Hindawi
International Journal of Intelligent Systems
Volume 2023, Article ID 6677622, 13 pages
https://doi.org/10.1155/2023/6677622

Research Article

A Fusion-Based Dense Crowd Counting Method for Multi-Imaging Systems

Jin Zhang, Luqin Ye, Jiajia Wu, Dan Sun, and Cheng Wu

School of Rail Transportation, Soochow University, 8 Jixue Road, Suzhou 215011, China

Correspondence should be addressed to Cheng Wu; cwu@suda.edu.cn

Received 21 July 2023; Revised 12 September 2023; Accepted 6 October 2023; Published 18 October 2023

Academic Editor: Surya Prakash

Dense crowd counting has become an essential technology for urban security management. The traditional crowd counting methods mainly apply to the scene with a single view and obvious features but cannot solve the problem with a large area and fuzzy crowd features. Therefore, this paper proposes a crowd counting method based on high and low view information fusion (HLIF) for large and complex scenes. First, a neural network based on an attention mechanism (AMNet) is established to obtain a global density map from a high view and crowd counts from a low view. Then, the temporal correlation and spatial complementarity between cameras are used to calibrate the overlap areas of the two images. Finally, the total number of people is calculated by combining the low-view crowd counts and the high-view density map. Compared to single-view crowd counting methods, HLIF is experimentally more accurate and has been successfully applied in practice.

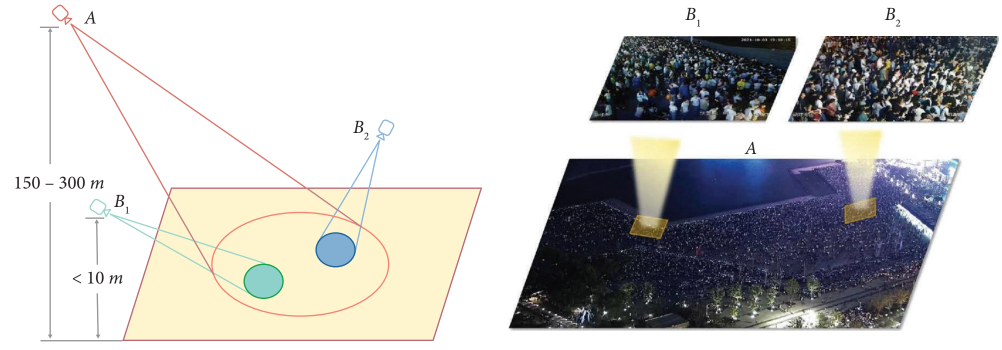
1. Introduction

Crowd counting and density estimation have essential applications in urban security governance, such as preventing dangerous events like crowd stampede and illegal assembly. With the continuous development of deep learning, the performance of single-camera dense crowd counting methods has gradually improved, achieving good results on existing single-view datasets [1–3]. However, for some scenes that are large and wide and require more robust perception as shown in Figure 1, such as parks, squares, stadiums, and large train stations, a single camera cannot cover the entire scene while improving the clarity of feature information. It is also challenging to address the impact of objects such as buildings, landscapes, and stairs in the location on the count. Therefore, multiple cameras with overlapping fields of view can provide complementary information about different features, effectively solving problems such as occlusion and perspective.

Traditional multi-view counting methods are mainly based on detection [4–7], regression [8, 9], and 3D cylinder [10]. The detection-based methods obtain the detection results of each view by a detector and then integrate the information from multiple perspectives. The regression-based methods extract the features of each view, such as size, shape, and key points, and construct a regression function between the feature vector and the crowd size. 3D cylinder-based methods determine the position of a person in a 3D scene by minimizing the distance between the 3D projected position of the person and the camera view. However, these methods depend on manual feature extraction and foreground extraction techniques, which are ineffective for high-density crowd counting. Based on the excellent performance of deep neural networks in single-view dense crowd counting, Zhang et al. [11] constructed the deep neural networks for multiview counting method, and more accurate results were achieved. The existing algorithms have made good progress for low- and medium-density crowd counting problems. However, further research is required for high-density or ultra-high-density scenarios.

Based on the above problems, we propose a crowd counting method based on high and low view information fusion (HLIF). First, under the premise that the high-altitude view should include the low-altitude view as shown in Figure 2(a), an attention mechanism-based low-altitude

Figure 1: Dense crowd in a large view scene.



(a)

(b)

Figure 2: (a) A framework of our multi-camera system: A is the high-view camera, and B is the low-view camera. There is a coincidence area between the two visual fields. (b) The high- and low-view images of CROWD_SZ dataset.

crowd counting model, AMNet, is established for accurate crowd counting in low-altitude view images. Second, the corresponding feature points between the high-altitude view and low-altitude view images are elected to achieve the fusion of the discrete density map of the population of high-altitude view images and low-altitude view images. Finally, the global number of people is derived from the low and high view crowd density map information in the overlap area.

The main existing multi-view datasets are PETS2009 [12], DukeMTMC [13], and City Street [14] as shown in Figure 3, but these datasets have some shortcomings. PETS2009 focuses on a sidewalk scene, not a large view scene. DukeMTMC focuses on multi-view tracking and human detection, so strictly speaking it is not dense crowd. City Street improves on the resolution and crowd size from the first two datasets, but the crowd density level is still low. To better explore the research idea of multi-view approach, we established the dataset CROWD_SZ for the landmark and large musical fountain square, which contains rich crowd density levels, multiple camera views with different spatial perceptions, etc., as shown in Figure 2(b).

In summary, the work and contributions of this paper are as follows.

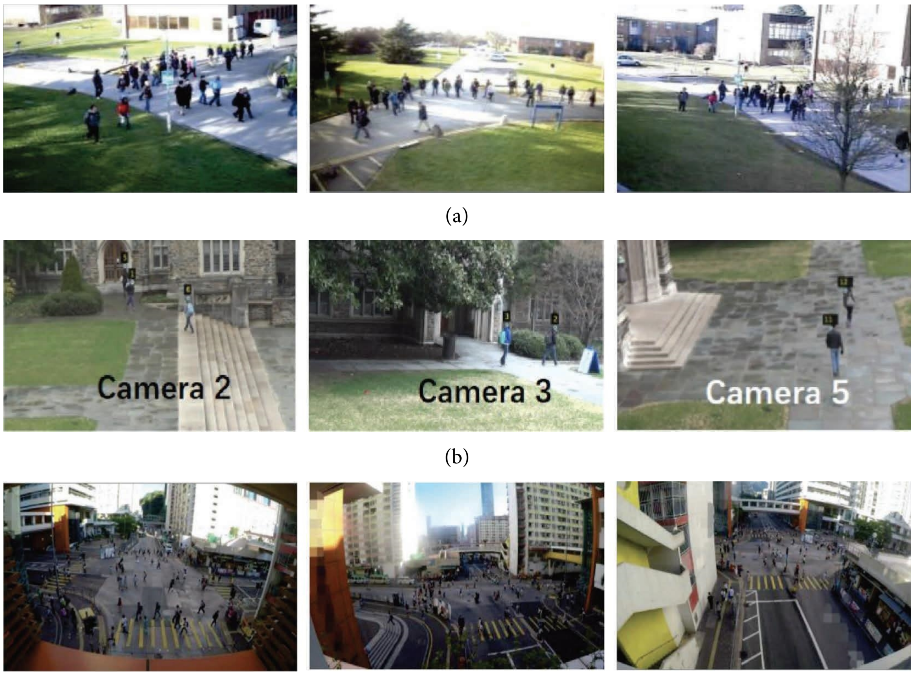
(i) For large-view scenes, a high- and low-altitude view image fusion mechanism is established to effectively utilize the spatial complementarity between high- and low-altitude image information to realize the alignment fusion of crowd density map of high-altitude view and low-altitude view images, which is experimentally verified to be highly accurate.

