast movement speed, sudden increase in acceleration, obvious concentration of movement in a certain direction or balance in multiple directions, and chaotic trajectory. The calculation of features such as speed, acceleration, direction, and motion amplitude is relatively simple and can be expressed by optical flow, while the extraction of features such as crowd expression is more complicated. In order to reduce the complexity of the proposed method, this paper extracts the trajectory characteristics of the crowd to detect the abnormal behavior.

As we all know, once the position of the pedestrian in the crowd is obtained, long-term tracking can be performed and the trajectory of the target can be calculated. However, due to the high density of pedestrians in the crowd and severe occlusion, it is impossible to track directly and accurately. Therefore, deep optical flow is introduced into counting the overall motion trajectory. Recently, the optical-flow calculation method based on convolutional neural network is the state-of-the-art algorithm. FlowNet uses a supervised method for deep learning training on the optical-flow prediction problem, and it is the first successful attempt to directly predict optical flow using a convolutional neural network [16]. As it involves pixel changes and predictions, the input of the optical-flow network is usually a pair of images, and the output is the corresponding optical-flow graph. Flow-Field refers to the two-dimensional field obtained by projecting the instantaneous speed of the corresponding pixel connection in the image-pair. The movement direction of the pixel connection is naturally distributed in the horizontal and vertical directions on the plane. The optical-flow map is an image representation of the optical-flow field. Obviously, the optical-flow map is a dual-channel image. The final output performance is that different colors mean the direction of movement, and the color depth means the speed of pixel movement. The structure of the deep optical flow for abnormal behavior detection is shown in Figure 4. The input of the contractive part of the optical-flow network is a pair of images, and the network uses a series of convolutional layers to extract feature maps. The simplest reduction method is to directly concatenate the channels of a pair of images. The number of input channels is 6. This contractive part of the network is composed of 9 convolutional layers, and the size of the convolution kernel decreases with the depth of the network, being set to 7×7, 5×5, 5×5, and 3×3, respectively.

The network of the expanding part in optical-flow network is mainly constructed by 4 deconvolution layers. It is mainly reflected in the magnified feature map. This part is similar to the full convolutional network. With backward deconvolution, the prediction is directly performed on the small feature map. After the prediction result is obtained, it is bilinearly interpolated and then concatenated on the deconvolved feature map and then forwarded back, repeating four times. The resolution of the predicted optical flow is still one-fourth of the input, and the direct bilinear interpolation obtains the optical-flow prediction map with the same resolution as the input.

The first layer of the contractive part of the network designed in this paper is a 7×7 convolution kernel, the second and third layers are 5×5 convolution kernels, and the following 7 layers are all 3×3 convolution kernels. The network expanding part is constructed by 4 deconvolution layers. The Deconvolution() is used to construct the deconvolution, the *Crop()* function is used to trim the data, and the Concat() function is used to concatenate the different channels. The deconvolution layer uses upsampling to restore original resolution, which is the opposite of subsampling in the convolution operation. The method generally uses 0 filling method, interpolation method, and other operations to enhance image resolution. Once we get the optical-flow trajectory, we can use a simple classification model to classify abnormal behavior.

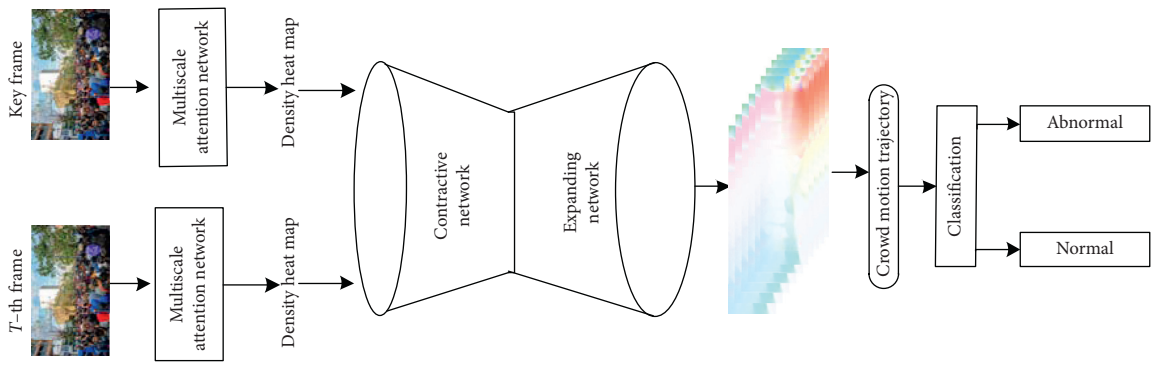

Figure 4: Deep optical flow for abnormal behavior detection.

