Hindawi
Scientific Programming
Volume 2022, Article ID 9863066, 21 pages
https://doi.org/10.1155/2022/9863066

Research Article

Research on 24-Hour Dense Crowd Counting and Object Detection System Based on Multimodal Image Optimization Feature Fusion

Guoyin Ren,[1] **Xiaoqi Lu**,[1,2] **and Yuhao Li**[1]

[1]*School of Mechanical Engineering, Inner Mongolia University of Science & Technology, Baotou 014010, China*
[2]*Inner Mongolia University of Technology, Hohhot 010051, China*

Correspondence should be addressed to Xiaoqi Lu; lan_tian1234@hotmail.com

Received 25 March 2022; Revised 7 August 2022; Accepted 27 August 2022; Published 16 September 2022

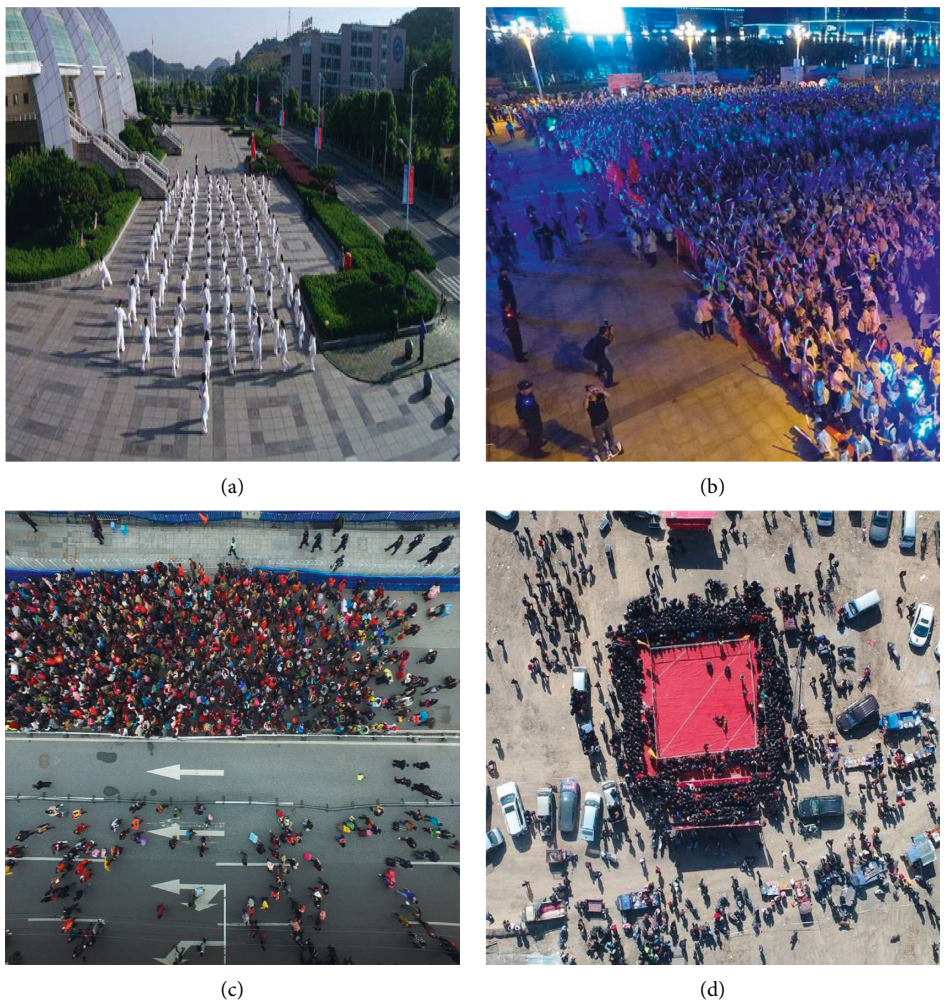
Academic Editor: Qianchuan Zhao

Motivation. In the environment of day and night video surveillance, in order to improve the accuracy of machine vision dense crowd counting and target detection, this paper designs a day and night dual-purpose crowd counting and crowd detection network based on multimode image fusion. *Methods*. Two sub-models, RGBD-Net and RGBT-Net, are designed in this paper. The depth image features and thermal imaging features are effectively fused with the features of visible light images, so that the model has stronger anti-interference characteristics and robustness to the light noise interference caused by the sudden fall of light at night. The above models use density map regression-guided detection method to complete population counting and detection. *Results*. The model completed daytime training and testing on MICC dataset. Through verification, the average absolute error of the model was 1.025, the mean square error was 1.521, and the recall rate of target detection was 97.11%. Night vision training and testing were completed on the RGBT-CC dataset. After verification, the average absolute error of the network was 18.16, the mean square error was 32.14, and the recall rate of target detection was 97.65%. By verifying the effectiveness of the multimode medium-term fusion network, it is found to exceed the current most advanced bimodal fusion method. *Conclusion*. The experimental results show that the proposed multimodal fusion network can solve the counting and detection problem in the video surveillance environment during day and night. The ablation experiment further proves the effectiveness of the parameters of the two models.

1. Introduction

Population estimation is the key to measuring population size. Not only can crowd estimation control the scene capacity of public scenes, but it is also an effective means to monitor emergencies. Although some advanced deep learning methods can realize the monitoring of multi-scale population and counting in complex scenes with good lighting environment, with the increase of population size and the influence of night light noise, it becomes difficult to pre-realize the robust day and night population counting and target positioning [1]. With the gradual popularization of depth imaging and thermal imaging equipment in the monitoring field, the depth imaging model based on depth learning can effectively solve the dense population count under daytime occlusion conditions [2]. The thermal imaging model based on deep learning helps to monitor, analyze, and count people at night. Therefore, how to solve the estimation of day and night dense population with depth imaging and thermal imaging becomes extremely challenging [3].

The deep learning model embedded in the traditional monitoring equipment can solve the scene with large-scale change and large light noise interference to a certain extent, but there are still some shortcomings to be improved. Traditional depth monitoring equipment cannot adapt to the multi-scale robustness brought by near targets, nor can it adapt to the light noise interference caused by the sudden drop of night light [4]. As shown in Figure 1, the scale and light intensity directly affect the population count. At present, there are few general counting models that can simultaneously deal with the problems of variable scale and night vision low light.



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

Figure 1: Images of dense crowds at different scales and different illuminations. (a) Multi-scale crowd. (b) Multi-scale crowd at night. (c) Single-scale crowd. (d) Single-scale distant crowd.

With the help of computer vision technology based on deep learning, target detection [5], face recognition [6], image segmentation [7], and visual tracking [8] can be realized. Different monitoring devices can achieve different image acquisition purposes. RGB, RGBD, and RGBT datasets are from three different types of cameras, including visible light cameras, depth camera sensors, and thermal imaging cameras. The visible light camera can collect the texture information of the crowd, but it can only work in the daytime. Thermal cameras can work during the day and at night and can avoid the interference of visible light noise [9–11]. But thermal imaging also has obvious disadvantages in population counting. For example, it is difficult to detect occluded crowd targets [12]. Based on the depth information and spatial layout of people and scenes provided by RGBD images, depth images can effectively distinguish the depth difference between objects in the scene, so as to locate the target and avoid the occlusion problem. Therefore, crowd counting has better performance than RGB images. Many researchers use depth images to complete crowd counting [13]. Both RGB and RGBD images are based on the visible light environment. However, the visible light data collected at night may have disadvantages, such as a large amount of light noise, which will make them ineffective under the weak light conditions at night. Fortunately, thermal imaging data have been shown to be effective in facilitating image analysis and allowing daytime and nighttime scene perception [14].

Existing methods of population counting include detection-based and regression-based methods [15–18]. Detection-based methods can count population in sparse scenes, but they are limited in crowded scenes. Detecting pedestrians to complete crowd counting is a simple solution. However, there is a serious problem. When the crowd is dense, the detector can easily fail due to the scale effect. Regression counting depends on some visual descriptors, such as texture features and edge features. Regression-based methods have been widely proved to be computationally feasible, parameter robust, and accurate in various challenging dense crowd scenarios. Regression includes direct regression and indirect regression. The method based on direct regression can estimate the number of people from the scene image using the linear regression function [19, 20]. Indirect regression regresses the population density map from the input image [21, 22] and then integrates the density

map to predict the population number. Since the density map can provide abundant spatial information, it reduces the difficulty of directly mapping the image to the estimation result. In general, the indirect regression estimation of the population number of the density map is proved to be more robust [23]. The proportion of the target in the image has changed significantly from near to far. For the proportion change problem, the viewpoint-aware multibranch CNN or switched CNN architecture can solve this challenge. These CNN structures are composed of multiple parallel CNN branches with different receptive field sizes and are used to process multi-scale feature regression from low-density images to high-density images. In a recent five-branch neural network, three branches are multiscale perceptual, while the remaining two branches act as density map estimators [24]. Zhou et al. [25] can realize multi-scale population counting in the complex scene of visible light by using the advanced adversarial learning model method.

In the aspect of multifeature fusion, the previous methods can only predict the density map from one-modal data. However, multimodal fusion can show the complementary multimodal features, so it is better than single-modal density map prediction results. Zhang et al. [26] proposed a multimodal fusion-based population counting model with multi-scale feature learning and fusion modules. All modules are jointly optimized and trained in an end-to-end manner. Effective extraction of low-level modal features and high-level modal fusion features can estimate more accurate density maps and more accurate population counts. Liu et al. [27] proposed an effective multimode fusion method, which combines depth data and thermal images with visible light features to improve the estimation results to a certain extent. This method discusses how to fuse the depth information and thermal imaging information with the visible light image and estimate the effective density map. The number of people in a crowded scene can be accurately calculated using Gaussian kernel. Learning features from the original data of depth images, thermal imaging images, and RGB images as inputs to realize the deep convolution network has significant advantages and is more robust to light noise, occlusion, and night lighting environment.

In terms of small target detection, the model proposed by Xia et al. [28] adds occlusion conditions to the correlation filtering algorithm, which can improve the robustness of small target detection position prediction. Finally, multi-target tracking and detection are realized by adaptive combination of multiple models. Chen et al. [29] proposed a learning method based on combinatorial representation, which uses the depth residual network and the depth neural network as generators and discriminators, respectively, and finally realizes the super-resolution reconstruction of small target faces, which is helpful to the detection and analysis of key targets in dense populations. Zhang et al. [30] designed a multifeature fusion method. The same weight is used to fuse various manual features, and then the adaptive weight is used to fuse manual features and depth features, which greatly improves the detection ability of the target object. Zhang et al. [31] proposed a method of integrating depth features and manual features in correlation filter learning and realized a robust target tracking and detection method with the help of correlation filter tracking and twin network.

The main contributions of this paper can be summarized in the following three aspects:

(i) Through the analysis of previous work, it is found that the two-stream multifeature fusion model has the possibility of further improvement. In this paper, RGBD-Net and RGBT-Net are proposed. The combination of these two models can realize the round-the-clock population counting. In this paper, three improved schemes of multimodal fusion are designed, which are early fusion, intermediate fusion, and late fusion. Extensive experiments and evaluations were performed on the MICC [32] and RGBT-CC [33] datasets. The medium-term fusion scheme performs best in the comprehensive evaluation of training time, counting error on small datasets, and detection recall rate; it is superior to some existing multimode fusion methods and has good generalization ability in the night vision environment crowd counting task.

(ii) Previous work rarely fused the RGBD depth features and thermal image features into the model at the same time. In this paper, two cross pattern fusion methods for crowd counting are established, namely, RGB and RGBD, and RGB and thermal image. In the first stage, we mainly focus on the capture of multimodal data. In the second stage, the RGBD multimode fusion network is used to solve the inter-day scale diversity and occlusion interference. The RGBT multimode fusion network solves the interference of light noise at night and realizes robust head counting when the light intensity changes suddenly. The final density maps and head detection of the two models were used as outputs.

(iii) In the past, there was little work on multi-scale small target detection in dense crowds. This paper attempts to combine adaptive Gaussian kernel with multimode fusion model to improve the night vision perception ability of the two-stream model for crowd images. In density map estimation, taking the depth feature and thermal image feature as the spatial priori of Gaussian kernel edge detection can enable the model to learn more spatial features and then complete the spatial target detection task of dense population.