(ii) We construct a multiview dense crowd counting dataset from real-world scenes, and it provides video frames of different scenes with good statistical dispersion, fully shows the actual situation in multiview scenes.

2. Related Work

Since single-view images are primarily close-up images with more obvious crowd characteristics, existing methods have improved on scale transformation, background noise, and other issues. However, they can only count the number of people in a single region and cannot reflect the global crowd dynamics changes in large scenes. Multi-view counting can compensate for the lack of information provided by single-view images and effectively solve the problems of occlusion and perspective in dense crowd scenes. Also, the reliability of single-view counting is beneficial to improve the perception of large scenes and the accuracy of global crowd calculation, which is the basis of multi-viewpoint research.

2.1. Single-View Perception Methods. To address the problems of scale variation and uneven distribution in crowd counting, Zhang et al. [1] constructed a multi-column convolutional neural network using convolutional kernels of different sizes to extract head scale information of various sizes, which is a pioneering work in crowd counting algorithms. PACNN [14] is a four-column perspective-aware convolutional neural network that integrates perspective information into density regression to provide additional knowledge of the scale variation of the person in the image, and the output density map is combined to adapt to scale variation. DeepCount [15] introduces multi-gradient fusion, where the backbone network receives gradients from



(a)

(b)

(c)

Figure 3: Existing multi-view datasets. (a) PETS2009 dataset. (b) DukeMTMC dataset. (c) City Street dataset.

multiple branches to learn density information. MMNet [16] is an end-to-end scale-aware network that handles the problem of head scale variation by integrating multi-scale features generated by filters of different sizes. SASNet [17] is a scale-adaptive selection network that automatically learns the internal correspondence between scale and feature level. STNet [18] consists of a scale tree diversity enhancer and multi-level auxiliaries and uses a tree structure for hierarchical parsing of coarse to delicate crowd regions to alleviate the scale variation problem. PFDNet [19] overlays multiple perspective views in the backbone network and introduces fractional expansion convolution to address the scale variation problem.

In a recent study, Gao et al. [20] analyzed crowd counting technologies in the fields of Internet of Things, healthcare systems, and cell counting, showing the trends and challenges of crowd counting technologies in various fields. Guo et al. [21] proposed a lightweight Ghost Attention Pyramid Network (GAPN) that combines the advantages of Ghost Convolution and Ghost Batch Normalization, reduces the network parameters and computation, and uses Pyramid Attention Mechanism to capture multi-scale crowd features. Scale Region Recognition Network (SRRNet) was proposed in [22]. The scale variation problem is solved by encoding multiple scale feature representations through Scale-Level Awareness module. The effect of background interference is suppressed by Object Region Recognition module. Zhai et al. [23] proposed Scale-Context Perceptive Network (SCPNet), which consists of Scale Perceptive (SP) module and Context Perceptive (CP) module. The scale variation problem is solved by a local-global branching structure, and the CP module uses a channel-space self-attention mechanism to suppress the effect of background interference. Zhai et al. [24] proposed Attentive Hierarchy ConvNet (AHNet) in which discriminative feature extractor is used to extract multi-level feature representation and hierarchical feature aggregator is used to mine semantic features in a coarse-to-fine manner. Zhai et al. [25] proposed Feature Pyramid Attention Network (FPANet), which uses a lightweight structure to extract features at multiple scales and uses an attention mechanism to focus on crowd regions and suppress background interference. Dense Attention Fusion Network (DAFNet) was proposed in [26]. DAFNet employs a partitioning strategy and designs two key modules: the Iterative Attention Fusion (IAF) module and the Dense Spatial Pyramid (DSP) module. The IAF module utilizes multi-scale channel attention units to mitigate the effect of background clutter, and the DSP module utilizes hierarchical information from different receptive fields to overcome the problem of object scale variation. Guo et al. [27] proposed a Triple Attention and Scale-Aware Network (TASNet) for object counting in remotely sensed images, where the feature pyramid module uses a lightweight structure to extract multi-scale features, the triple-weighted map attention module uses a three-dimensional attention manipulation to distinguish between the object region and the background region, and the pyramid feature aggregation module uses an adaptive weight fusion to generate the final density map. Guo et al. [28] proposed a Spatial-Frequency Attention Network (SFANet), where the spatial attention module is used to emphasize features at different locations in the spatial domain and adaptively selects regions containing individuals, and the multi-spectral channel attention module is used to obtain a more complete representation of each channel with frequency components in the frequency domain. Inspired by biology, Zhai et al. [29] proposed Grouped Segmentation Attention Network (GSANet), which reduces the computational cost by dividing the input feature map into multiple subgroups and processing the subfeatures of each subgroup in parallel. At the same time, it combines the information of spatial and channel dimensions to mitigate the estimation error of the background region. Finally, it employs a learning-based cross-group strategy to aggregate and facilitate the fusion of feature maps with different channel dimensions. Zhai et al. [30] proposed a Dual Attention Perception Network for robust crowd counting in dense crowd scenes with scale variations. The network consists of a Spatial Attention (SA) module and a Channel Attention (CA) module. The SA module focuses on spatial dependencies throughout the feature map to accurately localize the head. The CA module attempts to process the relationships between the channel maps and highlights discriminative information in specific channels. The interaction between the two modules provides synergy and helps in learning discriminative features with attention to the head region.

Although the multi-column structure can effectively handle the scale transformation problem, such models usually depend on the number of network branches. The model has many parameters, is difficult to train, and has poor real-time counting capability. The density map generated based on the attention mechanism has specific errors. It cannot be directly used for density estimation, and it is necessary to consider a way to perceive density in a way that minimizes errors.

Based on this, our low-view crowd counting method uses a single-column structure to ensure the accuracy of counting while simplifying the computational complexity of the network.

2.2. Multi-View Perception Methods. Due to the limitations of traditional multi-view counting methods, the powerful feature extraction capability of deep neural networks, and their success in single-view counting, more and more neural network-based multi-view counting methods are proposed. The multiview multiscale (MVMS) [11] model is the first DNN-based multi-view counting method. It uses the camera geometry, the feature maps from all cameras are projected onto the ground-plane in the 3D world so that the same person's features are approximately aligned across multiple views. The aligned single-view feature maps are fused together and used to predict the scene-level groundplane density map. Zhang and Chan [31] switched to a 3D density map and 3D projection to improve the counting performance. Zheng et al. [32] further enhanced the performance of the fusion model later in the MVMS by modeling the correlation between each team of views. Since the projection of single-view onto the ground plane for fusion requires camera calibration, it limits the method's applicability to scenes where camera calibration is impossible. Zhang et al. [33] proposed a cross-view cross-scene multi-view counting model (CSCV) that incorporates camera selection and noise into the training and can output density maps in different scenes with arbitrary camera layouts. In [34], a multi-view counting model without calibration (CF-MVCC) was proposed to obtain the whole scene person by weighting the confidence score of camera view content and distance information. To solve the view scale inconsistency problem in multi-view counting methods, Liu et al. [35] proposed a multi-view crowd counting model (SASNet) based on scale aggregation and spatially aware networks, in which a multi-branch adaptive scale aggregation module selects the appropriate scale for each pixel in each view based on the extracted features, which can ensure the scale consistency across views.

A key issue in multi-view counting is how to fuse information from multiple cameras. Existing methods primarily project features from the original image coordinates (2D) to the world coordinate system (3D). Then, the projected features of multiple views are fused to generate a scene-level density map. Therefore, it is necessary to obtain the internal and external parameters of the camera and the z-coordinate (height) in the world coordinate system, which has some limitations for the application of the methods in natural scenes and cannot be used across scenes. In addition, the validation datasets used by these methods only partially satisfy the requirements of large scenes. The crowd density needs to be higher to fully illustrate the effectiveness of dense crowd counting methods in large scenes.