4. Experiment Results and Analysis

4.1. Dataset. In order to verify the effectiveness of the model, this experiment uses three datasets commonly used in crowd counting research to conduct experiments, namely, ShanghaiTech, UCF_CC_50, and UCF-QNRF. The ShanghaiTech dataset contains 1198 crowd images, of which 330165 object pedestrians are marked. The dataset is divided into two parts, Part_A and Part_B. Part_A contains a total of 482 crowd images collected on the Internet, which are specifically divided into 300 training images and 182 testing images. Part_B contains a total of 716 crowd images taken in a pedestrian street in Shanghai, which are specifically divided into 400 training images and 316 testing images. Relatively speaking, the crowd in the image scene of Part_B is relatively sparse. UCF_CC_50 dataset contains a total of 50 images collected on the Internet, of which 63,075 object pedestrians are marked. The image contains an average of 1280 people, and a single image contains 94~4543 people. This dataset contains a small amount of crowd, so the 5-fold cross-validation method is used to conduct experiments. The UCF-QNRF dataset contains a total of 1535 crowd images, of which 1.25×10^6 object pedestrians are marked. The dataset is specifically divided into 1201 training images and 334 test images. A single image contains 49~12865 people. The basic information of the above three datasets is shown in Table 1.

4.2. Evaluation Criteria. This paper uses two evaluation indicators commonly used in crowd counting to evaluate the performance of the model, namely, Mean Absolute Error (MAE) and Mean Square Error (MSE). MAE reflects the accuracy of model prediction, and MSE reflects the robustness of model prediction. The lower the two values, the better the model performance.

4.3. Parameter Setting. The experimental environment used in this paper is Intel Xeon® Sliver 4110 2.10 GHz CPU, Quadro P5000 GP (16 G memory). The operating system used is Ubuntu 16.04, and the deep learning framework used is the PyTorch framework. This paper uses the VGG-16 model parameters pretrained on the ImageNet dataset to initialize the front end of the generation network, and the parameters of the remaining networks are randomly initialized with a Gaussian distribution with a mean value of 0 and a standard deviation of 0.01. The model is optimized by Adam algorithm, the learning rate is fixed at 0.0000001, and the total number of iterations is 30,000. For ShanghaiTech Part_A, UCF_CC_50, and UCF-QNRF datasets, this experimental uses geometrically adaptive Gaussian kernels to produce label density images [20]; for ShanghaiTech Part_B datasets, because of the sparse crowd in the images, this paper uses fixed Gaussian kernels to produce label density image. In addition, for the ShanghaiTech and UCF_CC_50 datasets, this experiment uses the original image size for training, sets the batch size to 1, and performs data through random horizontal flips. Since the UCF-QNRF dataset is all high-resolution images (such as 9000×6000), this paper follows the training method proposed in [17] and crops the original image into 16 nonoverlapping subimages with a size of 224×224, and the batch size is 16 for training.

4.4. Analysis of Experimental Results. The experimental results of the different benchmark dataset are shown in Table 2, where we select TEDNet [11], PACNN [16], SANet [18], CSRNet [21], ACSCP [22], S_CNN [23], and MCNN [24] as comparison algorithms. This experiment compares the model with 7 existing state-of-the-art methods for crowd counting research in recent years. For Part_A, the model has obtained the lowest MAE value, which is 1.1% lower than the TEDnet, and the MSE value of the proposed model is also close to the ACSCP method, which performs the best in this index. For Part_B, the model could have obtained the lowest MAE value and MSE value, respectively, in which the MAE index was the same as the TEDNet, and the MSE index was reduced by 3.9% compared to the TEDNet. The experimental results in the two parts of the ShanghaiTech dataset show that the proposed model has good performance in both crowded and sparse crowd scenarios.