2. Related Work

Early crowd counting methods [34–36] tended to rely on detection counting, i.e., detecting the head or body and then counting the population. However, in very crowded scenes, detection with high recall is difficult to achieve. Regression density map estimation methods have gradually replaced detection methods. For some large-scale scenes (stadiums, squares, etc.) or some large-scale gathering activities (festival parades, marathons, etc.), Paolanti proposed a method to

complete the crowd counting in images using the FCN detection framework; in [37, 57], FCN model that can be used for crowd counting is employed.

Some works use cameras to observe the dangerous situation of crowd gathering on the ground in real time to achieve the purpose of crowd evacuation; Miao proposed a lightweight CNN in [38], which can monitor the ground situation in real time to ensure the safety of crowd gathering. Compared with shallow learning methods, the crowd counting method based on deep CNN shows a significant performance improvement. Reference [39] proposed to combine the adaptive feature maps extracted from multiple layers to generate the final density map for the large variation of pedestrian scale. Zhou et al. [40] proposed a scale aggregation network to improve multi-scale representation and generate high-resolution density maps.

Due to the closeness of the acquisition target and the existence of angular distortion in monitoring, the crowd images acquired from oblique viewing angles have perspective distortion and scale changes. Highly dense crowds often suffer from severe occlusion and are characterized by nonuniform scaling: for example, individuals close to the camera are larger in scale, while individuals farther away from the camera show only small-scale head blobs. To address this issue, Boominathan proposed CrowdNet [41] using a combination of both shallow and deep structures, employing a data augmentation technique based on multi-scale pyramid representation. The model is robust to scale changes. However, these works did not prove to be equally effective in nighttime images.

Samuel et al. [42] addressed the problem of counting people from images. The method uses features based on thermal imaging and kernel density estimation, being more accurate and efficient than CNN-based methods for nighttime crowd counting. This method uses a fast feature detector to calculate the density map. Since the images used are taken from drones, the pedestrian targets in the images are small, and there is no occlusion problem caused by vertical and oblique viewing angles. There are also methods based on RGBD image CNN networks [43–45], because this method can reduce the occlusion problem caused by oblique viewing angles, and the error rate of crowd counting is low. However, the multicolumn structure leads to difficulty in training with many redundant parameters, and although the use of ensembles of CNNs can bring significant performance improvements, they come at the cost of a large amount of computation.

Considering the above shortcomings, many researchers began to use the method of multimodal data fusion combined with adaptive Gaussian kernel to replace the above scheme. Several methods based on two-modal fusion are used [46–48] to demonstrate the advantages of crowd counting in terms of day and night illumination, occlusion, and scale transformation by obtaining fused features. Two-stream models [49, 50] are proposed to fuse hierarchical cross-modal features to achieve fully representative shared features. In addition, there are methods [51] that explore the use of shared branches to map shared information into a common feature space. Furthermore, some recent works [52, 53] have been proposed to address RGBD and RGBT saliency fusion, which is also an example of a cross-modal dense crowd counting task.

Recently, multimodal images can be used to assist head counting and head localization [54–58]. However, methods utilizing depth images are ineffective at night or in low light conditions. In outdoor scenes, depth images often suffer from noise interference. At the same time, the detection range of depth detectors is limited, so depth-based methods have relatively limited deployment range.

Based on the above considerations, we find that thermal images are robust to illumination and can detect long perceptual distances. Liu proposed [34] a multimodal dynamic enhancement mechanism, which can make full use of more robust thermal modal images to increase the diversity of crowd counting features. Multimodal learning has received increasing attention in the field of computer vision. By integrating RGB and thermal image data, the model introduces soft cross-modal consistency among modalities and optimal query learning to improve robustness.

In order to advance the detection of pedestrians in various scenarios, multimodal learning can understand and represent cross-modal data by fusing models. There are various strategies for cross-modal feature fusion. These works have played a positive role in promoting the development of fusion models and are especially instructive for the work of day and night crowd counting.

3. Methods

3.1. Gaussian Kernel Density Map

3.1.1. Kernel Density Estimation Method for Adaptive Bandwidth. If the bandwidth of the Gaussian kernel is not fixed but varies according to the size of the head samples, this results in a particularly powerful method called adaptive or variable bandwidth kernel density estimation. Some parts of the population are highly aggregated, and some parts of the population are less aggregated. That is to say, different parts of the image should adopt different analysis scales, so this paper uses an unfixed bandwidth for kernel density estimation.

The kernel density estimation method of adaptive bandwidth is obtained by modifying the bandwidth parameters on the basis of the fixed bandwidth kernel density function, and its form is shown in the following formula:

$$
\begin{aligned}
k(x) &= \frac{1}{M}\sum_{j-1}^{M}\frac{1}{\left(wh_j\right)^n}K\left(\frac{x-x^{(j)}}{wh_j}\right),\\
K(x) &= \frac{1}{\sqrt{(2\pi)^n|S|}}\exp\left(-\frac{1}{2}x^T S^{-1}x\right),\\
h_j &= \left\{\left[\prod_{k=1}^{M} f\left(x^{(k)}\right)\right]^{\frac{1}{M}} f\left(x^{(j)}\right)\right\}^{\alpha},
\end{aligned}
\tag{1}
$$

where $k(x)$ is the Gaussian kernel density estimation function with bandwidth h_j, M is the number of heads in the crowd, and each Gaussian kernel density point j has a bandwidth h_j, so the bandwidth can be adaptive or variable. α is the sensitivity factor, $0 \le \alpha \le 1$, and is usually set to 0.5. When $\alpha = 0$, the Gaussian density estimation of adaptive bandwidth becomes the Gaussian density estimation of fixed bandwidth. The kernel density estimate of the fixed bandwidth is the kernel density estimate $k(x)$ mentioned earlier. ω represents the parameter of the bandwidth.

3.1.2. Depth and Thermal Image Adaptive Gaussian Kernel Density Map. The adaptive Gaussian kernel can make the density map regression clearer and produce a regression density map that is closer to the true density map. The adaptive Gaussian kernel can be closer to the real head size, and the density map generated by regression can provide the deep network with the prior knowledge of head detection, which can guide the position and size of the detection frame.

The center point of the human head image label is calculated with a standard Gaussian kernel function and converted into a crowd density map. Assuming that $C = \{x_1, x_2, \ldots, x_n\}$ is a dataset in d-dimensional space, an image has n head instances, and the number of instances is n, then the distribution density of the data can be expressed as follows:

$$\widehat{f}(x) = \frac{1}{nh^d} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right). \tag{2}$$

The multivariate Gaussian kernel function is given as follows:

$$K\left(\frac{x - x_i}{h}\right) = 12\pi^{d/2} \cdot \exp\left(-\frac{\|x - x_i\|}{2h^2}\right). \tag{3}$$

Among them, the Euclidean distance between x and x_i is $\|x - x_i\|$, h is the bandwidth, and the dimension is d. When the bandwidth h is equal to the head diameter h in the depth image, the estimated data amount of the dense population is n, and it can be expressed as follows:

$$n = \frac{4}{h^{(d+4)}(d+2)}. \tag{4}$$

When the bandwidth h is equal to the corresponding head diameter $h = R_{deep}$ in the depth image,

$$n = \frac{4}{R_{\text{deep}}^{(d+4)}(d+2)}. \tag{5}$$

When the bandwidth h is equal to the diameter $h = R_{thermal}$ of the corresponding head in the thermal image,

$$n = \frac{4}{R_{\text{thermal}}^{(d+4)}(d+2)}. \tag{6}$$

For a multimodal dataset $X = \{x_1, x_2, \ldots, x_n\}$, x_n represents each instance. Let its class label set be $F = \{c_1, c_2, \ldots, c_f\}$, where the number of classes is N_{ci}. Then, the density of instance x_i with respect to category c_i is calculated as follows:

$$f_{ci}(x_i) = \frac{1}{N_{ci} - 1} \sum_{i \ne j, i=1}^{n} \frac{1}{R_{de\,ep}^{d}\left(R_{thermal}^{d}\right)} \cdot L(x_i) \cdot K\left(\frac{x_i - x_j}{h}\right), L\left(x_j\right) = c_j. \tag{7}$$

Among them, $L(x_i)$ and $L(x_j)$ represent the labels of instances x_i and x_j, respectively. The adaptive Gaussian kernel formula is also applicable to the depth image and thermal image.

3.2. Network Design

3.2.1. Design of RGBD Fusion Network. In theory, a multimodal CNN with the above-mentioned early fusion can learn the features of the two-stream model, and SFVB is the fusion machine, as shown in Figure 2. Therefore, early fusion is usually more expressive than mid-level fusion, which can exploit correlations between modalities already on low-level CNN computations. However, the higher expressiveness comes at the cost of requiring more data for training. The benefit of late fusion is that most of the network initialization can be reused directly without adjusting the network weights based on additional inputs. Unfortunately, it does not allow the network to learn about such high-level interdependencies between individual input modalities, since only the resulting scores at the classification level are fused.

First, for the deep branch, we preprocess the dataset and perform initial training. The depth perceptron consists of two branch modules; the depth feature branch consists of four convolutional layers and three downsampling layers; the visible light feature branch consists of six convolutional layers, five downsampling layers, and one upsampling layer; the size of the convolution kernel is shown in Figure 2. This depth sensor is called RGBD-Net.

We believe that the filters needed for depth data are quite different from those obtained by training on RGB data. For example, we want the edge and speckle filters to be wider to be robust to noisy depth estimation.

In general, the training effect of fully supervised learning is the best, because fully supervised learning can annotate most of the head posture, lighting, and image perspective in the picture. However, the main disadvantage of full supervised learning is that the cost of labeling is too high in the face of a dense population. This paper chooses semi-supervised learning point labeling method to complete the training. The advantage is that the use of point annotations instead of full annotations can reduce the annotation cost on the premise of ensuring accurate positioning.

Deep MLP contains convolutional layers and max-pooling layers to quickly reduce the spatial resolution. This part is followed by further pooling layers, each of which halves the spatial dimension, as shown in Figure 3. We identify SEVB fusion points connecting depth and RGB networks. First, the RGB and depth inputs can be directly connected at the SEVB fusion point, and we call this model midterm fusion. The scores for the RGB network and the depth branch include two SEVB fusion points and finally use

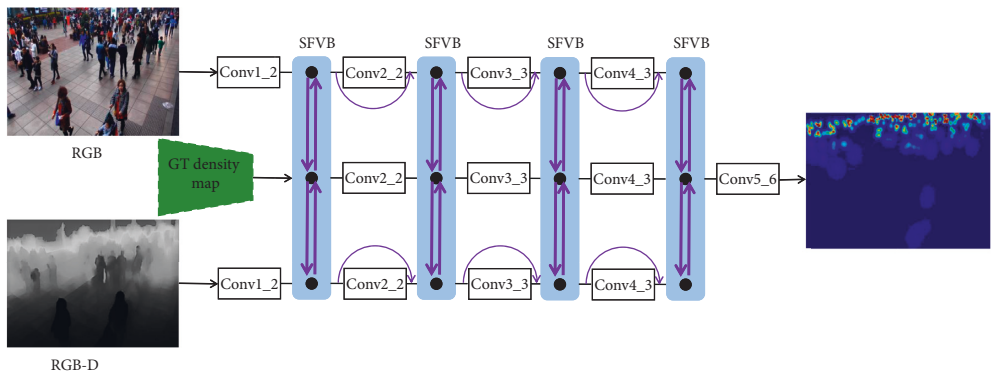


Figure 2: Structure diagram of early fusion depth perception network.

1×1 convolution as the classifier. The amount of upsampling used in this mid-level fusion method is determined by the desired spatial dimension in the RGB network.

The pipeline of the multimodal crowd counting network is shown in Figures 3 and 4. Our network has two specific modules, multi-scale feature learning module and SFVB module (including modality alignment module and adaptive fusion module). The feature learning module is used to extract general and modal features of the input data. The extracted pair of features are sent to the SFVB module to further extract high-level semantic features, and each pair of semantic features is aligned to the same feature space at the same time. After using high-level semantic features to regress the number of crowds, the pipeline fuses the prediction output of the multimodal data through the adaptive fusion module to obtain the final result.

3.2.2. Design of Thermal Imaging Fusion Network. By understanding the characteristics of different imaging principles, the study found that the feature learning of different modal data is different. An intuitive idea is to extract their distinguishing features separately. However, this will increase the parameters of the network and may reduce the efficiency. Furthermore, both streams also ignore modal sharing feature learning. In order to reduce the parameters, we use the feature extractor to obtain the information of the two branches before using the midterm fusion, use the modality extractor to extract the modal features, and then fuse them, as shown in Figures 3 and 4.