Based on the above problems, we propose a new crowd counting method based on the fusion of high- and low-altitude information and validate it using a new large-scene dense crowd dataset.

3. Method

Aiming at the problems of incomplete low-view coverage and unclear high-view texture characteristics in dense crowd estimation, global density information is corrected using the number of low-view angles. As shown in Figure 4, high and low view information fusion (HLIF) consists of two modules, local information processing and information fusion. By inputting a high-view image into the network, it can obtain multiple density levels of a discrete density map which directly reflects the density similarity of each region. At the same time, the overlap area with the low-view image is calculated to obtain its crowd count. The information fusion module combines the person number of low-view and the pixel count of high-view density maps in the overlap area, using the similarity of density distribution to compensate for blind areas, establishes the proportion relation, and deduces the accurate crowd count value.

3.1. Low-View Crowd Counting

3.1.1. AMNet Architecture. In this method, we select the first 13 layers of VGG-16 [36] as the front-end network of the structure, as shown in Figure 5. VGG-16 has advantages of simple structure, strong feature extraction ability, and strong transfer learning ability. It has performed well in many fields, such as image recognition and target detection. Take a continuous $3*3$ convolutional kernel to ensure the reception field and reduce the parameters. Select the pooling size of $2*2$ to extract the main features, reduce the size of the

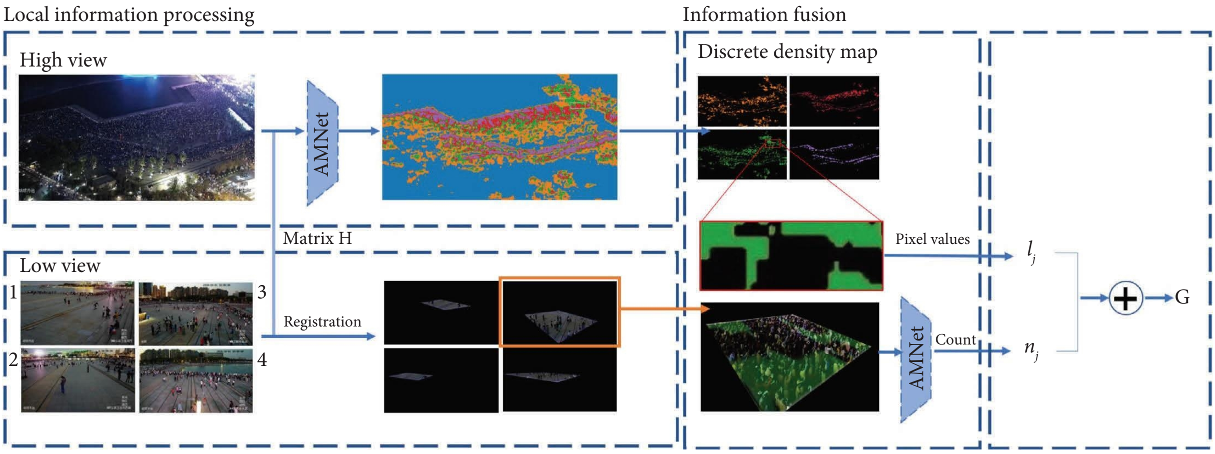


Figure 4: Flowchart of fusion algorithm.

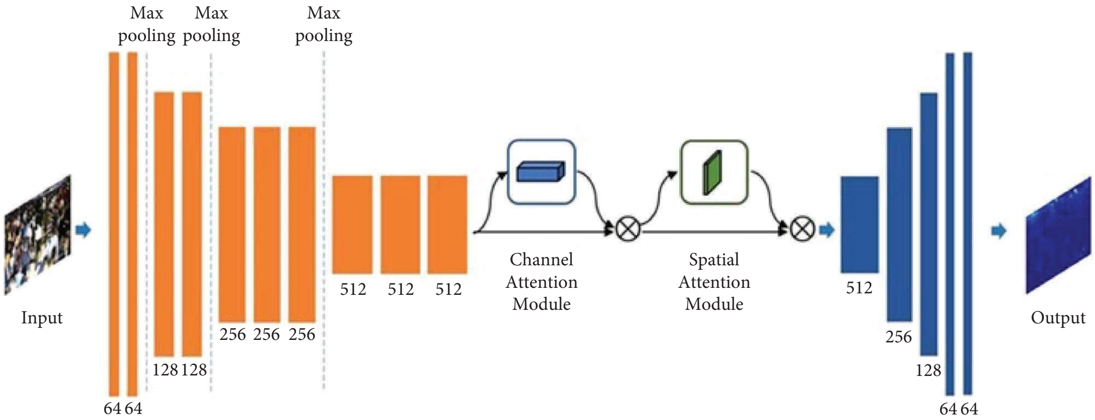


Figure 5: AMNet network structure.

output feature graph, and simplify the computational complexity of the network. The back-end of the network introduces the Convolutional Block Attention Module (CBAM) [37]. The spatial attention module takes the feature map output from the front-end network as input and assigns weights to the pixels for different locations, and the channel attention module realizes the assignment of weights to the foreground and background information in the feature channel, so that the model pays more attention to the head location information. The attention module generates the corresponding dimensional weights for the original feature maps in channel and space, respectively, then multiplies them with the original feature maps of the input, respectively, which makes the network pay more attention to the more important feature information, and finally outputs the density map of the crowd distribution using successive convolutional layers.

3.1.2. Ground Truth Generation. Before training, we first generate the ground truth to calculate the real number of people in the image and generate the density map by the normalized Gaussian Blur head annotation convoluting Gaussian kernel. The specific process is as follows. If there is a head at a particular position x_i, it can be represented as $\delta(x - x_i)$. If there are N heads in a picture, the number of people in the picture can be expressed as follows:

$$H(x) = \sum_{i=1}^{N} \delta(x - x_i). \tag{1}$$

According to the known head position, the size of the head is often associated with adjacent K personal head of center distance, so we can get the distribution by using the geometric adaptive Gaussian kernel, and density map $F(x)$ can be expressed as

$$F(x) = H(x) * G_{\sigma_i}(x) \quad \text{with } \sigma_i = \beta \overline{d_i}, \tag{2}$$

where G_{σ_i} is a Gaussian kernel, σ_i represents the average adaptive distance of the head within a certain range, β is usually 0.3, and $\overline{d_i}$ is the average distance of k neighbors of the current head.

3.1.3. Data Augmentation. We crop four patches from each image at different locations and gray scale with a quarter size of the original image. At the same time, adaptive kernel is applied to the annotation files, weighted sum of the original image is carried out, and the ground truth of the image is obtained by traversing, which is involved in the subsequent training model.

3.1.4. Training Details. The model's input is the crowd image, and the output is the density map. Stochastic Gradient Descent (SGD) is applied with a fixed learning rate at $1e-6$ during training. We choose the Euclidean distance to measure the difference between the ground truth and the estimated density map. The loss function is given as

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

where N is the size of training batch, $F(X_i;\Theta)$ is the output generated by AMNet with parameters shown as Θ, X_i represents the input image, and F_i^{GT} is the ground truth of the input image X_i.

3.2. Fusion Algorithm

3.2.1. Discrete Density Map Generation. Firstly, the image was input into the AMNet to obtain the density map. Secondly, using the colormap in Matplotlib, it was discretely divided into four density levels: low density, medium density, high density, and ultra-high density.