The experimental results of the UCF_CC_50 dataset are also shown in Table 2. This experiment also compares the model with 7 existing state-of-the-art methods for population counting research in recent years. The model achieved the lowest values on both MAE and MSE indicators. Compared with the TEDNet, the MAE indicator was reduced by 9.1%, and the MSE indicator was reduced by

Table 1: Benchmark dataset for crowd counting and abnormal behavior detection.

Dataset	Number	Size	Min_Crowd	Max_Crowd	Density
Part_A	482	—	24	2144	High
Part_A	716	1024×768	8	550	Low
UCF_CC_50	45	—	64	3530	Extremely high
UCF-QNRF	1500	—	48	870	Extremely high

Table 2: Quantitative result for different benchmark datasets.

Models	Part_A		Part_B		UCF_CC_50	
	MAE	MSE	MAE	MSE	MAE	MSE
TEDNet	64.85	109.25	88.25	12.85	249.25	354.62
PACNN	66.18	106.24	21.552	3.56	267.9	357.06
SANet	66.39	104.51	8.54	13.65	258.44	334.86
CSRNet	68.27	115.27	21.62	33.52	266.15	397.17
ACSCP	76.58	102.75	17.55	27.58	292.08	404.87
S_CNN	90.2	135.22	8.922	135.18	318.27	439.57
MCNN	110	173.25	8.28	12.84	377.78	509.21
Proposed	63.4	106.23	8.21	12.66	266.71	310.65

12.4%. The dataset contains a small number of samples, only 50 images. The experimental results show that the model can also show good adaptability to small sample data.

The UCF-QNRF dataset is one of the latest datasets released in 2018 [25]. Currently, there are relatively few evaluation methods using this dataset. Therefore, we compare the model with four existing state-of-the-art methods, which are MCNN, SCNN, CL, and TEDNet, respectively. The results are shown in Table 3. The model obtains a competitive MAE value, while obtaining the lowest MSE value. Compared with the TEDNet, the MAE index of the model is reduced by 15.3%, and the MSE index is also close to it. This dataset has the characteristics of a large number of samples and complex scenes. In this case, the prediction accuracy of the model needs to be improved. In addition, the prediction robustness of the proposed model is good, indicating that it has good generalization ability.

Table 3: Quantitative result for UCF-QNRF dataset.

Models	UCF-QNRF	
	MAE	MSE
TEDNet	113.85	188.25
S_CNN	230.2	435.22
MCNN	270.25	423.25
Proposed	113.41	156.23

4.5. Ablation Analysis. In order to further verify the effectiveness of the structure of each part of the proposed model, this paper designs a model structure ablation analysis on the basis of the Part_A dataset, specifically focusing on three factors of the model structure: embedded GAN structure, the number of scale submodules, and skipped-connection settings. In order to balance model performance and resource overhead, the maximum number of scale submodules is limited to 2 [26]. Specifically, this paper constructs 10 models with different structures based on the principle of permutation and combination and shows the specific description and corresponding results of each model in Table 4, where the scale submodule is denoted as E; the skipped-connection is recorded as S. These combinations of models are described as follows. (a) Only the generation network is contained, which is denoted as$G$. (b) On on the basis of model (a), a discriminant network is added to form a generative confrontation network, which is recorded as GAN. (c ~ f) The model structures are all nonembedded GAN structures (respectively corresponding to the embedded GAN structures of (g ~ j)), denoted as GAN∗. In this type of model, this paper combines the original generation network with the scale module and treats the combined overall structure as an independent generation network, and it uses the discriminant network to directly supervise the final output of the model; (g) on the basis of embedded GAN structure, a standard scale submodule is connected. (h) On the basis of model (g), a skipped-connection setting is added. (i) On the basis of embedded GAN structure, two standard scale submodules are connected. (j) On the basis of model (i), a skipped-connection setting is added, which is the multiscale enhanced network model proposed in this paper.