For the thermal image branch, we adjust the dataset for initial training. The thermal image perceptron consists of two branch modules; the thermal image feature branch consists of six convolutional layers, five downsampling layers, and one upsampling layer; and the visible light feature branch consists of four convolutional layers and three downsampling layers. The kernel size of the convolutional layer is shown in Figure 4. This thermal image perceptron is called RGBT-Net.

3.2.3. Advantages of Two-Stream Medium-Term Fusion Model. The feature fusion of depth image, thermal image, and visible image is a cross-modal problem, and the feature data may have high intra-modal variability, which makes the task of cross-modal image feature fusion very challenging. Medium-term integration can solve this problem to a certain extent. The SFVB modules, RGBD-Net and RGBT-Net, can first train each mode to a certain extent and then extract the midterm modal features of the two branches and realize feature fusion, as shown in Figures 3 and 4. In the two-flow model, SFVB module can not only train some high-level interdependent features between modes, but also ignore the shared feature learning of modal parts. This can not only reduce parameters, but also improve the fusion efficiency and enhance the ability of the network to judge the hierarchical characteristics of image depth and the ability to judge the characteristics of night crowd thermal images.

In contrast, early fusion and late fusion have obvious disadvantages.

Early fusion is also called feature level fusion. The principle is to extract the distinctive features of cross-modal images and complete the feature level fusion before the training process. The advantage of early fusion is that it can fully understand the high-level interdependency between various input modes. However, the disadvantage of doing so is that it will increase the parameters of the network and may reduce the efficiency of model parameter generation.

Late fusion is also called decision level fusion. The principle is to train each mode separately and then integrate the classification prediction scores of the model output layer after the training to generate the final decision. Therefore, it does not allow the network to understand the high-level interdependency between each input mode and the variability within the mode.

3.3. Density Map Guided Detection. Detection-based models such as RetinaNet cannot detect small/tiny heads because the detection subnet cannot adaptively adjust the anchors of these heads. However, our network benefits from adapting the Gaussian kernel density map with depth. The density map shows the distribution of the head in relation to the pixels of the Gaussian kernel in the density map. Therefore, we propose to feed the estimated density map into the detection network to improve the performance of detection of small/miniature heads. Heads of different scales are detected according to the decoding layers of the head RGBD and thermal images fed back by our network learning. In

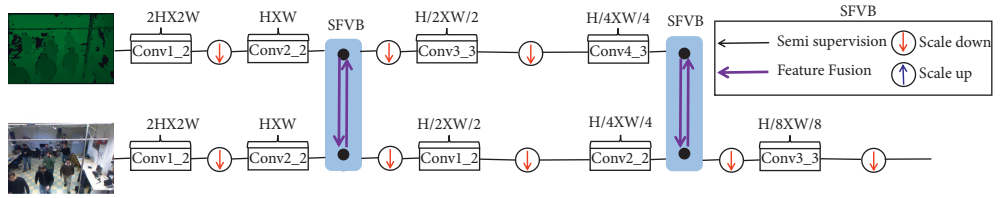


Figure 3: Structure of deep fusion perceptron.

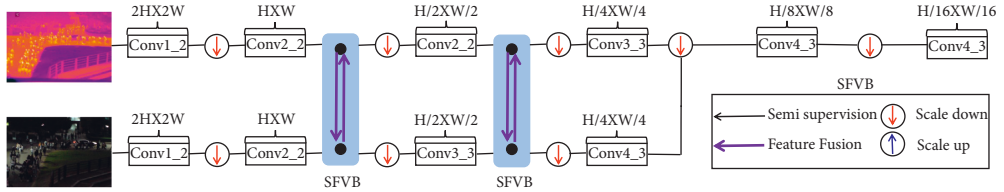


Figure 4: Thermal image fusion perceptron structural diagram.

depth image and thermal image preprocessing, the head RGBD map and thermal image are downsampled to the same size as the density map. For each head in M_l, RGBD and thermal image pixel values are used to reinforce our estimated density map. Specifically, for a given training head RGBD depth and thermal image size, it is assumed that the size of the head to be detected is marked with a rectangular box as the head size. Then, we generate a head frame feature matrix M_l through training and further fuse M_l with the density map function $D^A(x)$ to generate a density map constrained by RGBD and thermal image adaptive Gaussian kernel:

$$D_l^A(x) = D^A(x) \otimes M_l, \tag{8}$$

where $\otimes$ denotes feature fusion, $D_l{}^A(x)$ is the density map regression of RGBD and thermal image adaptive Gaussian kernel constraints, M_l is the head box feature matrix, and $D^A(x)$ is the adaptive Gaussian density map function.

3.4. Multimodal Fusion Loss Function. RGBD and thermal image multilayer perceptron combined with adaptive Gaussian kernel can solve the problem of adaptive perception of head size in day and night environment, where depth perceptron and thermal image perceptron are composed of RGB image and RGBD, and RGB image and thermal image, respectively. Dual-modal image feature fusion network is composed, fusing multimodal features through the import of the middle layer of the model to achieve feature alignment and adaptive fusion.

In this paper, an effective fusion of RGB features, RGBD features, and thermal image features is proposed to improve the accuracy of object recognition. We convert RGB image and RGBD features, and RGB image and thermal image to raw data vector input, and concatenate them; the input data can be represented as $\{x_{r1}, x_{r2}, \ldots, x =; x_{d1}, x_{d2}, \ldots, x_{dn}\}$, where $\{x_{r1}, x_{r2}, \ldots, x_{rn}\}$ represent the connection features of RGB images and RGBD, and $\{x_{d1}, x_{d2}, \ldots, x_{dn}\}$ represent the connection features of RGB images and thermal images. Then, the parameter matrix A corresponding to the input data can be expressed as follows:

$$A = \begin{bmatrix} A_{11}, A_{12} \cdots A_{1r_n} & A_{1(r_n+1)}, \cdots, A_{1(r_n+r_d)} & A_{1(r_n+r_d+1)}, \cdots, A_{1(r_n+r_d+(r_r))} \\ A_{21}, A_{22} \cdots A_{2r_n} & A_{2(r_n+1)}, \cdots, A_{2(r_n+r_d)} & A_{2(r_n+r_d+1)}, \cdots, A_{2(r_n+r_d+(r_r))} \\ \vdots & \vdots & \vdots \\ A_{k1}, A_{k2} \cdots A_{kr_n} & A_{k(r_n+1)}, \cdots, A_{k(r_n+r_d)} & A_{k(r_n+r_d+1)}, \cdots, A_{k(r_n+r_d+(r_r))} \end{bmatrix}. \tag{9}$$

The first half of A is the following:

$$A_{RGB} = \begin{bmatrix} A_{11}, A_{12} \cdots A_{1r_n} \\ A_{21}, A_{22} \cdots A_{2r_n} \\ \vdots \\ A_{k1}, A_{k2} \cdots A_{kr_n} \end{bmatrix}. \tag{10}$$

The second half of the parameter corresponding to the RGB is image vector A_{Depth}:

$$A_{\text{Depth}} = \begin{bmatrix} A_1(r_n+1), \cdots, A_1(r_n+r_d) \\ A_2(r_n+1), \cdots, A_2(r_n+r_d) \\ \vdots \\ A_k(r_n+1), \cdots, A_k(r_n+r_d) \end{bmatrix}. \tag{11}$$

$A_{Thermal}$ is the parameter corresponding to the RGBT thermal image vector, and k represents all possible class labels.

$$A_{\text{Thermal}} = \begin{bmatrix} A_{1(r_n+1)}, \cdots, A_{1(r_n+r_r)} \\ A_{2(r_n+1)}, \cdots, A_{2(r_n+r_r)} \\ \vdots \\ A_{k(r_n+1)}, \cdots, A_{k(r_n+r_r)} \end{bmatrix}, \tag{12}$$

$$\begin{aligned} J_{\text{sparse}} = {} & \frac{1}{m}\sum_{i=1}^{m}\left(\frac{1}{2}\left\|h_{w,b}\left(x^i\right)-x^i\right\|^2\right) \\ & +\frac{1}{2}\sum_{l=1}^{n_l-1}\sum_{i=1}^{s_l^r}\sum_{j=1}^{s_{l+1}^r}\left(\lambda_{RGB}\left(A_{RGB}^l\right)_{ij}\right)^2 \\ & +\frac{1}{2}\sum_{l=1}^{n_l-1}\sum_{i=1}^{s_l^d}\sum_{j=1}^{s_{l+1}^d}\left(\lambda_{RGB}\left(A_{\text{Depth (Termal)}}^l\right)_{ij}\right)^2+\beta P(x). \end{aligned} \tag{13}$$

J_{sparse} is the parameter corresponding to the depth image vector, and k represents all possible class labels.

$$\lambda_{RGB} = \begin{bmatrix} \lambda_{r1} \\ \lambda_{r2} \\ \ddots \\ \lambda_{rk} \end{bmatrix}, \tag{14}$$

$$\lambda_{Depth} = \begin{bmatrix} \lambda_{d1} \\ \lambda_{d2} \\ \ddots \\ \lambda_{dk} \end{bmatrix}, \tag{15}$$

$$\lambda_{\text{thermal}} = \begin{bmatrix} \lambda_{t1} \\ \lambda_{t2} \\ \ddots \\ \lambda_{tk} \end{bmatrix}. \tag{16}$$

The overall cost function is shown in the following equation:

$$\begin{aligned} J_{\text{sparse}} = {} & \frac{1}{m}\sum_{i=1}^{m}\left(\frac{1}{2}\left\|h_{w,b}\left(x^i\right)-x^i\right\|^2\right) \\ & +\frac{1}{2}\sum_{l=1}^{n_l-1}\sum_{i=1}^{s_l^r}\sum_{j=1}^{s_{l+1}^r}\left(\lambda_{RGB}\left(A_{RGB}^l\right)_{ij}\right)^2 \\ & +\frac{1}{2}\sum_{l=1}^{n_l-1}\sum_{i=1}^{s_l^d}\sum_{j=1}^{s_{l+1}^d}\left(\lambda_{\text{Depth}}\left(A_{\text{Depth}}^l\right)_{ij}\right)^2 \\ & +\left(\frac{1}{2}\sum_{l=1}^{n_l-1}\sum_{i=1}^{s_l^l}\sum_{j=1}^{s_{l+1}^l}\left(\lambda_{\text{thermal}}\left(A_{\text{thermal}}^l\right)_{ij}\right)^2+\beta P(x)\right). \end{aligned} \tag{17}$$

The weight decay parameter λ_{RGB} corresponding to the RGB feature of the object is initialized to a smaller value to reduce its penalty and extract more RGB features. The weight attenuation parameter λ_{depth} corresponding to the depth feature is initialized to a larger value, and the penalty is increased to extract fewer depth features. The weight attenuation parameter $\lambda_{thermal}$ corresponding to the thermal imaging feature of the object is initialized to a larger value, and the penalty is increased to extract fewer thermal image features.

3.5. Overview of Methods. The system mainly includes two-part method: dense crowd counting and head target detection. The pseudocode of the crowd counting part is shown in Algorithm 1.

3.5.1. Training Parameter Setting. The head target detection is shown in Figure 5. First, we mark the center point of the head, train the crowd counting network, generate the density map, and set the mean value of the initial parameters of the multivariate Gaussian function μ to 0.5 and the standard deviation σ to 0.02. The Gaussian multivariate function approximates the head size of the thermal image with the Gaussian kernel function bandwidth according to the center of the thermal image annotation. The pseudocode of the crowd head detection part is shown in Algorithm 2.

3.5.2. Head Scale Judgment. When the bandwidth h of the Gaussian kernel function is equal to the head size of the thermal image or the head size of the thermal image, the Gaussian kernel expansion is stopped and the Gaussian kernel density map is generated. Otherwise, the head size continues to be matched. As shown in Figure 5, this network designs three fusion modes, namely, early fusion, intermediate fusion, and late fusion, to complete the experiment. In this paper, the three fusion schemes are tested separately, so the type of current fusion mode needs to be preset before the network runs. After the fusion mode is selected, the density map constrained by the adaptive Gaussian kernel of the thermal image is further fused by M_l and the density map function D^A, and the counting is completed. Finally, the head detection frame is generated using the boundary of the adaptive Gaussian kernel constrained by the density map.