We determined thresholds for different density classes based on the statistics in CROWD_SZ. Specifically, we took the number of people at each pixel as the density value and then calculated the mean and standard deviation of the density values at all pixels of the images. We used a mean plus or minus standard deviation as the dividing line between medium and high density, and two mean plus or minus standard deviations as the dividing line between low and medium density, and high and ultra-high density. Finally, as shown in Figure 6, the density map displays different colors according to the change in density level, namely, orange, green, red, and purple. Color can more intuitively reflect the correlation between regions with the same density level on the same image. In the image from the high view, there are some regions that overlap with the image from the low view. As a result, the density distribution in this region can be observed more clearly. However, there are no low-view cameras with overlapping fields of view in some high-view regions. Therefore, analogous reasoning can be applied to other high-view regions with the same density.

3.2.2. High- and Low-View Image Region Registration. The images to be processed are images from high view and low view. The choice of perspective should follow the following principles:

(i) The high-view range must include the area covered by the low-view range.

(ii) Low-view camera equipment should be able to capture a set of richer human features.

(iii) The selection of low-view areas should include different gathering forms at this time.

(iv) High- and low-view images should adhere to time consistency.

There is an overlap area between high- and low-view images that may reflect the distribution of crowds in high-view images and the specific number of people in low-view images. The corresponding relationship is shown in Figure 7.

To achieve more accurate regional registration, as shown in Figure 8, this paper finds the corresponding feature points in the high-view images and registers them according to the homography theory. The position of feature points in the low-view image is represented by (x, y), the position coordinates of multiple feature points are represented by (x_r, y_r), and the corresponding feature points in the high-view region are represented by (X_r, Y_r). The transformation relation of the two images is as follows:

$$\begin{gathered}(X_r, Y_r) = H * (x_r, y_r),\\ H = \begin{bmatrix} h_{11} & h_{12} & h_{13}\\ h_{21} & h_{22} & h_{22}\\ h_{31} & h_{32} & h_{33}\end{bmatrix},\end{gathered} \tag{4}$$

where H is the transformation matrix.

To verify the registration accuracy, ground marker lines were selected as reference lines, as shown in Figure 9. The red line in Figure 9(a) is the ground marker line, and the verification result of the fifth column is obtained by a transformation; the two are highly overlapping. After verifying the matrix accuracy, the rectangular region to be studied is selected from the high-view image and converted to obtain the corresponding region of the overlapping region in the low-view image. In Figure 9(a), the red border is the selected overlap region, and the blue quadrangle in Figure 9(c) is obtained after the homography matrix transformation.

3.2.3. Information Fusion. The information fusion process is as follows:

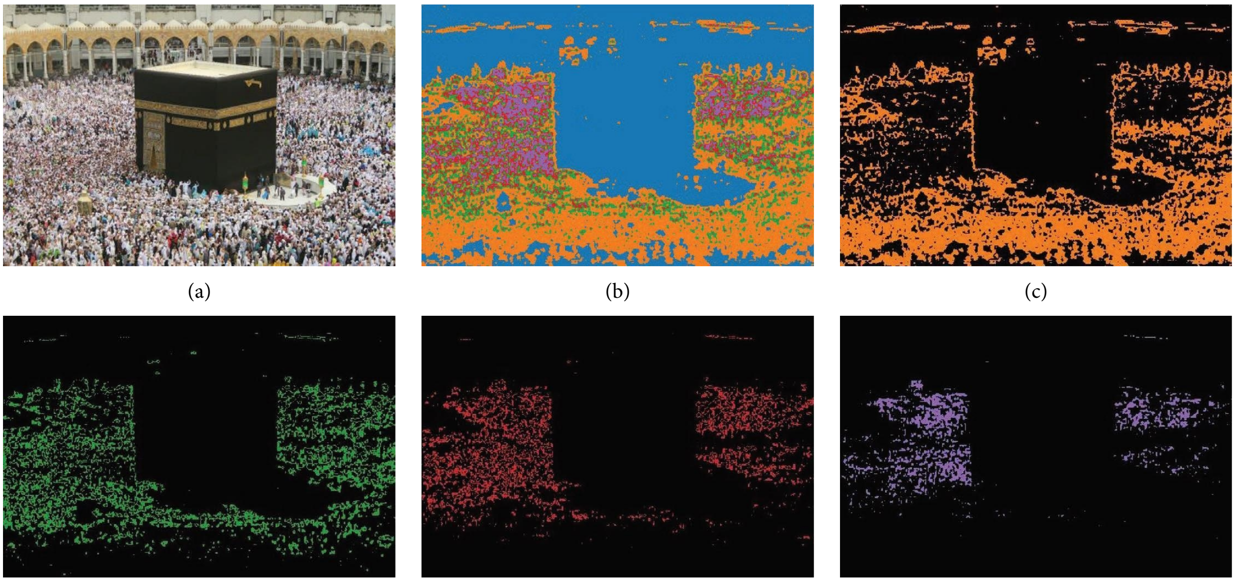
(i) Image acquisition: at the same time, select 1 high-view equipment called A and m low-view equipment called B_1, B_2, $B_3 \ldots B_m$.

(ii) Region registration: select m rectangular areas (in overlap areas) on device A as L_1, L_2, $L_3 \ldots L_m$ and use the matrix H to obtain the overlap area on device B_i corresponding to the rectangular region.

(iii) Information processing: Input the high-view image to get the discrete density map and calculate the pixel number b_t of each density level, where t represents the density level. The number of pixels at each density level in the region L_m is denoted as n_j. Through the registration matrix, the discrete density map is projected onto the corresponding low-view image and the AMNet is used to calculate the specific number of people in the covered area l_j.

(iv) At last, establish a proportional relationship to calculate the global number G.

$$G = \sum_{t=1}^{k}\left(\frac{1}{M}\sum_{j=1}^{M}\frac{l_j}{n_j}\right) * b_t, \quad j = 1, 2, 3..m, t = 1, 2, 3..m, \tag{5}$$



(a)

(b)

(c)

(d)

(e)

(f)

Figure 6: Discrete density diagram. (a) The original image. (b) Discrete density map. (c) Low density. (d) Medium density. (e) High density. (f) Ultra-high density.

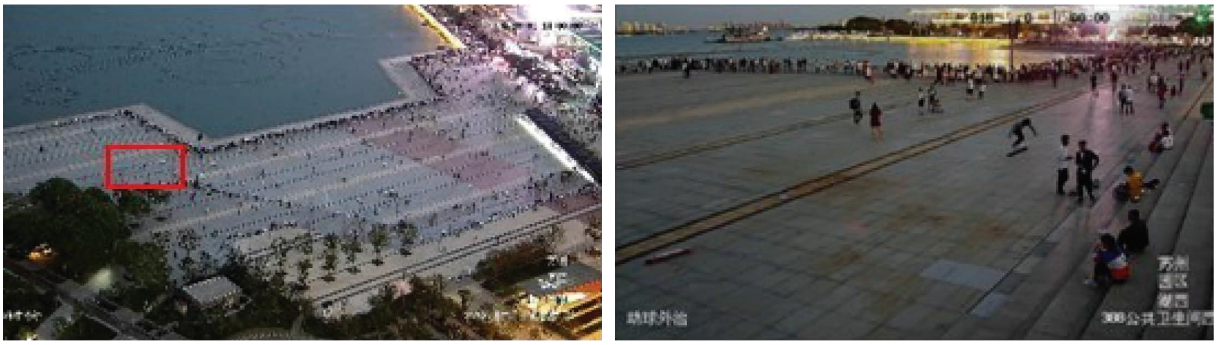


Figure 7: Schematic diagram of overlap area of high- and low-view images.