It can be seen from Table 4 that the performance of model (b) is better than model (a), indicating that the introduction of regional discriminant network can optimize the local correlation of images and improve the accuracy of crowd counting; the performance of models (d) and (h) is better than that of models (e) and (g), which shows that the use of skipped-connection settings is helpful to reconstruct the structure and global context information of the input image; the performance of model (9) is better than that of model (g), indicating that the use of two-scale submodules is more conducive to the multiscale local features of each region in crowd image; under the premise of having the same configuration, the performance of the model using the embedded GAN structure is better than the corresponding nonembedded GAN structure model, and the models (e) and (f) have the worst performance among all

Table 4: Ablation analysis for different models.

Sequence number	Structural description	Mode	Number of MSEs	Skipped connection	MAE
a	G	—	—	—	67.51
b	GAN	—	—	—	65.65
c	$GAN{*}GAN{*}(E\times 1)$	—	1	—	65.31
d	$GAN{*}GAN{*}(E\times 1+S)$	—	1	√	66.59
e	$GAN{*}GAN{*}(E\times 2)$	—	2	—	66.47
f	$GAN{*}GAN{*}(E\times 2+S)$	—	2	√	65.05
g	$E\_GAN+E\times 1$	√	1	—	64.78
h	$E\_GAN+E\times 1+S$	√	1	√	64.13
i	$E\_GAN+E\times 2$	√	2	—	64.75
j	$E\_GAN+E\times 2+S$ (proposed)	√	2	√	63.52

(a) (b) (c)

Figure 5: Predicted results for our proposed model. (a) Single image. (b) Density map. (c) Label result for crowd.

models. The reason may be that the structure of the independent generation network formed by the combination of the original generation network and the scale module is more complicated, and the amount of parameters is too large, which makes the overall model difficult to converge during training [27]. Therefore, it also proves that the use of embedding GAN structure can effectively improve the performance of the model.

The predicted results of the crowd counting algorithm proposed in this paper are shown in Figure 5, where Figure 5(a) is the original image. Figure 5(b) is the result of the density map; Figure 5(c) shows the labeled results of the crowd. Two representative images are selected to carry out crowd counting. The first image is the street scene in the Part_A dataset. The object in the crowd is heavily occluded. The pedestrians close to the camera are more scattered, so the density is lower, and the pedestrian density is higher in the position far from the camera [28]. The crowd density map predicted by the network in this paper has a smaller error compared with the real density map, and the number of pedestrians predicted by our proposed network is greatly improved compared with the traditional method. The real number is 227, and the predicted value is 224.24, which is only 3 pedestrians away from the real value. The second image is a photo of the assembly, full of people. The crowd far away from the lens is very blurry. It can be seen from the results in Figure 5(c) that the crowd counting proposed in this paper can basically detect pedestrians, and the number of pedestrians is closer to the true value.

In addition, in order to further prove the effectiveness of connecting the multiscale module after the GAN module, this paper compares the results of the model (b) and the model (j) predicting the image in Figure 6. The structures of the two are the GAN structure and the proposed structure, where the only difference is whether the model includes a scale module. It can be seen that the image predicted by the model (j) can better reflect the density map of the crowd distribution, and the number of the calculated people is closer to the number of people actually contained in the label image according to the predicted image. So, the effectiveness of the proposed multiscale module is further proved.

In order to facilitate the analysis of the behavior detection performance of different algorithms, this experiment mainly divides the behavior into two states, normal and abnormal, and the detection results are shown in Figure 7. It