4. Experiments

4.1. Dataset Introduction and Evaluation Criteria. The depth perceptron of this crowd counting method has been experimentally evaluated on the MICC dataset (RGBD). The thermal image perceptron has been evaluated experimentally on the RGBT-CC dataset (RGBT), and the feasibility and applicability of our proposed method have been verified through experimental comparison. We first give the parameters of the key datasets used in the experiments. In each dataset, the corresponding method in this paper is compared with the most advanced crowd counting method in this dataset, and the density map and the real and estimated values of the crowd on the pictures of each dataset are given.

Input: **Training** density map **image** $\S = \{D_1, \ldots, D_N\}$, **training**
Epochs N_{ci}, and Adaptive Gaussian kernel initialization input initial mean $\mu = 0.5$ and standard deviation $\sigma = 0.02$
Onput: A dense crowd detection model with parameters θ **AND crowd head detection box**
Initializing density map **image** D and **parameters** θ
for each epoch **do**
Step 1:
If $(h == h_{\text{thermal}})$
Gauss kernel matching stop;
else Continue to match head size according to (5);
If $(h == h_{\text{Depth}})$
Gauss kernel matching stop;
else Continue to match head size according to (6);
Step 2: Multivariate Gauss matching with all head sizes;
Step 3: Multimode features for midterm fusion, according to (9–17);
Step 4: If (If (8) is true)
Generating density maps with adaptive Gaussian kernel constraints;
else Regenerate Gaussian density map according to (7)
Step 5: Density map guided generation of crowd head detection box;
Step 6: Update D according to (4).
end

Algorithm 1: Training density map guided detection for RGBD-Net/RGBT-Net.

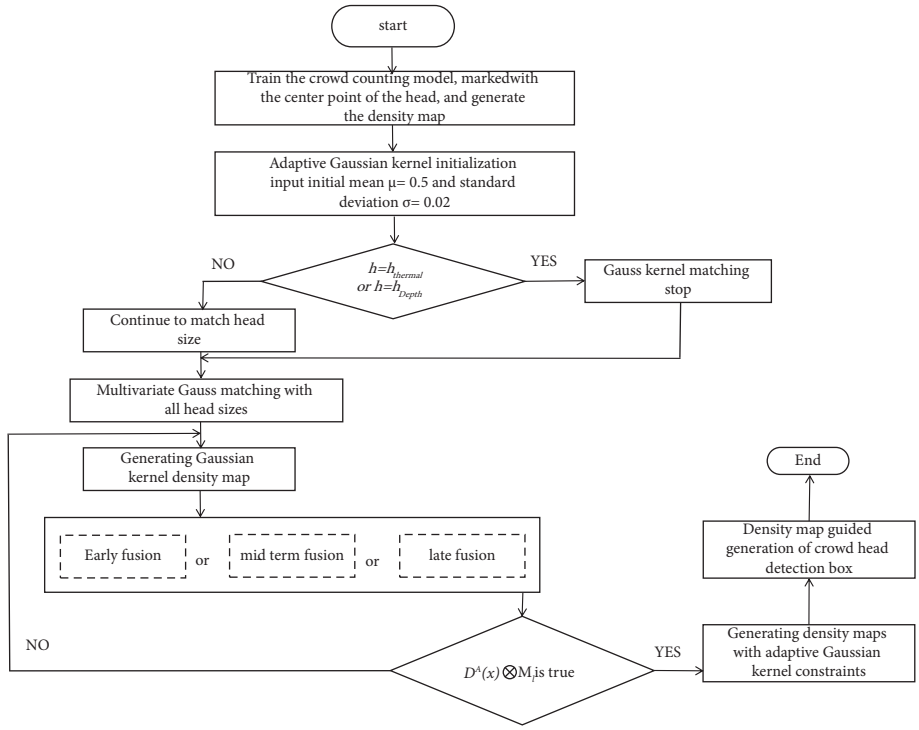


Figure 5: System global design flowchart.

Finally, this paper conducts ablation experiments to demonstrate the independent effectiveness of each method unit in our method ensemble and finally gives the effect diagram of day and night crowd counting.

MICC dataset. The MICC dataset is a crowd image dataset taken from an indoor fixed scene. A total of 3,358 frames of crowd RGB + depth images are obtained. The video resolution is 480×640 pixels. The maximum number of head annotations in a single frame is 11, and the minimum number of head annotations in a single frame is 0. There are 17,630 head annotations. The MICC dataset contains three kinds of video sequences: stream sequence, team sequence, and group sequence. The flow sequence includes a total of 1,260 frames with 3,542 pedestrian bounding box annotations, the queue sequence contains 918 frames with 5,031 pedestrian bounding box annotations, and the group sequence includes 9,057 pedestrian

Input: **Training** density map **image** $\S = \{D_1, \ldots, D_N\}$**, training**
Epochs N_{ci}, and Adaptive Gaussian kernel initialization input initial mean $\mu = 0.5$ and standard deviation $\sigma = 0.02$
Onput: A dense crowd detection model with parameters θ **AND crowd head detection box**
Initializing density map **image** D and **parameters** θ
for each epoch **do**
Step 1:
If $(h == h_{\text{thermal}})$
Gauss kernel matching stop;
else Continue to match head size according to (5);
If $(h == h_{\text{Depth}})$
Gauss kernel matching stop;
else Continue to match head size according to (6);
Step 2: Multivariate Gauss matching with all head sizes;
Step 3: Multimode features for midterm fusion, according to (9–17);
Step 4: If (If (8) is true)
Generating density maps with adaptive Gaussian kernel constraints;
else Regenerate Gaussian density map according to (7)
Step 5: Density map guided generation of crowd head detection box;
Step 6: Update D according to (6).
end

Algorithm 2: Training density map guided detection for RGBD-Net/RGBT-Net.

bounding box annotations in 1,180 frames. RGB images all correspond to RGBD images. In streaming sequences, people walk from one location to another, and the frame rate is lower. In the queuing sequence, the acquisition frame rate is larger and pedestrians move slowly. In the group sequence, people are constrained to the area of action, as shown in Table 1.

RGBT-CC dataset. The RGBT-CC dataset is a large-scale RGB-thermal crowd counting dataset proposed by Sun Yat-Sen University. This dataset contains a large number of RGB-thermal images collected using photothermal cameras in scenes such as shopping malls, streets, and subway stations. The RGB images are cropped and preprocessed by the dataset publisher from a high resolution of $2{,}048 \times 1{,}536$ and resized to a resolution of 640×480, and the thermal image resolution is 640×480. The dataset includes 2,030 pairs of annotated RGB-thermal images: 1,013 pairs of daytime crowd images and 1,017 pairs of nighttime crowd images. The dataset has 138,389 pedestrians labeled with point annotations, with an average of 68 pedestrians per thermal image. The RGBT-CC dataset is a close-range collection of urban populations with a wide range of densities, as shown in Table 1. Therefore, this dataset is more challenging.

Metrics. We use mean absolute error (MAE) and mean squared error (MSE) to evaluate different methods based on commonly used metrics in existing crowd counting work:

$$MAE = \frac{1}{N}\sum_{1}^{N}\left|z - \widehat{z}_i\right|,$$
$$MSE = \sqrt{\frac{1}{N}\sum_{1}^{N}\left(z_i - \widehat{z}_i\right)^2}, \tag{18}$$

where N is the total number of test images, z_i is the actual number of people in the ith test image, and Z_i is the estimated number of people in the ith image.

4.2. Dataset Parameter Setting and Training

MICC dataset. Since the three sequences of flow, group, and queue have the same scene and participant characteristics, 20% of the RGB images of the flow, group, and queue scenes and their corresponding RGBD images can be input as the training set, and the remaining 80% of the RGB images and their corresponding RGBD images are used as the test set. This dataset is for training a crowd counting network with depth perception.

RGBT-CC dataset. RGBT-CC is the benchmark RGBT dataset, which includes two types of images: bright and dark. The bright training set includes 510 pairs of images, the validation set includes 97 pairs of images, and the test set includes 406 pairs of images; the dark training set includes 520 pairs of images, the validation set includes 103 pairs of images, and the test set includes 394 pairs of images; these images were randomly assigned to the final three new sets: training, validation, and test. Finally, 1,030 image pairs were used for training, 200 image pairs were used for validation, and 800 images were used for testing.

Training settings. We train the RGBD-Net and RGBT-Net models end to end. The Gaussian parameters of the adaptive Gaussian kernel are set by the mean value to 0.5 and the standard deviation to 0.02. The RGBD-Net and the RGBT Net are trained on the MICC dataset and the RGBT-CC dataset. The network selects the momentum random gradient descent (SGD) during training, sets the initial learning rate to 0.005, and sets the momentum to 0.85. Its

Table 1: Parameter information for MICC dataset and RGBT-CC dataset.

Dataset	Resolution	Color	Num	Max	Min	Ave	Total	Modality
MICC	480 × 640	RGB + D	3,358	11	0	5.2	17,630	RGB + depth
RGBT-CC	640 × 480	RGB + T	4,060	82	45	68	138,389	RGB + thermal

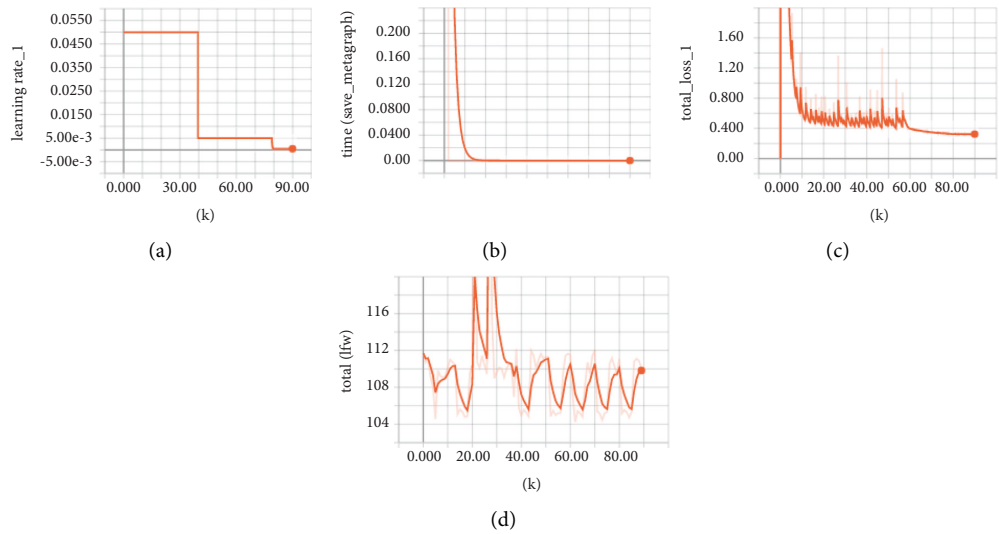


Figure 6: The training process. (a) Learning rate setting curve. (b) Time variation curve of training and saving weights. (c) Loss function variation curve. (d) Total training time variation curve.

convergence speed is fast, and the training process is shown in Figure 6. Our methods are all implemented under the PyTorch framework. In terms of hardware, three NVIDIA 1080 Ti GPU graphics cards and four Intel® E5-2630 v4 CPUs are used to ensure the performance requirements of graphics cards and computing units.

4.3. Comparison with State-of-the-Art Methods

4.3.1. Crowd Counting

(1) MICC dataset. We collect the specific data of the state-of-the-art methods in the field of crowd counting, give the performance of these methods on the MICC dataset, and give the comparison results between the method used in this paper and the current state-of-the-art crowd counting methods in different categories. From Table 2, it can be found that the performance of different categories of advanced methods shows regular changes, so this paper only compares the results of methods in recent literature, as shown in Table 2.

Object detection method. RetinaNet [59] utilizes ResNet and efficient pyramid feature network and adopts anchor boxes to detect crowd size error: MAE = 1.641, MSE = 2.554. The structure of DetNet [60] is improved on the basis of ResNet50 because ResNet50 itself has excellent performance and improves detection, so the results are in the MICC dataset. The MAE and MSE indicators achieved an improvement of 0.1 and 0.172. Idress et al. [61] used multi-source features including SIFT and head detection to achieve a 0.145 improvement in the MICC dataset MAE metric compared to DetNet for detecting crowd numbers. Although these detection frameworks can detect some small-scale targets, because there is no effective strategy designed, the target detection method is obviously unable to cope with the dense and small crowds with serious occlusion.