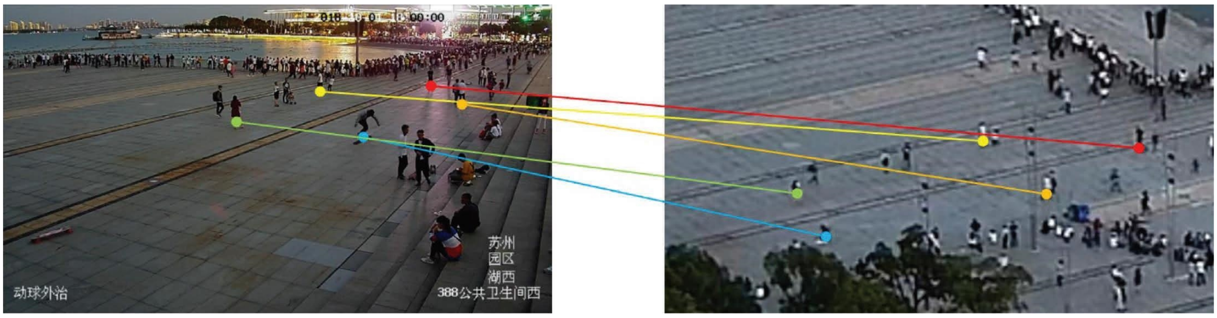


Figure 8: Feature point matching.

where G is the global number of the high-view image, n_j represents the pixels of same areas in the discrete density map, l_j is the number of people in the area covered by the discrete density map, b_t is the number of all pixels contained in each density level, m is the number of overlap areas, and t is the density level.

4. Experiments

4.1. Evaluation Metrics. In this paper, we use the Mean Absolute Error (MAE) and the Mean Squared Error (MSE) to evaluate the performance of the model on the test set, which are defined as follows:

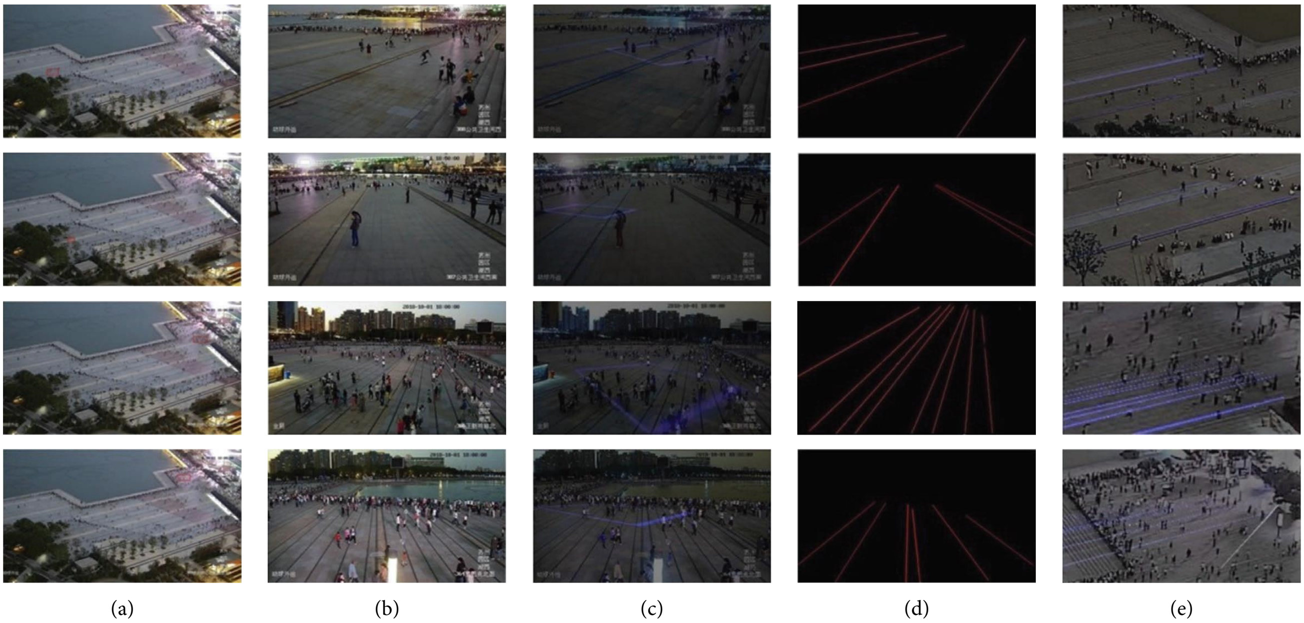


Figure 9: (a) The overlap region in the high-view image. (b) Corresponding low-view images. (c) The result of fusion. (d) Ground mark line. (e) The result of fusion accuracy validation.

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}\left|z_i - \widehat{z}_i\right|,$$
$$\mathrm{MSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left|z_i - \widehat{z}_i\right|^2}, \tag{6}$$

where N is the number of test images and Z_i and $\widehat{Z}_i$ represent the ground truth and estimated values of the *ith* image.

Table 1: Estimation errors on ShanghaiTech dataset.

Method	Part_A		Part_B	
	MAE	MSE	MAE	MSE
PCC Net [38]	124.0	73.5	19.0	11.0
IG-CNN [39]	118.2	72.5	21.1	13.6
CP-CNN [40]	106.4	73.6	30.1	20.1
RReg (MCNN) [41]	114.3	72.6	23.1	15.5
ACSCP [42]	**102.7**	75.7	27.4	17.2
AMNet	113.0	**71.7**	**17.0**	**11.0**

The bold value shows the best result in every column.

4.2. Evaluation and Comparison

4.2.1. AMNet Performance Evaluation. Our model was trained on two benchmark datasets, ShanghaiTech and UCF_CC_50. Experimental results show that this structure is superior to most existing crowd counting networks.

As shown in Table 1, the ShanghaiTech dataset contains 1,198 images, with a total count of 330,165 people, divided into two parts, sparse and crowded scenes. The optimal MAE and sub_optimal MSE were obtained, and the MAE was 1.1% lower than that of IG-CNN. In Part_B, AMNet achieves the optimal MAE and MSE, the MAE is reduced by 23.6% compared to IG-CNN, and the MSE is reduced by 11.8% compared to PCC Net.

UCF_CC_50 specifically collects high-density crowd scenes, which compensates for the shortcomings of the existing dataset. The number of people in a single image ranges from 94 to 4,543, with an average of 1,280. However, this dataset only contains only 50 images, so we used 5x cross-validation. As shown in Table 2, the MAE was 2.8% lower than that of the previous optimal SANet.

4.2.2. Validation of Fusion Algorithm. At present, there is no high view and low view dataset to test the proposed method, so we collect a crowd dataset CROWD_SZ. The dataset was collected from the video surveillance, located in the Jinji Lake Urban Life Square in Suzhou, Jiangsu Province, China. The Jinji Lake Urban Life Square is famous for its large musical fountain and attracts a large number of visitors during the opening of the fountain. Thus, it is a typical large-scale, high-density scene. As shown in Figure 10, one high view and four low views are selected to determine their overlapping regions. Choose October 1, 2018, from 18:00 to 21:00, during which the crowd flow, gathering, and dispersal can be comprehensively observed, including before the fountain opens and after the fountain closes.

First, AMNet was used to obtain the discrete density map of the high-view image, as shown in Figure 10. Color components of different density levels were extracted, and the pixel numbers of color components of different density levels and the pixel numbers of color components of each density level contained in the overlapping area were counted.

Using 19:35:00 as an example, the process of estimating the number of people in the medium density area is as

Table 2: Estimation errors on UCF_CC_50 dataset.

Method	UCF_CC_50	
	MAE	MSE
SANet [43]	258.4	334.9
SAAN [44]	271.6	391.0
PACNN [14]	267.9	357.8
Onoro-Rubio and López-Sastre [45]	465.7	371.8
PGCNet [46]	259.4	**317.6**
CSRNET [47]	266.1	397.5
AMNet	**251.3**	350.7

The bold value shows the best result in every column.