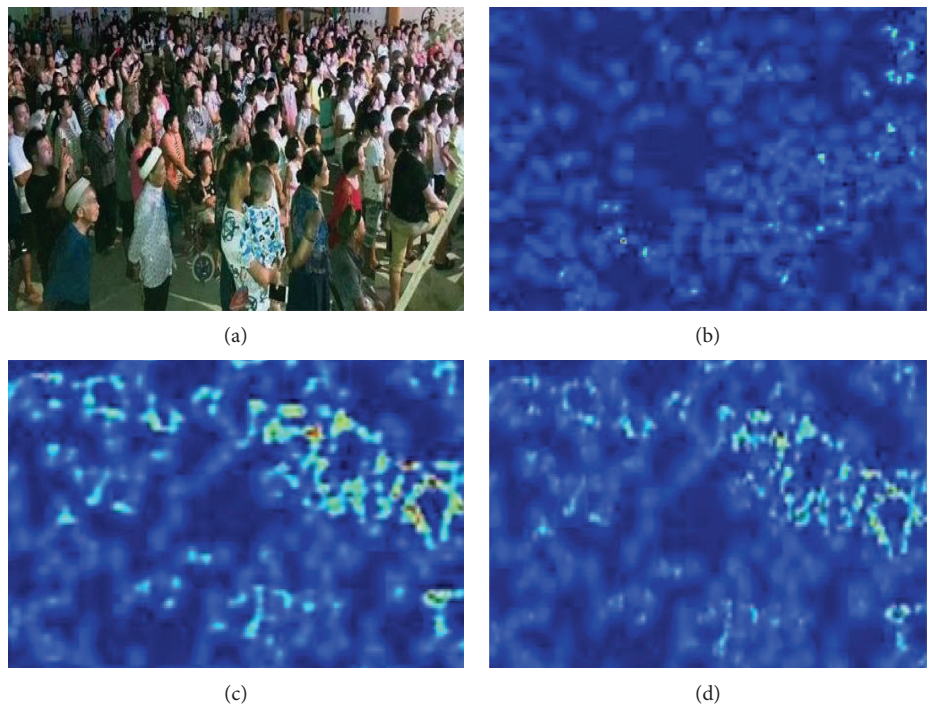

Figure 6: Crowd counting results under different module combinations. (a) Single image. (b) GAN. (c) Proposed. (d) Ground-truth density map.

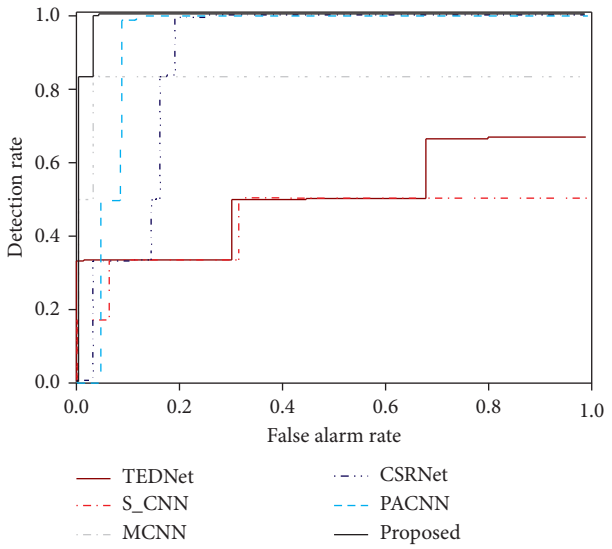

Figure 7: Comparative analysis of abnormal behavior detection results.

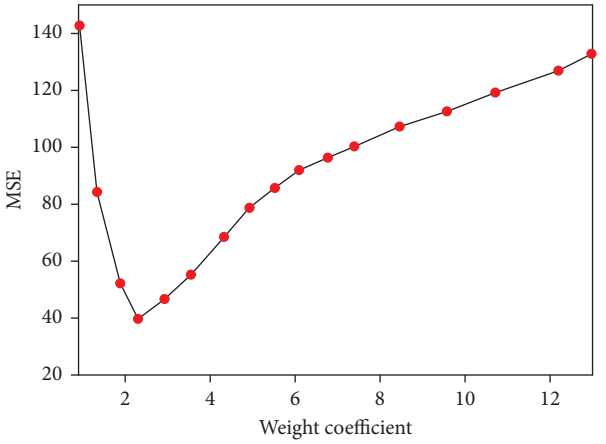

Figure 8: Performance of the model under different parameter weights.

can be seen that although we mainly focus on crowd counting, the behavior classification results based on crowd trajectories still have high accuracy, which fully demonstrates the effectiveness of the model.