Table 2: Comparison of the different state-of-the-art methods on MICC dataset.

Model	MAE	MSE
RetinaNet [59]	1.641	2.554
DetNet [60]	1.541	2.382
Idrees et al. [61]	1.396	2.642
MCNN [62]	1.5	2.259
CSRNet [63]	1.359	2.125
Cascaded-DCNet [64]	0.836	1.031
MCNN-adaptive [61]	1.489	2.114
CSRNet-adaptive [61]	1.343	2.007
RDNet [65]	1.38	2.551
Ours (RGBD-Net-adaptive)	1.025	1.521

Density regression method. The advanced regression methods are all methods based on CNN density map regression [62–64]. The accuracy of advanced methods based on CNN density map regression is significantly higher than that of detection-based methods. For example, the results of CSRNet [63] are better than those of Idrees et al. [61] with 0.011 and 0.145 improvement in MAE and MSE metrics for MICC dataset, and MCNN is better than DetNet in terms of MAE and MSE metrics with 0.041 and 0.123 improvement.

However, the density regression method lacks deeper spatial features, so the method will be more likely to fail and cannot localize pedestrian spatial locations. But Cascaded-DCNet [64] improves on spatial feature extraction; it is a cascaded depth-aware counting network that jointly performs head segmentation and density map regression on MICC datasets. In terms of MAE and MSE indicators, it achieved excellent scores of 0.836 and 1.031.

Density regression-guided detection method. The density regression-guided detection method makes full use of the RGBD image dataset to add depth perception and adaptive Gaussian kernel to the deep network to realize the crowd counting method of regression-guided detection. MCNN-adaptive and CSRNet-adaptive are based on the density map regression. The depth perception branch is added, thus achieving 0.1 and 0.172 improvement over MCNN [62] and 0.016 and 0.118 improvement over CSRNet [63] in terms of MAE and MSE metrics on the MICC dataset. The accuracy of crowd counting can be significantly improved, and the position of the detection frame can be obtained. However, such methods do not surpass Cascaded-DCNet on the MICC dataset, because Cascaded-DCNet performs maximum pooling on the head region in the depth image. Therefore, the failure of the depth map can be avoided. However, the expensive labeling cost limits the further application of monitoring domain counting and detection tasks. The advanced RDNet [65] utilizes box annotations for head detection. It is worth noting that the original MICC dataset contains head box annotations and we only use the center of each box as its point annotation.

As shown in Figure 7, we select images from the MICC dataset with the most to the few head counts for testing. We use a depth-aware fusion network and an adaptive Gaussian kernel and estimate the head density map. By observing the results on the MICC dataset, we can see that the final result of our network achieves MAE = 1.025 and MSE = 1.521, surpassing all the above methods, because Cascaded-DCNet [64] uses strong head segmentation. Supervised method, while we only use the center point of each head annotation as a weakly supervised method for training.

The analysis of the qualitative results shows that our method performs well on the MICC database. The main reason is that our proposed network learns more spatial context information using a better depth-sensing network and an adaptive Gaussian kernel, which is consistent with our original motivation. The results verify the effectiveness of our method.

(2) RGBT-CC dataset. The RGBT-CC dataset is a heatmap crowd counting dataset published by Sun Yat-Sen University. As shown in Table 3, the comparison results between the method used in this paper and the current advanced crowd counting methods of different categories are given. Because the current experimental new method of the Sun Yat-Sen University team is the first to propose RGBT-CC and use the dataset to complete the comparison of different fusion models, on the basis of this scheme, our method is improved and compared with experiments. We still use classical counting models such as MCNN [33], SANet [69], CSRNet [33], and Bayesian Loss [70] as the backbone networks for experimental reference. Compared with the best result of Bayesian Loss [70], our model has improved MAE and MSE by 0.54 and 0.53 on the RGBT-CC dataset. Similarly, our method is compared with multiple best-performing multimodal fusion models of UCNet [66], HDFNet [67], and BBSNet [68], and compared with the best-performing BBSNet [7], it is found that our model has 1.4 and 0.34 improvement in MAE and MSE on the RGBT-CC dataset. In the literature of the Sun Yat-Sen University team, the integration of the IADM "early fusion" mechanism into the classical counting model networks such as MCNN [33], SANet [69], CSRNet [33], and Bayesian Loss [70] can improve the performance of the model. The MAE and MSE of MCNN + IADM [33] have an improvement of 2.12 and 2.14; the MAE and MSE of SANet + IADM [33] have an improvement of 3.81 and 7.88; the MAE and MSE of CSRNet + IADM [33] have an improvement of 2.56 and 4.35; and the MAE and MSE of Bayesian Loss + IADM [33] have a 3.09 and 4.49 improvement. The difference is that instead of using the "early fusion" method to take the concatenation of RGB and thermal images as input, we use the "midterm fusion" method to make a comparison with the "early fusion" method. In addition, we finally complete the density map regression using an adaptive Gaussian kernel. After comparison, it is found that our "midterm fusion" model and "early fusion" have an almost equivalent improvement in MAE and MSE on the RGBT-CC dataset.

We demonstrate the effectiveness of our method in generating density maps from thermal images under different lighting conditions, and we select images with different lighting conditions as count objects on the RGBT-CC dataset. The light conditions in the first column of pictures gradually become darker from top to bottom. What is most obvious is that the first picture is a daytime indoor scene and the bottom one is a nighttime street scene, as shown in Figure 8. The second column of pictures is the thermal map under different lighting conditions. The third column of pictures is the crowd density map of BL + IADM [33] under different lighting conditions. The fourth column of pictures is the population density map of BL + RDNA under different lighting conditions.

Figure 8

In order to verify the effectiveness of our method in generating density maps under different crowd conditions, we selected images with different crowd conditions as the counting objects on the MICC dataset. The first column of images from top to bottom gradually decreases, as shown in Figure 7. The second column of pictures is the depth map under the condition of different numbers of people. The third column of pictures is the crowd density map of CSRNet-adaptive under different lighting conditions. The fourth column of pictures is the population density map of CSRNet + RDNA under different population density conditions. Figure 7 shows the results on three real datasets. We pretrain the depth perceptron and regressor. From the generated density map results on the MICC dataset, the density map of our method is very close to the ground-truth

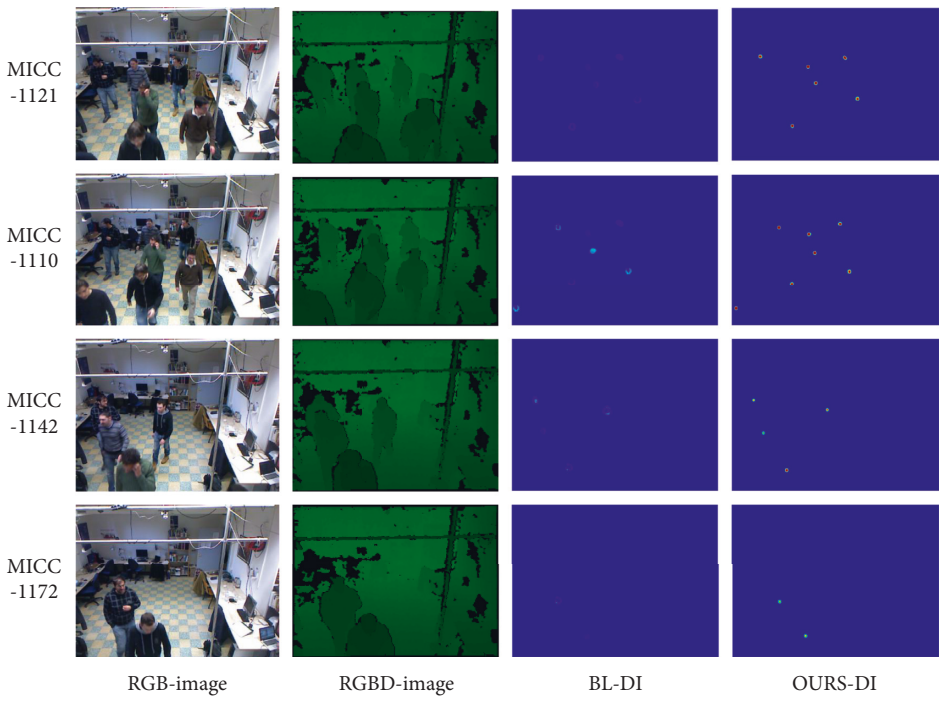


Figure 7: Visualization results from ShanghaiTech, UCF_CC_50, and UCF-QNRF datasets. From left to right: the input images, ground truth, and results of SCLNet.

Table 3: Comparison of the different state-of-the-art methods on RGBT-CC dataset.

Model	MAE	MSE
UCNet [66]	33.96	56.31
HDFNet [67]	22.36	33.93
BBSNet [68]	19.56	32.48
MCNN [33]	21.89	37.44
SANet [69]	21.99	41.6
CSRNet [33]	20.4	35.26
Bayesian Loss [70]	18.7	32.67
MCNN + IADM [33]	19.77	30.34
SANet + IADM [33]	18.18	33.72
CSRNet + IADM [33]	17.94	30.91
Bayesian Loss + IADM [33]	15.61	28.18
Ours (RGBT-Net)	18.16	32.14

density map of CSRNet-adaptive, which to some extent proves the effectiveness and counting performance of our depth perceptron. The fusion feature of the features of the depth map plays an effective role when the number of crowds becomes smaller and smaller, thus proving the effectiveness of this method.

4.3.2. Crowd Detection. At present, it is difficult for the head detection of dense crowds to detect the human head under obvious occlusion and different lighting; especially at night, the detection will fail. In order to solve the occlusion problem and the effectiveness of night head detection on small pixel heads, this paper uses two perceptrons: depth perceptron and thermal image perceptron. Head counting and regression-guided detection can be accomplished with the help of an adaptive Gaussian kernel, as shown in Figure 9. In Figure 9, the first row is the result of density regression-guided head detection based on depth perceptron + adaptive Gaussian kernel. The first image is from the MICC dataset, the second image is the corresponding depth image, the third image is the original annotation frame, and the fourth picture is the annotation frame estimated by the method in this paper. In Figure 9, the second row is the result of density regression-guided head detection based on thermal image perceptron + adaptive Gaussian kernel. The first image is from the RGBT-CC dataset, and the second image is the corresponding thermal image. The first three pictures are the original annotation boxes, and the fourth picture is the annotation boxes estimated by the method in this paper. The method used in this paper is to use the center of the annotation box from the dataset as the training target to locate the boundary of the estimated box according to the adaptive Gaussian kernel and the corresponding perceptron. First, we extract the head position points (green points) from the ground-truth annotations, and then we extract the head position points (red points) in the density map using the layer perceptron and RGBD depth fusion. From Figure 9, the depth sensor can be used to detect all heads in the room, the thermal image sensor can detect people's heads in the nighttime environment, and the heads of people occluded under the dark image can still be detected. The ROC curves of front fusion, midrange fusion, and end fusion of the RGBD-Net and RGBT-Net networks combined with the adaptive Gaussian kernel are shown in Figures 10(a) and 10(b).

This paper collects the target detection methods related to the dense population in the past three years, classifies the

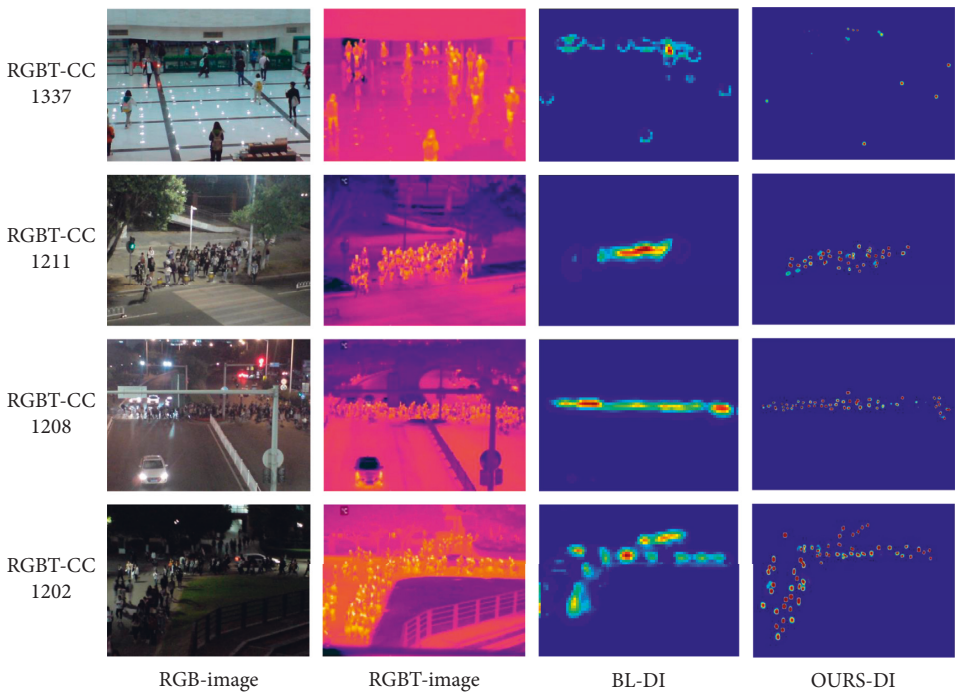


Figure 8: Visualization results from ShanghaiTech, UCF_CC_50, and UCF-QNRF datasets. From left to right: the input images, ground truth, and results of SCLNet.