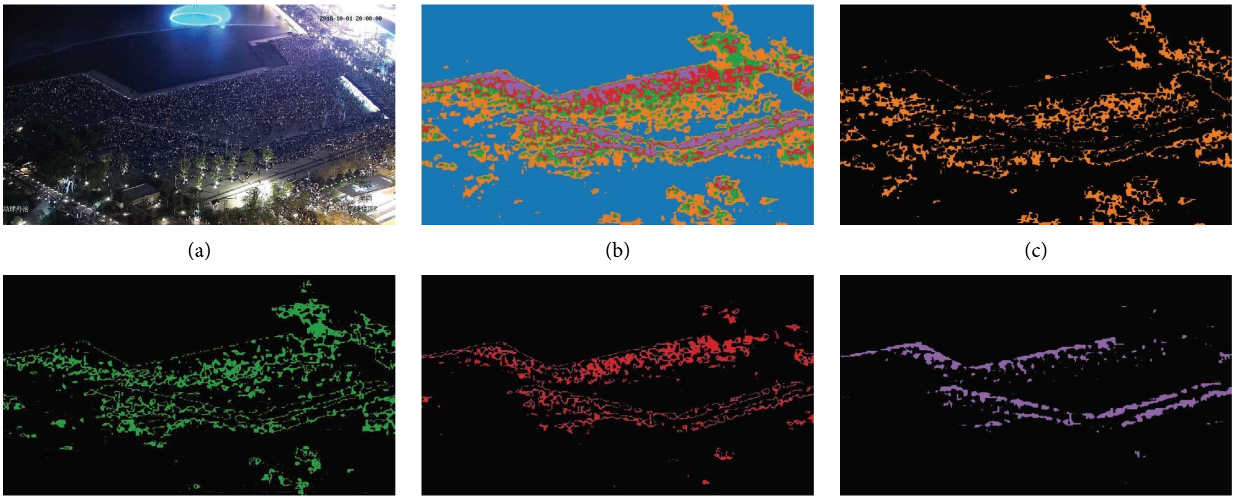

(a)
(b)
(c)
(d)
(e)
(f)

Figure 10: Discrete density map in CROWD_SZ. (a) The original image. (b) Discrete density map. (c) Low density. (d) Medium density. (e) High density. (f) Ultra-high density.

follows: the first column of Figure 11 is the color component diagram of the medium density level, and the pixel value b_2 is 169,124. The second column of Figure 11 lists the two selected overlapping regions, and the pixel values n_j are 2,592 and 1,224, respectively. Through the homography matrix, the perspective is mapped to the corresponding region of the low-view image, namely, the green mask part in the image. Through AMNet, the crowd counts l_j are 42.6 and 18.2, respectively, c is the scale factor, and g is the crowd count in the current density level. Table 3 shows the specific calculation content. For each density level, the global number can be added up.

As shown in Table 4, the global count calculation process at 20:05:00 is taken as an example to illustrate the content of the global crowd calculation. The global number G can be obtained by summing up the inferred crowd of each density level. At 20:05:00, the global crowd count in this scene is 5,981.4331.

We selected the video clips from 19:28:40 to 20:45:00 to well observe the changes in crowd density. During this period, the number of people was calculated every five minutes, as shown in Figure 12, the average error is about 258 people, and the estimation accuracy is up to 93.8%. From 19:28:40, the number of people continues to increase until 19:45:00 to 20:20:00 with relatively stable fluctuations. After 20:20:00, the crowd count continues to reduce. However, since the selected scene is an open space, there are errors in manual labeling, and the ground truth here is approximate, basically consistent with the actual changes.

In addition, we use several popular methods to count the number of people in this dataset, such as CSRNET, MCNN, and AMNet. The performance of the methods is poor in the scene, as shown in Figure 13. Although the crowd distribution can be captured, the calculated results are very different from the ground truth.

Due to the distance of the high-altitude camera equipment, the crowd information is displayed as pixels, and the effective characteristic information of the human body is not obvious, so we need to add the low-altitude information to supplement. In addition, in the CROWD_SZ dataset, the musical fountain performance has strong light and shadow changes, which leads to overexposure or insufficient light in the global image, and more information is lost, which is also compensated by the low-altitude information. For example, in the first three columns of Figure 13, neural network models such as MCNN are affected by the light and estimate the exposed area as a blank area. However, the use of AMNet on low-altitude images can effectively detect scenes with changing black light and compensate for the lost information globally by more accurate low-altitude crowd

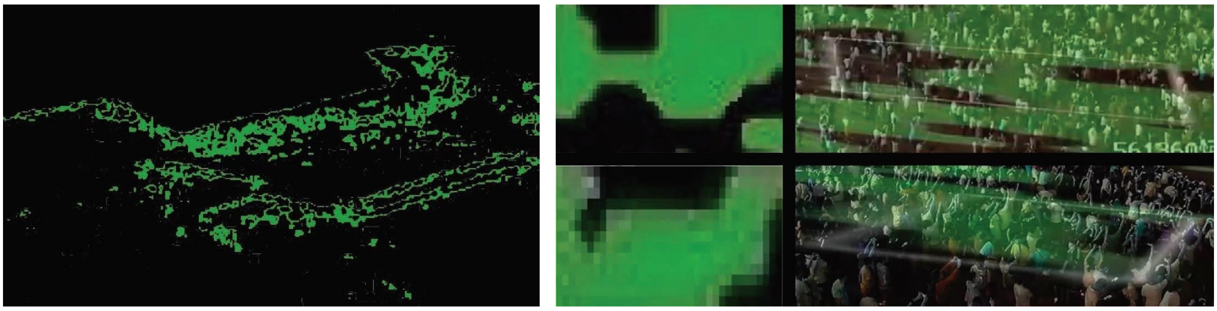


FIGURE 11: The first column shows the two selected regions, the second column shows the discrete density maps generated by AMNet, and the third column shows the corresponding low-view images of these regions.

TABLE 3: 19:35:00 medium density crowd calculation process.

Density levels	b_2	n_j	l_j	c	g
Medium density	169,124	2,592 1,224	42.6 18.2	0.016	**2,705.984**

The bold value shows the crowd count in the medium density level.

TABLE 4: 20:05:00 global crowd calculation process.

Density levels	b_t	n_j	l_j	c	g	G
Low density	224,538	669 974	0.9 1.8	0.0016	359.2608	**5,981.4331**
Medium density	213,261	2,019 910	29.6 17.2	0.0168	3,582.7848	
High density	22,187	2,828 2,780	62.1 41.7	0.0185	410.4595	
Ultra-high density	45,248	261 417	10.9 12.6	0.0360	1,628.9280	

The bold value shows the global crowd count.