4.6. Loss Function Weight Selection Experiment. In order to explain the basis of the weights in the loss function, this experiment also analyzes the performance of the model under different parameter weights [29]. From the perspective of simplifying the model training process, this experiment firstly compares the size of the gradient of each loss function and sets the weight α to 2.1, and then 6 representative values are selected as the testing value of the weight β, which is determined by comparison experiments for its final value. The experimental results are shown in Figure 8. As the value of β increases, the MAE index of the model continues to decrease. In addition, the weight of L_1 and L_E in the loss function was equal when $\beta = 1$, and the model obtained the lowest MAE index. When the value of β continues to increase, the MAE indicator increases rapidly. In other words, the weight gap between L_1 and L_E increases gradually, and the performance of the model begins to

decline. Therefore, when the value of β is equal to 1, the model performance is the best.

5. Conclusion

In order to solve the problem that the local correlation of image is ignored and the ability of multiscale feature extraction is limited to crowd counting, a crowd counting model based on multiscale network is proposed in this paper. The multibranch generation network and the regional discrimination network are combined to form an embedded GAN module and then connected to the multiscale module based on pyramid pooling structure. Three loss functions are used to train the whole model, so that the model can improve the local correlation of the predicted image and the multiscale feature extraction ability, so as to improve the final counting accuracy and robustness of the model. In this paper, a large number of qualitative and quantitative experiments on the public dataset of 3 crowd counts have proved the effectiveness of the model, which is suitable for engineering applications in the field of security surveillance. In order to achieve real-time detection results, in the next step, we will optimize the model in parallel and transplant embedded devices to improve the level of intelligent detection of security monitoring applications.

Data Availability

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

Conflicts of Interest

The authors declare that they have no conflicts of interest regarding the publication of this paper.

References

[1] G. Xiong, J. Cheng, X. Wu, Y.-L. Chen, Y. Ou, and Y. Xu, "An energy model approach to people counting for abnormal crowd behavior detection," *Neurocomputing*, vol. 83, pp. 121–135, 2012.

[2] X. Tian, H. H. Li, and H. Deng, "Object tracking algorithm based on improved context model in combination with detection mechanism for suspected objects," *Multimedia Tools and Applications*, vol. 78, no. 12, pp. 16907–16922, 2019.

[3] Z. Xu, S. Zhu, and D. Jin, "Abnormal behavior detection in crowd scenes," in *Proceedings of the 30th China Control and Decision Conference*, pp. 1523–1531, Shenyang, China, December 2020.

[4] J. Shao, C. C. Loy, K. Kang et al., "Slicing convolutional neural network for crowd video understanding," in *Proceedings of the Computer Vision & Pattern Recognition*, pp. 25–38, IEEE, Las Vegas, NV, USA, June 2016.

[5] M. Sun, "Crowd abnormal behavior detection based on label distribution learning," in *Proceedings of the International Conference on Intelligent Computation Technology & Automation*, pp. 1230–1245, Xi'an, China, May 2016.

[6] X. Tian, H. Li, and H. Deng, "Object tracking algorithm based on improved siamese convolutional networks combined with deep contour extraction and object detection under airborne platform," *Journal of Imaging and Technology*, vol. 16, pp. 2525–2538, 2020.

[7] J. Cai, X. Zhang, and S. Xie, "Video crowd detection and abnormal behavior model detection based on machine learning method," *Neural Computing and Applications*, vol. 31, no. 8, pp. 759–763, 2019.

[8] F. Zhao and J. Li, "Pedestrian motion tracking and crowd abnormal behavior detection based on intelligent video surveillance," *Journal of Networks*, vol. 9, pp. 049–068, 2014.

[9] J. J. Lee, G. J. Kim, and M. H. Kim, "Trajectory extraction for abnormal behavior detection in public area," in *Proceedings of the International Conference & Expo on Emerging Technologies for A Smarter World*, pp. 2598–2608, Incheon, South Korea, November 2013.

[10] D. G. Lee, I. Heung, and S. W. Lee, "Computer science applications information systems signal processing," in *Proceedings of the 11th IEEE international Conference On Advanced Video and Signal-Based Surveillance, AVSS 2014*, pp. 259–268, Seoul, South Korea, August 2014.

[11] F. Xu, Y. Rao, and Q. Wang, "An unsupervised abnormal crowd behavior detection algorithm," in *Proceedings of the International Conference on Security, Pattern Analysis, and Cybernetics 0*, Shenzhen, China, December 2017.