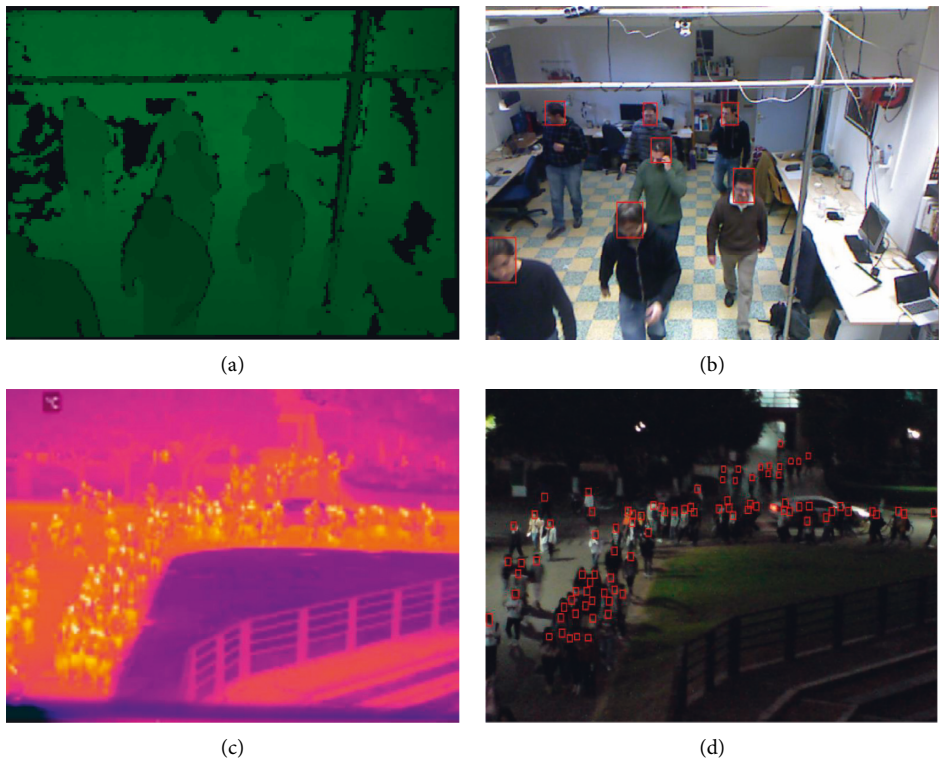


Figure 9: Crowd detection results on deep datasets and crowd detection results on thermal image datasets. (a) RGBD crowd. (b) RGBD crowd detection. (c) RGBT crowd. (d) RGBT crowd detection.

methods into RGBD and RGBT, and conducts comparative tests on the ShanghaiTechRGBD and KAIST datasets. The application scenarios of the methods are divided into three types: day, night, and day and night. The effectiveness of the method proposed in this paper is evaluated by comparing the accuracy of all the methods in three different lighting scenarios. After analyzing the results in Table 4, it is found that the method used in this paper has the best accuracy in

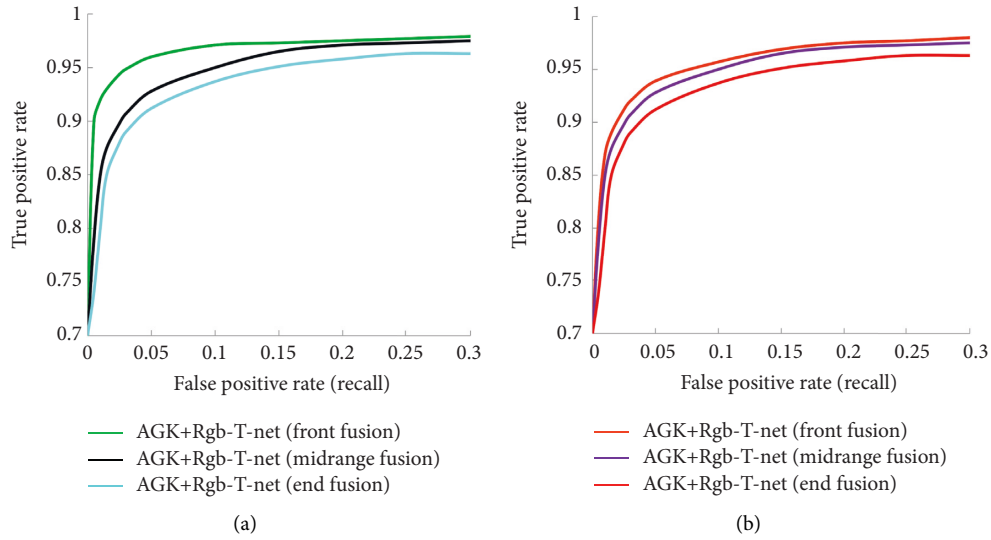


Figure 10: (a) Depth perceptron + adaptive Gaussian kernel early fusion; depth perceptron + adaptive Gaussian kernel late fusion; depth perceptron + adaptive Gaussian kernel mid fusion precision-recall curves of all object classes and all object classes and average precision-recall curves of our method. (b) Thermal image perceptron + adaptive Gaussian kernel early fusion; thermal image perceptron + adaptive Gaussian kernel mid fusion; thermal image perceptron + adaptive Gaussian kernel late fusion precision-recall curves for all object classes and all object classes and mean precision-recall curves of our method.

both RGBD and RGBT modes and performs best especially in the environment of day and night.

As shown in Table 5, compared with the most advanced methods (DetNet, RDNet, and CGD), the one of this paper has the best detection accuracy on the ShanghaiTechRGBD dataset. These methods are based on the detection module, and there is a certain gap in performance with the feature fusion and regression guidance detector in this paper. The performance of this detector is better than DetNet, RDNet, and CGD. It can be observed from the results that the detector (YOLO-T) that only processes thermal images does not perform as well as YOLO4-RGBT. After using RGBT, it can be observed that the performance of YOLO4-RGBT is similar to YOLO4-T, but YOLO4-T is slightly lower. After analysis, it is believed that this may be because the use of multimodal post-fusion feature learning can improve the target judgment of YOLO4-RGBT network. Among them, the average accuracy (AP) of YOLO4-RGB detector is the worst, which proves the advantage of data fusion in day and night target detection.

Table 4: Comparison of the different state-of-the-art methods on RGBT-CC dataset.

Model	+	MAE	MSE
MCNN [33]	N/A	21.89	37.44
SANet [69]	N/A	21.99	41.6
CSRNet [33]	N/A	20.4	35.26
Bayesian Loss [70]	N/A	18.7	32.67
MCNN + IADM [33]	N/A	19.77	30.34
MCNN [33]	+RTNA	18.04	29.16
SANet + IADM [33]	N/A	18.18	33.72
SANet [69]	+RTNA	17.98	32.04
CSRNet + IADM [33]	N/A	17.94	30.91
CSRNet [33]	+RTNA	17.82	30.14
Bayesian Loss + IADM [33]	N/A	15.61	28.18
Bayesian Loss [70]	+RTNA	15.48	27.96
Ours (RGBT-Net + AGK)	—	18.16	32.14

4.4. Ablation Study

4.4.1. Effectiveness of Depth Perceptrons. We conduct ablation experiments on the effectiveness of depth perceptrons for crowd counting. As shown in Table 6, four different variables are selected for qualitative analysis, and we construct a depth-sensing network from RGB and depth images using "early fusion," "medium fusion," and "late fusion." There will be obvious differences in the fusion networks of different stages. We only use the adaptive Gaussian kernel AGK in the reference comparison: MAE = 1.367, MSE = 2.458. But compared to the use of only the adaptive Gaussian kernel AGK to finally make the regression of the density map, the depth-aware network can constrain the edge expansion of each Gaussian kernel to be more effective for dense crowds, which means that the combination of the adaptive Gaussian kernel function AGK with RGBD depth information awareness is helpful in crowd counting and MAE and MSE are smaller. MAE = 1.004 and MSE = 1.489 for AGK + RGBD-Net (front fusion) on the MICC dataset. MAE = 1.025 and MSE = 1.521 for AGK + RGBD-Net (midrange fusion) on the MICC dataset. MAE = 1.256 and MSE = 1.925 for AGK + RGBD-Net (end fusion) on the MICC dataset. The reason for this result is as follows:

(a) In the process of "early fusion" of the RGB branch and the depth branch, the RGB image and the depth image can be directly connected from the initial input stage to build the first convolutional layer. We refer to this form of fusion as early fusion. The CNNs

Table 5: Comparison of the different state-of-the-art target detection methods in day and night.

Methods	Type	Dataset	Day (AP)	Night (AP)	Day + night (AP)
DetNet [71]	RGBD	ShanghaiTechRGBD	0.383	—	—
RDNet [71]	RGBD	ShanghaiTechRGBD	0.610	—	—
CGD [71]	RGBD	ShanghaiTechRGBD	0.727	—	—
YOLO4-RGB [72]	RGBT	KAIST	0.684	0.298	0.465
YOLO4-T [72]	RGBT	KAIST	0.641	0.617	0.625
YOLO4-RGBT [72]	RGBT	KAIST	0.648	0.609	0.618
Ours	RGBD	ShanghaiTechRGBD	0.825	0.806	0.816
Ours	RGBT	KAIST	0.726	0.715	0.721

Table 6: Comparison of different feature fusion and normalization methods for the MICC dataset.

Method	MAE	MSE
AGK	1.367	2.458
AGK + RGBD-Net (front fusion)	1.004	1.489
AGK + RGBD-Net (midrange fusion)	1.025	1.521
AGK + RGBD-Net (end fusion)	1.256	1.925

built in the RGB branch and depth in the early fusion can enhance the dependencies of independent network flows by fully learning the features of the two modalities. Therefore, early fusion can theoretically fully understand the high-level interdependencies and accuracy advantages between various input modalities more than intermediate fusion. However, the cost of the higher accuracy advantage is that training may require more data, but the current RGBD dataset is not as large as the RGB dataset, so the early fusion is limited by the amount of data.

(b) The weights of the RGBD branch are first trained using midterm fusion, and the RGBD depth branch can be merged into a 1×1 convolutional layer before one max-pooling layer of the RGB network. This is done so that the RGBD CNN branch can use weights that have been pretrained on the RGB branch and the network can fully understand the high-level interdependencies between the various input modalities. This will greatly reduce the total training time and does not require more RGBD training data, which also satisfies the actual problem of insufficient RGBD datasets. Although the final accuracy is not as small as the error of early fusion, the difference is not much. MAE and MSE are only 0.021 and 0.032, so the sacrificed accuracy is completely acceptable to us compared with training more data and consuming more time.

(c) The RGB network and the RGBD depth branch can be combined into a multimodal classifier at the last concatenated layer of the network. This fusion mode is called late fusion. The advantage of late fusion is that the network weights do not have to be initialized repeatedly, which can be reused despite additional input network weights. Unfortunately, it does not allow the network to learn about such high-level interdependencies between individual input modalities, since only the resulting scores at the classification level are fused.

Therefore, the midterm fusion + AGK method of the depth perceptron used in this paper can meet the premise of lack of time and training data, and the accuracy of MAE = 1.025 and MSE = 1.521 is also satisfactory to us.