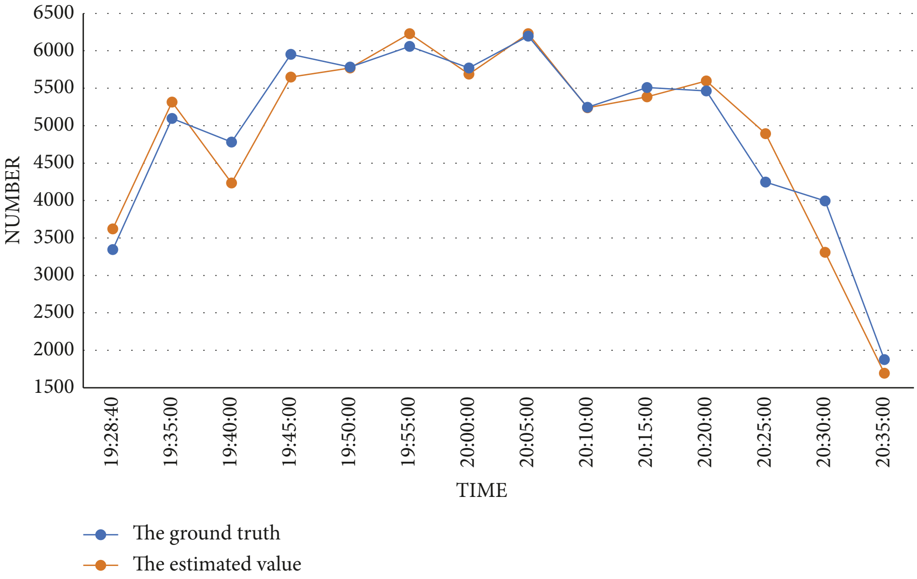


FIGURE 12: Test results of high and low view information fusion algorithm on CROWD_SZ.

counts, resulting in a more accurate global crowd count estimate. Thus, the HLIF structure effectively reduces the impact of light on the global crowd counting. In the last three columns, the light source is closed and the overall illumination of the square is weak, resulting in unclear features of the area farthest from monitoring and the worst

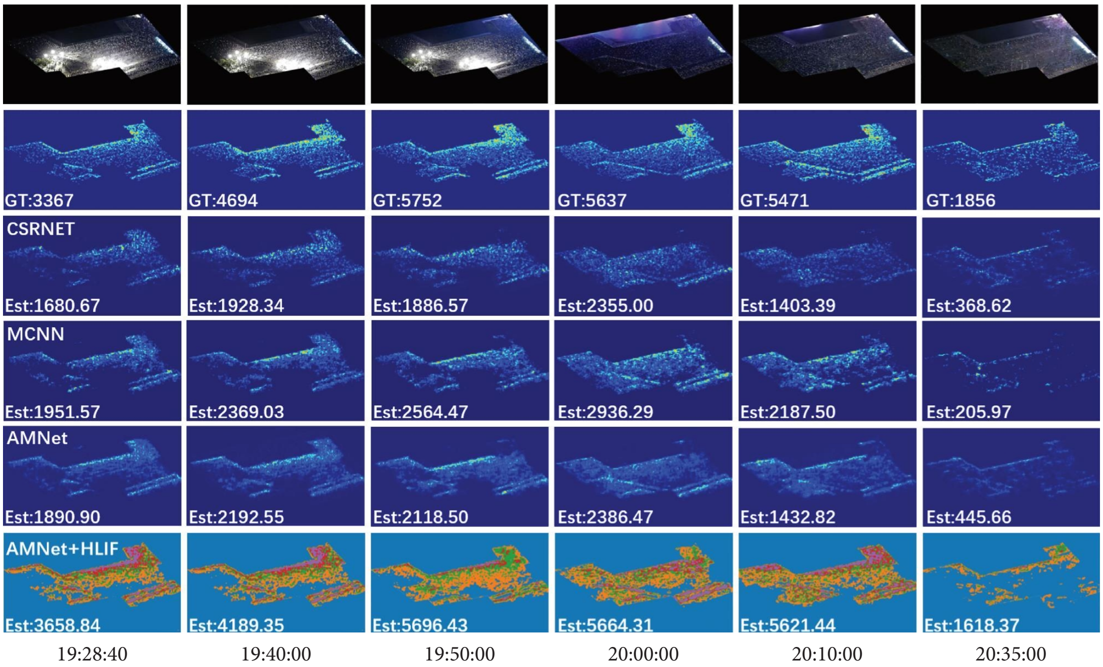


Figure 13: The first row shows the samples of the test set in CROWD_SZ dataset. The second row shows the ground truth for each sample. The third, fourth, and fifth rows are density maps generated by CSRNET, MCNN, and AMNet, respectively. The sixth row is the test result of AMNet + HLIF.

recognition effect of CSRNET. In the fifth column, the estimated value and actual value of MCNN on the image are 2,187.50 and 5,471, respectively, with an error of nearly 3,300 people. After introducing the high-altitude perspective image information fusion module, the estimated result is 5,621.44, with an error reduced to 150 people. Therefore, a more accurate number of people can be obtained by the similarity of density distribution.

5. Conclusion

Crowd counting in public places, especially large-scale high-density spaces, is obviously a very challenging task for public security. The video surveillance system is a good tool to monitor and manage the crowd. The rich video information can provide a more accurate estimate of the number of people, which becomes an important foundation for security management decisions.

In this paper, we propose a crowd counting method based on high and low view information fusion, which effectively solves the problem of dense crowd counting in large-view scenes. The crowd counting network structure based on an attention mechanism (AMNet) is established and helps us to obtain a discrete density map. The overlapping areas are defined by the registration mechanism from high-altitude view and low-altitude view in the scene. The global number of people was calculated by combining the density distribution information of high-altitude view and the number of people in low-altitude view in the overlapping areas.

Compared with the traditional crowd counting algorithm, our proposed high-altitude and low-altitude information fusion algorithm (HLIF) can effectively use the low-altitude counting results to compensate and optimize the global crowd counting, adapt to the drastic changes of light at night, and improve the overall counting accuracy.

Data Availability

The data 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.

Acknowledgments

This study was supported in part by the National Natural Science Foundation of China under grant nos. 62201372 and 62275186, in part by the Scientific Research Project Funding of Suzhou Rail Transit Group Company Ltd. under grant no. SZZG06YJ6000017, and in part by the Scientific Research Project Funding of Suzhou Public Security Bureau Industrial Park Branch under grant no. SZYX2018-Y-C-002.

References

[1] 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 IEEE conference on computer vision and pattern recognition*, pp. 589–597, Las Vegas, NV, USA, June 2016.

[2] H. Idrees, I. Saleemi, C. Seibert, and M. Shah, “Multi-source multi-scale counting in extremely dense crowd images,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2547–2554, Portland, OR, USA, June 2013.

[3] H. Idrees, M. Tayyab, K. Athrey et al., "Composition loss for counting, density map estimation and localization in dense crowds," in *Proceedings of the European conference on computer vision (ECCV)*, pp. 532–546, Munich, Germany, September 2018.

[4] F. Dittrich, L. E. de Oliveira, A. S. Britto, and A. L. Koerich, "People counting in crowded and outdoor scenes using a hybrid multi-camera approach," https://arxiv.org/abs/1704.00326.

[5] J. Li, L. Huang, and C. Liu, "People counting across multiple cameras for intelligent video surveillance," in *Proceedings of the 2012 IEEE Ninth International Conference on Advanced Video and Signal-Based Surveillance*, pp. 178–183, IEEE, Beijing, China, September 2012.

[6] H. Ma, C. Zeng, and C. X. Ling, "A reliable people counting system via multiple cameras," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 3, no. 2, pp. 1–22, 2012.

[7] L. Maddalena, A. Petrosino, and F. Russo, "People counting by learning their appearance in a multi-view camera environment," *Pattern Recognition Letters*, vol. 36, pp. 125–134, 2014.

[8] D. Ryan, S. Denman, C. Fookes, and S. Sridharan, "Scene invariant multi camera crowd counting," *Pattern Recognition Letters*, vol. 44, pp. 98–112, 2014.

[9] N. C. Tang, Y.-Y. Lin, M.-F. Weng, and H.-Y. M. Liao, "Cross-camera knowledge transfer for multiview people counting," *IEEE Transactions on Image Processing*, vol. 24, no. 1, pp. 80–93, 2015.

[10] W. Ge and R. T. Collins, "Crowd detection with a multiview sampler," in *Proceedings of 11th European Conference on Computer Vision*, pp. 324–337, Heraklion, Crete, Greece, September 2010.