[12] J. Jiang, Y. Tao, W. Zhao, and X. Tang, "Abnormal crowd motion detection using double sparse representations," *Lecture Notes in Computer Science,Intelligence Science and Big Data Engineering. Image and Video Data Engineering*, vol. 14, no. 8, pp. 528–539, 2015.

[13] J. M. Grant and P. J. Flynn, "Crowd scene understanding from video: a survey," *Acm Transactions on Multimedia Computing Communications & Applications*, vol. 13, pp. 19–35, 2017.

[14] L. I. Fei, "Crowd abnormal behavior detection based on motion similar entropy," *Telecommunications*, vol. 24, no. 8, pp. 258–269, 2017.

[15] H. Madhura, V. Vyas, and Y. M. Vaidya, "Abnormal crowd behavior detection based on combined approach of energy model and threshold," *International Conference on Pattern Recognition and Machine Intelligence*, vol. 10597, pp. 137–147, 2017.

[16] Q. Wang, Q. Ma, C. H. Luo et al., "Hybrid histogram of oriented optical flow for abnormal behavior detection in crowd scenes," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 30, no. 2, pp. 14–25, 2016.

[17] T. Senst, V. Eiselein, A. Badii, M. Einig, I. Keller, and T. Sikora, "A decentralized privacy-sensitive video surveillance framework," *International Conference on Digital Signal Processing IEEE*, vol. 4pp. 58–698, , Fira, Greece, July 2013.

[18] H. Fradi and J.-L. Dugelay, "Spatial and temporal variations of feature tracks for crowd behavior analysis," *Journal on Multimodal User Interfaces*, vol. 10, no. 4, pp. 307–317, 2016.

[19] H. Rabiee, H. Mousavi, M. Nabi, and M. Ravanbakhsh, "Detection and localization of crowd behavior using a novel tracklet-based model," *International Journal of Machine Learning and Cybernetics*, vol. 9, no. 12, pp. 1999–2010, 2018.

[20] Y. Rui, L. Bing, H. Ye-Lin et al., "A method for abnormal behavior recognition based on deep learning," *Journal of Wuyi University (Natural Science Edition)*, vol. 12, no. 11, pp. 25–31, 2018.

[21] B. Huang, "Abnormal behavior detection method for aerial monitoring platform," *Ence Technology and Engineering*, vol. 10, pp. 307–317, 2018.

[22] W. Zhaojing, "An abnormal behavior classification detection algorithm based on crowd density," *Video Engineering*, vol. 24, no. 5, pp. 63–69, 2018.

[23] D. Jin, H. Wang, G. Feng, B. Li, and M. Jia, "Sparse representation and weighted clustering based abnormal behavior detection," in *Proceedings of the 2018 24th International Conference on Pattern Recognition (ICPR)*, pp. 3697–3699, Beijing, China, August 2018.

[24] N. Marir, "Distributed abnormal behavior detection approach based on deep belief network and ensemble SVM using spark," *IEEE Access*, vol. 6, 2018.

[25] S. Wang and J. Huo, "Bayesian framework with non-local and low-rank constraint for image reconstruction," *Journal of Physics Conference Series*, vol. 12, no. 5, pp. 787–796, 2017.

[26] H. Tsushita and T. T. Zin, "A study on detection of abnormal behavior by a surveillance camera image," in *Proceedings of the International Conference on Big Data Analysis and Deep Learning Applications*, pp. 688–702, Saint Petersburg, Russia, June 2018.

[27] Z. Fang, F. Fei, Y. Fang et al., "Abnormal event detection in crowded scenes based on deep learning," *Multimedia Tools and Applications*, vol. 75, no. 22, pp. 14617–14639, 2016.

[28] D. Xu, "Learning deep representations of appearance and motion for anomalous event detection," 2015, http://arxiv.org/abs/1510.01553.

[29] M. Ravanbakhsh, "Abnormal event detection in videos using generative adversarial nets," in *Proceedings of the 2017 IEEE International Conference on Image Processing (ICIP)*, pp. 1577–1581, Beijing, China, September 2017.