4.4.2. Effectiveness of Thermal Image Perceptrons. We conduct ablation experiments on the effectiveness of thermal image perceptrons. As shown in Table 7, four different variables are selected for qualitative analysis. We use the methods of "early fusion," "mid-phase fusion," and "late fusion" to construct a thermal image perception network from RGB and thermal images, and fuse them at different stages. There will be obvious differences in the network of AGK, and we only use MAE = 22.46 and MSE = 38.97 in the reference comparison of AGK. But compared to the use of only AGK to finally make the complete density map regression, the thermal image-aware network can constrain the edge expansion of each Gaussian kernel to be more effective for dense crowds, which means that the adaptive Gaussian kernel function AGK with RGBT thermal image information awareness is more effective. The combination is helpful in crowd counting, and the MAE and MSE errors are smaller. MAE = 18.01 and MSE = 31.49 for AGK + RGBT-Net (front fusion) on the MICC dataset. The MAE = 18.16 and MSE = 32.14 for AGK + RGBT-Net (midrange fusion) on the MICC dataset. The MAE = 19.35 and MSE = 34.71 for AGK + RGBT-Net (end fusion) on the MICC dataset. Therefore, the midterm fusion + AGK method of the thermal image perceptron used in this paper can meet the premise of lack of time and training data, and the accuracy of MAE = 18.16 and MSE = 32.14 is also satisfactory to us.

4.4.3. Effectiveness of Advanced Networks Based on Depth Perceptron. We use MCNN, CSRNet, RetinaNet, DetNet, Idrees et al., and RDNet as backbone networks for comparative experiments. The performance of all the comparative methods is shown in Table 8. It can be observed that all our new attempts to add RDNA on all backbones consistently outperform the corresponding backbones. For example, compared with the "mid-stage fusion" model of the backbone network MCNN, MCNN + RDNA on the MICC dataset has a 0.295 and 0.45 improvement in MAE and MSE, respectively; compared with the "mid-stage fusion" model of

Table 7: Comparison of different feature fusion and normalization methods for the RGBT-CC dataset.

Method	MAE	MSE
AGK	22.46	38.97
AGK + RGBT-Net (front fusion)	18.01	31.49
AGK + RGBT-Net (midrange fusion)	18.16	32.14
AGK + RGBT-Net (end fusion)	19.35	34.71

Table 8: Comparison of the different state-of-the-art methods on MICC dataset.

Model	MAE	MSE	+	MAE	MSE
MCNN [62]	1.5	2.259	+RDNA	1.205	1.719
MCNN-adaptive [61]	1.489	2.114	N/A	—	—
CSRNet [63]	1.359	2.125	+RDNA	1.195	1.723
CSRNet-adaptive [61]	1.343	2.007	N/A	—	—
RetinaNet [59]	1.641	2.554	+RDNA	1.489	1.962
DetNet [60]	1.541	2.382	+RDNA	1.356	1.861
Idrees et al. [61]	1.396	2.642	+RDNA	1.299	1.849
RDNet [65]	1.38	2.551	+RDNA	1.289	1.787
Cascaded-DCNet [64]	0.836	1.031	N/A	—	—
Ours (RGBD-Net + AGK)	1.025	1.521	—	—	—

the backbone network CSRNet, on the MICC dataset, CSRNet + RDNA has an improvement of 0.164 and 0.402 in MAE and MSE, respectively; compared with the "midterm fusion" model of the backbone network RetinaNet, on the MICC dataset, RetinaNet + RDNA has 0.152 and 0.592 improvement in MAE and MSE, respectively. Compared with the "midterm fusion" model of the backbone network DetNet, DetNet + RDNA has an improvement of 0.185 and 0.521 in MAE and MSE, respectively, on the MICC dataset; compared with the "fusion" model, the backbone network + RDNA of Idrees et al. has an improvement of 0.097 and 0.793 in MAE and MSE, respectively, on the MICC dataset; on the MICC dataset, the backbone network + RDNA of RDNet has an improvement of 0.091 and 0.764 in MAE and MSE, respectively; our method is more accurate than MCNN-adaptive and CSRNet-adaptive networks in terms of RGBD-Net + AGK fusion, because our method explicitly learns the interdependency and complementarity of RGB and RGBD, while MCNN-adaptive and CSRNet-adaptive simply add depth-adaptive discriminative capabilities without mutually enhancing network features. The reason why RGBD-Net + AGK does not exceed Cascaded-DCNet is that Cascaded-DCNet uses "early fusion." Our fusion method is midterm fusion. The reason for early fusion superiority over midterm fusion has been given in the previous section. But Cascaded-DCNet outperforms our method only by 0.189 and 0.49 in MAE and MSE, respectively. This comparison demonstrates the effectiveness of the state-of-the-art networks based on depth perceptrons.

4.4.4. Thermal Image Perceptron-Based Advanced Network Effectiveness. The performance of all comparative methods is shown in Table 4. It can be observed that all instances of our method consistently outperform the corresponding backbone networks. For example, MCNN + IADM and SANet + IADM have an 18.9% relative performance improvement on RMSE compared to their "early fusion" model. Furthermore, our CSRNet + IADM and BL + IADM achieve better performance on all evaluation metrics compared to other advanced methods (i.e., UCNet, HDFNet, and BBSNet). This is because our method explicitly learns specific shared representations and mutually enhances each other, while other methods simply fuse multimodal features without mutual enhancement. Therefore, our method can better capture the complementarity of RGB images and thermal images. This comparison demonstrates the effectiveness of our RGBT crowd counting method.

First, we use MCNN, SANet, CSRNet, and Bayesian Loss as the backbone networks to participate in comparative experiments. The performance of all comparative methods is shown in Table 4. It can be observed that all our new attempts to add RDNA on all backbones consistently outperform the corresponding backbones. For example, compared with the backbone network MCNN model, MCNN + RDNA on the MICC dataset has a 2.12 and 8.28 improvement in MAE and MSE, respectively; compared with the model of the backbone network CSRNet, CSRNet + RDNA has an improvement of 2.58 and 5.12 in MAE and MSE, respectively, on the MICC dataset; compared with the model of the backbone network Bayesian Loss, Bayesian Loss + RDNA has an improvement of 3.22 and 4.71 in MAE and MSE, respectively, on the MICC dataset; for MCNN + IADM, SANet + IADM, and Bayesian Loss + IADM, our method is in RGBD-Net + AGK fusion. The accuracy exceeds these two networks because IADM uses "early fusion." Our fusion method is midterm fusion. The reason why early fusion exceeds midterm fusion has been given in the previous section. This comparison demonstrates the effectiveness of the state-of-the-art networks based on thermal image perceptrons.

Figure 11: Performance of model transfer learning.

5. Conclusions

In this paper, we propose a heatmap fusion network (RGBD-Net and RGBT-Net) for crowd counting in daytime and night vision environments, and a guidance detection method combined with adaptive Gaussian kernel. A population counting and density estimation method based on cross-modal fusion of RGB, RGBD, and thermal images is established. A large number of experiments and evaluations have been carried out on the MICC and RGBT-CC public datasets. This method is superior to the existing multimodal two-stream fusion population counting method in terms of training time and retrieval recall. At the same time, it has a good promotion effect on the RGB population counting task. In the multimode fusion model, the medium-term fusion model can be used to extract the channel features of the image, and the Gaussian model can be used to extract the spatial edge constraint features of the image for the final population count estimation. This method has achieved satisfactory results in population counting. In terms of guidance detection, RGBD-Net + AGK and RGBT-Net + AGK can realize day and night vision counting and detection of dense population. From the results, the RGBD-Net + AGK model has completed the daytime training and testing on the MICC dataset. Through verification, the average absolute error of the model is 1.025, the mean square error is 1.521, and the target detection recall rate is 97.11%. The average absolute error of the RGBT-Net + AGK model in the RGBT-CC open dataset is 18.16, the mean square error is 32.14, the detection recall rate is 97.65%, and the robustness to occlusion and night complex scenes is good. This day and night counting method can solve the application of some actual scenarios. One possible application in the future is to use UAVs to count the number of people and locate special targets in the battlefield background, and help the UAVs (or UAV groups) integrated with search and strike to form effective attacks (or reasonable distribution of the attacked objects) above the crowd, so as to maximize the attack power.

6. Discussion

The comprehensive use of RGBD-Net and RGBT-Net can realize the day and night crowd counting and detection. However, the method in this paper has the disadvantage that the generalization ability of target detection in other datasets is not good. In this paper, the trained model is used to complete the target detection on the Shanghai Science and Technology dataset, and it is found that as long as there is an occluded head, it is not detected, as shown in Figure 11. This is one of the important reasons that, in the future, we will extend our proposed method to the field of UAV night vision attack and video crowd counting and detection, especially to improve the real-time processing ability of the entire algorithm.

Data Availability

The MICC and RGBT-Net datasets used to support the findings of this study can be obtained from the corresponding author upon request.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This work was supported by National Natural Science Foundation of China (nos. 61179019 and 81571753), CERNET Innovation Project of China (no. NGII20170705), and Baotou Youth Innovative Talent Project of China (no. 0701011904).

References

[1] L. Deng, S. H. Wang, and Y. D. Zhang, "Fully optimized convolutional neural network based on small-scale crowd," in *Proceedings of the 2020 IEEE International Symposium on Circuits and Systems (ISCAS)*, pp. 1–5, IEEE, Seville, Spain, October 2020.

[2] M. Xu, Z. Ge, X. Jiang et al., "Depth information guided crowd counting for complex crowd scenes," *Pattern Recognition Letters*, vol. 125, pp. 563–569, 2019.

[3] B. Zhang, Y. Du, and Y. Zhao, "I-MMCCN: improved MMCCN for RGB-T crowd counting of drone images," in *Proceedings of the 2021 7th IEEE International Conference on Network Intelligence and Digital Content (IC-NIDC)*, pp. 117–121, IEEE, Beijing, China, 17-19 November 2021.

[4] Z. Fan, H. Zhang, Z. Zhang, G. Lu, Y. Zhang, and Y. Wang, "A survey of crowd counting and density estimation based on convolutional neural network," *Neurocomputing*, vol. 472, pp. 224–251, 2022.

[5] Y. Liu, P. Wang, and H. Wang, "Target tracking algorithm based on deep learning and multi-video monitoring," in *Proceedings of the 2018 5th International Conference on Systems and Informatics (ICSAI)*, pp. 440–444, IEEE, Nanjing, China, 10-12 November 2018.

[6] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," *Learning*, vol. 23, p. 23746, 2015.

[7] Z. Mahmood, N. Muhammad, N. Bibi, and T. Ali, "A review on state-of-the-art face recognition approaches," *Fractals*, vol. 25, no. 02, p. 1750025, 2017.

[8] J. W. Johnson, "Adapting mask-rcnn for automatic nucleus segmentation," 2018, http://arxiv.org/abs/1805.00500.

[9] T. Peng, Q. Li, and P. Zhu, "Rgb-t crowd counting from drone: a benchmark and mmccn network," in *Proceedings of the Asian Conference on Computer Vision*, Singapore, Novomber 1-4 2020.

[10] T. Zhang, X. Liu, and Q. Zhang, "SiamCDA: complementarity-and distractor-aware RGB-T tracking based on Siamese network," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 56, p. 129, 2021.

[11] S. S. Shivakumar, N. Rodrigues, and A. Zhou, "Pst900: rgb-thermal calibration, dataset and segmentation network," in *Proceedings of the 2020 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 9441–9447, IEEE, 2020.

[12] K. Rassels and P. French, "Accurate body temperature measurement of a neonate using thermography technology," in *Proceedings of the 2021 Smart Systems Integration (SSI)*, pp. 1–5, IEEE, Grenoble, France, April 2021.

[13] G. Ren, X. Lu, and Y. Li, "Research on local counting and object detection of multiscale crowds in video based on time-frequency analysis," *Journal of Sensors*, p. 634, 2022.

[14] D. Lian, J. Li, and J. Zheng, "Density map regression guided detection network for rgb-d crowd counting and localization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1821–1830, Long Beach, CA, USA, June 2019.

[15] A. Shehzed, A. Jalal, and K. Kim, "Multi-person tracking in smart surveillance system for crowd counting and normal/abnormal events detection," in *Proceedings of the 2019 international conference on applied and engineering mathematics (ICAEM)*, pp. 163–168, IEEE, Taxila, Pakistan, August 2019.

[16] Z. Tu, T. Xia, C. Li, X. Wang, Y. Ma, and J. Tang, "RGB-T image saliency detection via collaborative graph learning," *IEEE Transactions on Multimedia*, vol. 22, no. 1, pp. 160–173, 2020.