[11] Q. Zhang and A. B. Chan, "Wide-area crowd counting via ground-plane density maps and multi-view fusion cnns," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8297–8306, Long Beach, CA, USA, June 2019.

[12] J. Ferryman and A. Shahrokni, "Pets2009: dataset and challenge," in *Proceedings of the 2009 Twelfth IEEE International Workshop on Performance Evaluation of Tracking and Surveillance*, pp. 1–6, IEEE, Snowbird, UT, USA, December 2009.

[13] E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. Tomasi, "Performance measures and a data set for multi-target, multi-camera tracking," in *Proceedings of the Computer Vision–ECCV 2016 Workshops*, pp. 17–35, Springer, Amsterdam, The Netherlands, October 2016.

[14] M. Shi, Z. Yang, C. Xu, and Q. Chen, "Revisiting perspective information for efficient crowd counting," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7279–7288, Long Beach, CA, USA, June 2019.

[15] Z. Chen, J. Cheng, Y. Yuan, D. Liao, Y. Li, and J. Lv, "Deep density-aware count regressor," https://arxiv.org/abs/1908.03314.

[16] L. Dong, H. Zhang, Y. Ji, and Y. Ding, "Crowd counting by using multi-level density-based spatial information: Crowd counting by using multi-level density-based spatial information: A Multi-scale CNN framework multi-scale cnn framework," *Information Sciences*, vol. 528, pp. 79–91, 2020.

[17] Q. Song, C. Wang, Y. Wang et al., "To choose or to fuse? scale selection for crowd counting," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 3, pp. 2576–2583, 2021.

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

[19] Z. Yan, R. Zhang, H. Zhang, Q. Zhang, and W. Zuo, "Crowd counting via perspective-guided fractional-dilation convolution," *IEEE Transactions on Multimedia*, vol. 24, pp. 2633–2647, 2022.

[20] M. Gao, A. Souri, M. Zaker, W. Zhai, X. Guo, and Q. Li, "A comprehensive analysis for crowd counting methodologies and algorithms in internet of things," *Cluster Computing*, pp. 1–15, 2023.

[21] X. Guo, K. Song, M. Gao, W. Zhai, Q. Li, and G. Jeon, "Crowd counting in smart city via lightweight ghost attention pyramid network," *Future Generation Computer Systems*, vol. 147, pp. 328–338, 2023.

[22] X. Guo, M. Gao, W. Zhai, Q. Li, and G. Jeon, "Scale region recognition network for object counting in intelligent transportation system," *IEEE Transactions on Intelligent Transportation Systems*, pp. 1–10, 2023.

[23] W. Zhai, M. Gao, X. Guo, and Q. Li, "Scale-context perceptive network for crowd counting and localization in smart city system," *IEEE Internet of Things Journal*, vol. 1, 2023.

[24] W. Zhai, M. Gao, A. Souri et al., "An attentive hierarchy convnet for crowd counting in smart city," *Cluster Computing*, vol. 26, no. 2, pp. 1099–1111, 2023.

[25] W. Zhai, M. Gao, Q. Li, G. Jeon, and M. Anisetti, "Fpanet: Feature Pyramid Attention Network for Crowd Counting," *Applied Intelligence*, vol. 53, pp. 1–18, 2023.

[26] X. Guo, M. Gao, W. Zhai, Q. Li, K. H. Kim, and G. Jeon, "Dense attention fusion network for object counting in iot system," *Mobile Networks and Applications*, vol. 28, pp. 359–368, 2023.

[27] X. Guo, M. Anisetti, M. Gao, and G. Jeon, "Object counting in remote sensing via triple attention and scale-aware network," *Remote Sensing*, vol. 14, no. 24, p. 6363, 2022.

[28] X. Guo, M. Gao, W. Zhai, J. Shang, and Q. Li, "Spatial-frequency attention network for crowd counting," *Big Data*, vol. 10, no. 5, pp. 453–465, 2022.

[29] W. Zhai, M. Gao, M. Anisetti, Q. Li, S. Jeon, and J. Pan, "Group-split attention network for crowd counting," *Journal of Electronic Imaging*, vol. 31, no. 04, Article ID 041214, 2022.

[30] W. Zhai, Q. Li, Y. Zhou et al., "Da 2 net: a dual attention-aware network for robust crowd counting," *Multimedia Systems*, vol. 29, pp. 3027–3040, 2022.

[31] Q. Zhang and A. B. Chan, "3d crowd counting via multi-view fusion with 3d 3D Crowd Counting via Multi-View Fusion with 3D Gaussian Kernelsaussian kernels," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 7, pp. 12837–12844, 2020.

[32] L. Zheng, Y. Li, and Y. Mu, "Learning factorized cross-view fusion for multi-view crowd counting," in *Proceedings of the 2021 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6, IEEE, Shenzhen, China, July 2021.

[33] Q. Zhang, W. Lin, and A. B. Chan, "Cross-view cross-scene multi-view crowd counting," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 557–567, Nashville, TN, USA, June 2021.

[34] Q. Zhang and A. B. Chan, "Calibration-free multi-view crowd counting," in *Proceedings of the Computer Vision–ECCV 2022: 17th European Conference*, pp. 227–244, Springer, Tel Aviv, Israel, October 2022.

[35] C. Liu, Y. Chen, X. He, and T. Xu, "A scale aggregation and spatial-aware network for multi-view crowd counting," *IEEE Access*, vol. 10, pp. 108604–108613, 2022.

[36] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," https://arxiv.org/abs/1409.1556.

[37] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: convolutional block attention module," in *Proceedings of the European conference on computer vision (ECCV)*, pp. 3–19, Munich, Germany, September 2018.

[38] J. Gao, Q. Wang, and X. Li, "Pcc net: PCC Net: Perspective Crowd Counting via Spatial Convolutional Networkerspective crowd counting via spatial convolutional network," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 10, pp. 3486–3498, 2020.

[39] D. B. Sam, N. N. Sajjan, R. V. Babu, and M. Srinivasan, "Divide and grow: capturing huge diversity in crowd images with incrementally growing cnn," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3618–3626, Salt Lake City, UT, USA, June 2018.

[40] V. A. Sindagi and V. M. Patel, "Generating high-quality crowd density maps using contextual pyramid cnns," in *Proceedings of the IEEE international conference on computer vision*, pp. 1861–1870, Venice, Italy, October 2017.

[41] J. Wan, W. Luo, B. Wu, A. B. Chan, and W. Liu, "Residual regression with semantic prior for crowd counting," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4036–4045, Long Beach, CA, USA, June 2019.

[42] Z. Shen, Y. Xu, B. Ni, M. Wang, J. Hu, and X. Yang, "Crowd counting via adversarial cross-scale consistency pursuit," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5245–5254, Salt Lake City, UT, USA, June 2018.

[43] X. Cao, Z. Wang, Y. Zhao, and F. Su, "Scale aggregation network for accurate and efficient crowd counting," in *Proceedings of the European conference on computer vision (ECCV)*, pp. 734–750, Munich, Germany, June 2018.

[44] M. Hossain, M. Hosseinzadeh, O. Chanda, and Y. Wang, "Crowd counting using scale-aware attention networks," in *Proceedings of the 2019 IEEE winter conference on applications of computer vision (WACV)*, pp. 1280–1288, IEEE, Waikoloa, HI, USA, January 2019.

[45] D. Onoro-Rubio and R. J. López-Sastre, "Towards perspective-free object counting with deep learning," in *Proceedings of the Computer Vision–ECCV 2016: 14th European Conference*, pp. 615–629, Springer, Amsterdam, The Netherlands, October 2016.

[46] Z. Yan, Y. Yuan, W. Zuo et al., "Perspective-guided convolution networks for crowd counting," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 952–961, Seoul, Korea, October 2019.

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