[17] D. Zhou and Q. He, "Cascaded multi-task learning of head segmentation and density regression for RGBD crowd counting," *IEEE Access*, vol. 8, pp. 101616–101627, 2020.

[18] B. Zhang, Y. Du, and Y. Zhao, "I-MMCCN: improved MMCCN for RGB-T crowd counting of drone images," in *Proceedings of the 2021 7th IEEE International Conference on Network Intelligence and Digital Content (IC-NIDC)*, pp. 117–121, IEEE, Beijing, China, November 2021.

[19] C. Liu, Y. Huang, and Y. Mu, "DRENet: giving full scope to detection and regression-based estimation for video crowd counting," *International Conference on Artificial Neural Networks*, pp. 15–27, Springer, Cham, 2021.

[20] Y. Shi, J. Sang, and J. Tan, "GC-MRNet: gated cascade multi-stage regression network for crowd counting," *International Conference on Artificial Neural Networks*, pp. 53–66, Springer, Cham, 2021.

[21] X. Jiang, Z. Xiao, and B. Zhang, "Crowd counting and density estimation by trellis encoder-decoder networks," *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6133–6142, 2019.

[22] X. Liu, J. Yang, and W. Ding, "Adaptive mixture regression network with local counting map for crowd counting," *European Conference on Computer Vision*, pp. 241–257, Springer, Cham, 2020.

[23] J. Wan and A. Chan, "Adaptive density map generation for crowd counting," *IEEE/CVF International Conference on Computer Vision*, pp. 1130–1139, 2019.

[24] D. Wu, P. Yan, and Y. Guo, "A gear machining error prediction method based on adaptive Gaussian mixture regression considering stochastic disturbance," *Journal of Intelligent Manufacturing*, pp. 1–19, 2021.

[25] Y. Zhou, J. Yang, H. Li, T. Cao, and S. Y. Kung, "Adversarial learning for multiscale crowd counting under complex scenes," *IEEE Transactions on Cybernetics*, vol. 51, no. 11, pp. 5423–5432, 2021.

[26] S. Zhang, H. Li, and W. Kong, "A cross-modal fusion based approach with scale-aware deep representation for RGB-D crowd counting and density estimation," *Expert Systems with Applications*, vol. 180, p. 115071, 2021.

[27] L. Liu, J. Chen, and H. Wu, "Cross-modal collaborative representation learning and a large-scale rgbt benchmark for crowd counting," *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4832-4833, 2021.

[28] R. Xia, Y. Chen, and B. Ren, "Improved anti-occlusion object tracking algorithm using Unscented Rauch-Tung-Striebel smoother and kernel correlation filter," *Journal of King Saud University-Computer and Information Sciences*, vol. 34, p. 345234, 2022.

[29] Y. Chen, V. Phonevilay, J. Tao et al., "The face image super-resolution algorithm based on combined representation learning," *Multimedia Tools and Applications*, vol. 80, no. 20, pp. 30839–30861, 2021.

[30] J. Zhang, W. Feng, T. Yuan, J. Wang, and A. K. Sangaiah, "SCSTCF: spatial-channel selection and temporal regularized correlation filters for visual tracking," *Applied Soft Computing*, vol. 118, p. 108485, 2022.

[31] J. Zhang, J. Sun, J. Wang, Z. Li, and X. Chen, "An object tracking framework with recapture based on correlation filters and Siamese networks," *Computers & Electrical Engineering*, vol. 98, p. 107730, 2022.

[32] D. Lian, J. Li, and J. Zheng, "Density map regression guided detection network for rgb-d crowd counting and localization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1821–1830, Long Beach, CA, USA, June 2019.

[33] Z. You, K. Yang, and W. Luo, "Iterative correlation-based feature refinement for few-shot counting," 2022, http://arxiv.org/abs/2201.08959.

[34] D. Liu, K. Zhang, and Z. Chen, "Attentive cross-modal fusion network for RGB-D saliency detection," *IEEE Transactions on Multimedia*, vol. 23, pp. 967–981, 2021.

[35] D. Lian, X. Chen, and J. Li, "Locating and counting heads in crowds with a depth prior," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, p. 8734, 2021.

[36] C. Luo, J. Zhang, J. Yu, C. W. Chen, and S. Wang, "Real-time head pose estimation and face modeling from a depth image," *IEEE Transactions on Multimedia*, vol. 21, no. 10, pp. 2473–2481, 2019.

[37] M. Paolanti, R. Pietrini, A. Mancini, E. Frontoni, and P. Zingaretti, "Deep understanding of shopper behaviours and interactions using RGB-D vision," *Machine Vision and Applications*, vol. 31, no. 7-8, pp. 66–21, 2020.

[38] Y. Miao, J. Han, Y. Gao, and B. Zhang, "ST-CNN: spatial-temporal convolutional neural network for crowd counting in videos," *Pattern Recognition Letters*, vol. 125, pp. 113–118, 2019.

[39] X. Liu, J. Yang, and W. Ding, "Adaptive mixture regression network with local counting map for crowd counting," *European Conference on Computer Vision*, pp. 241–257, Springer, Cham, 2020.

[40] Y. Zhou, J. Yang, H. Li, T. Cao, and S. Y. Kung, "Adversarial learning for multiscale crowd counting under complex scenes," *IEEE Transactions on Cybernetics*, vol. 51, no. 11, pp. 5423–5432, 2021.

[41] L. Boominathan, S. S. S. Kruthiventi, and R. V. Babu, "Crowdnet: a deep convolutional network for dense crowd counting," in *Proceedings of the 24th ACM International Conference on Multimedia*, pp. 640–644, China, 2016.

[42] M. Samuel, M. A. Samuel-soma, and F. F. Moveh, "AI driven thermal people counting for smart window facade using portable low cost miniature thermal imaging sensors," *Window Facade*, p. 45, 2020.

[43] M. Gochoo, S. A. Rizwan, Y. Y. Ghadi, A. Jalal, and K. Kim, "A systematic deep learning based overhead tracking and counting system using RGB-D remote cameras," *Applied Sciences*, vol. 11, no. 12, p. 5503, 2021.

[44] Y. Yao, X. Zhang, and Y. Liang, "A real-time pedestrian counting system based on rgb-d," in *Proceedings of the 2020 12th International Conference on Advanced Computational Intelligence (ICACI)*, pp. 110–117, IEEE, Dali, China, 14-16 August 2020.

[45] M. Xu, "An efficient crowd estimation method using convolutional neural network with thermal images," in *Proceedings of the 2019 IEEE International Conference on Signal, Information and Data Processing (ICSIDP)*, pp. 1–6, IEEE, Chongqing, China, 11-13 December 2019.

[46] Z. Tang, T. Xu, and H. Li, "Exploring fusion strategies for accurate RGBT visual object tracking," 2022, http://arxiv.org/abs/2201/08673.

[47] G. Xu, X. Li, and X. Zhang, "Loop closure detection in RGB-D SLAM by utilizing siamese ConvNet features," *Applied Sciences*, vol. 12, no. 1, p. 62, 2022.

[48] C. Li, X. Wu, N. Zhao, X. Cao, and J. Tang, "Fusing two-stream convolutional neural networks for RGB-T object tracking," *Neurocomputing*, vol. 281, pp. 78–85, 2018.

[49] W. Zhang, X. Guo, J. Wang, N. Wang, and K. Chen, "Asymmetric adaptive fusion in a two-stream network for RGB-D human detection," *Sensors*, vol. 21, no. 3, p. 916, 2021.

[50] W. Zhou, J. Jin, and J. Lei, "CEGFNet: common extraction and gate fusion network for scene parsing of remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, p. 1232, 2021.

[51] H. Chen, Y. Li, and D. Su, "Discriminative cross-modal transfer learning and densely cross-level feedback fusion for RGB-D salient object detection," *IEEE Transactions on Cybernetics*, vol. 50, no. 11, pp. 4808–4820, 2020.

[52] Y. Fang, S. Gao, J. Li, W. Luo, L. He, and B. Hu, "Multi-level feature fusion based Locality-Constrained Spatial Transformer network for video crowd counting," *Neurocomputing*, vol. 392, pp. 98–107, 2020.

[53] A. G. Abuarafah, M. O. Khozium, and E. AbdRabou, "Real-time crowd monitoring using infrared thermal video sequences," *Journal of American Science*, vol. 8, no. 3, pp. 133–140, 2012.

[54] X. Zhang, G. Yu, and T. Chen, "Occlusion region searching and segmentation for multi-human detection based on RGB-D information," in *Proceedings of the 2018 IEEE Fourth International Conference on Multimedia Big Data (BigMM)*, pp. 1–4, IEEE, Xi'an, China, 13-16 September 2018.

[55] C. Orrite-Uruñuela and D. Vicente-Dueñas, "Counting people by infrared depth sensors," in *Proceedings of the 2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pp. 1–6, IEEE, Auckland, New Zealand, 27-30 November 2018.

[56] C. W. Liu, A. Breakspear, D. Guan et al., "NIN acts as a network hub controlling a growth module required for rhizobial infection," *Plant Physiology*, vol. 179, no. 4, pp. 1704–1722, 2019.

[57] A. Ben-Cohen, E. Klang, and S. P. Raskin, "Cross-modality synthesis from CT to PET using FCN and GAN networks for improved automated lesion detection," *Engineering Applications of Artificial Intelligence*, vol. 78, pp. 186–194, 2019.

[58] M. Yamazaki, A. Kasagi, and A. Tabuchi, "Yet another accelerated sgd: resnet-50 training on imagenet in 74.7 seconds," 2019, http://arxiv.org/abs/1903.12650.

[59] H. Zhang, H. Chang, and B. Ma, "Cascade retinanet: maintaining consistency for single-stage object detection," 2019, http://arxiv.org/abs/1907.06881.

[60] Z. Li, C. Peng, and G. Yu, "Detnet: a backbone network for object detection," 2018, http://arxiv.org/abs/1804.06215.

[61] F. Xiong, X. Shi, and D. Y. Yeung, "Spatiotemporal modeling for crowd counting in videos," *Proceedings of the IEEE International Conference on Computer Vision*, pp. 5151–5159, 2017.

[62] Y. Zhang, D. Zhou, and S. Chen, "Single-image crowd counting via multi-column convolutional neural network," *The IEEE conference on computer vision and pattern recognition*, pp. 589–597, 2016.

[63] Y. Li, X. Zhang, and D. Chen, "Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes," *IEEE conference on computer vision and pattern recognition*, pp. 1091–1100, 2018.

[64] X. Liu, J. Sang, W. Wu, K. Liu, Q. Liu, and X. Xia, "Density-aware and background-aware network for crowd counting via multi-task learning," *Pattern Recognition Letters*, vol. 150, pp. 221–227, 2021.

[65] F. Hong, C. Lu, W. Jiang, W. Ju, and T. Wang, "RDNet: regression dense and attention for object detection in traffic symbols," *IEEE Sensors Journal*, vol. 21, no. 22, pp. 25372–25378, 2021.

[66] L. Liu, J. Chen, and H. Wu, "Cross-modal collaborative representation learning and a large-scale RGBT benchmark for crowd counting," 2020, http://arxiv.org/abs/2012.04529.

[67] Z. Liu, Z. He, and L. Wang, "VisDrone-CC2021: the vision meets drone crowd counting challenge results," *Proceedings of the IEEE/CVF International Conference on Computer Vision*, vol. 179, no. 4, pp. 2830–2838, 2021.

[68] W. Zhou, Q. Guo, and J. Lei, "ECFFNet: effective and consistent feature fusion network for RGB-T salient object detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 45, pp. 4789–4801, 2021.

[69] J. Fan, X. Yang, and R. Lu, "Design and implementation of intelligent inspection and alarm flight system for epidemic prevention," *Drones*, vol. 5, no. 3, pp. 38–73, 2021.
[70] M. Fischer and A. Vignes, "An imprecise bayesian approach to thermal runaway probability," *International Symposium on Imprecise Probability: Theories and Applications*, pp. 150–160, PMLR, 2021.
[71] N. Japar, V. J. Kok, and C. S. Chan, "Coherent group detection in still image," *Multimedia Tools and Applications*, vol. 80, no. 14, pp. 22007–22026, 2021.
[72] K. Roszyk, M. R. Nowicki, and P. Skrzypczyński, "Adopting the YOLOv4 architecture for low-latency multispectral pedestrian detection in autonomous driving," *Sensors*, vol. 22, no. 3, p. 1082, 2022.