Sheep Counting Method Based on Multiscale Module Deep Neural Network

JIANMING XU, WEICHUN LIU, YANG QIN, AND GUANGRONG XU

Provincial Key Laboratory of Informational Service for Rural Area of Southwestern Hunan, Shaoyang University, Shaoyang, Hunan 422000, China

Corresponding author: Jianming Xu (3099523591@qq.com)

This work was supported by the Scientific Research Project of Hunan Education Department under Grant 19C1652.

ABSTRACT Due to the uneven distribution and large scale change of sheep in the pasture, it is not conducive to the counting and statistics of sheep in animal husbandry. The traditional target counting algorithm has low counting accuracy in the field of animal husbandry, and there are fewer sheep data sets for research. To solve these problems, the data set of sheep density estimation was established, and a method of grassland sheep number estimation based on multi-scale residual visual information fusion Network (MRVIFNet) was proposed. This method extracts multi-scale features of sheep targets by using multiple parallel hole convolutions with different hole rates, and designs a depth neural network that is more suitable for live counting of sheep, so as to reduce the grid effect caused by hole convolution and better adapt to multi-scale changes of sheep. In the sheep density data set, the method obtained the lowest mean absolute error (MAE) and root mean square error (RMSE). In addition, a convolutional neural network model based on view branch sharing is also studied. Compared with the five popular methods, this method can achieve better performance. It is applied to solve the problem of pedestrian scale change and chaotic distribution in complex scenes; The performance of this method is better than that of comparison method, and the application results in actual scenarios verify the effectiveness of this method.

INDEX TERMS Sheep counting, convolutional neural network, feature reuse, Csrnet network, density estimation, deep learning.

I. INTRODUCTION

As one of the best natural pastures, the grassland is the most widely distributed area in China. However, due to the lack of efficient sheep counting methods in pastoral areas, relevant departments cannot supervise and guide the grazing intensity, resulting in overgrazing. An efficient sheep counting method can also improve the working efficiency of herdsmen in the breeding process. In ranch management, the estimation of sheep quantity is one of the most important tasks in aquaculture management and asset estimation. Accurate sheep counting can help ranch managers evaluate living assets, thereby improving ranch breeding efficiency, and timely detect theft, helping enterprises or individuals reduce unnecessary losses [1]. In the actual sheep farm, it is very challenging to accurately count sheep in the sheep farm due to the mutual occlusion between sheep and the change of illumination. Traditional pasture management usually adopts manual counting method, but this method is very time-consuming, labor-consuming and error prone [2]. At present, the management of pasture sheep mainly uses hardware equipment with electronic ear tags as the core information carrier to count [3], but the hardware equipment has not been applied on a large scale due to the high maintenance cost and complex operation in animal husbandry and breeding; At the same time, wearing ear tags will also bring some physical injuries to livestock, such as wound infection and shock [4], [5], [6], [7].

Vision based target counting has received widespread attention in recent years. Many scenes use visual information to count targets, such as dense crowd counting, crop target counting, cell counting, small target counting, etc. But for animal husbandry, there is still little research on counting method based on vision. At present, there are three kinds
of vision based target counting methods, namely, detection based, regression based and density based. Some scholars use the method of target detection to count livestock. For example, Xu et al. [8] used the images collected by the four-axis aircraft to detect and segment livestock through maskr CNN, and realized the counting of sheep living under the grassland and the livestock farm. Li Qi et al. [9] combined yolov3 target detection algorithm with deep sort target tracking algorithm and realized automatic counting of grassland sheep based on double line counting method. The above methods are only suitable for scenes with relatively sparse targets. In the actual scene of the ranch, due to the different distance between the sheep and the camera, the scale of the sheep changes greatly, and most of the time, the sheep are gathered together, which has serious occlusion. In this case, the counting error of the target detection method is large. Tian et al. [10] estimated the number of pigs in the pigsty by predicting the density of pigs in the image. This method is accurate when the density of pigs is small, but the prediction error is large when the number exceeds 10.

With the continuous improvement of the computing power of image processing unit (GPU) and the continuous development of deep learning, the density map crowd counting method based on deep learning came into being. This method makes full use of computer storage and computing power, greatly improves the accuracy of crowd counting, and improves the real-time performance of crowd counting. To a certain extent, it solves the problem of manually selecting features in traditional crowd counting methods, and expands the scope of applicable scenarios. Zhang et al. [11] extracted multi-dimensional features through three columns of neural networks containing convolution kernels of different sizes, and fused these extracted multi-dimensional features to obtain the final population density map. After this method was proposed, a series of crowd counting methods based on multi column convolutional neural network emerged in endlessly. For example, on the basis of fine-tuning the method of Zhang et al. [13], Sam et al. [12] added a picture crowd density level classifier at the front end, and input the pictures classified into different density levels into different columns to get the corresponding crowd density map. Lu Jingang et al. [14] added jump operation on the basis of fine-tuning the method of Zhang et al. [15], and then fused the feature maps of each column of networks to obtain the final population density map. On the basis of fine-tuning the method of Zhang et al. [16], sindagi et al. Added two columns of networks to obtain global and local context information respectively, so as to strategically fuse the features of multiple columns of networks and obtain the final population density map. The method based on multi column convolutional neural network described above has improved the accuracy of crowd counting to a certain extent, but Li et al. [17] proposed the disadvantage of multi column network in the condensed scene recognition network (csrcnet): Although the multi column network uses convolution kernels of different sizes to extract multi-dimensional features, the resulting feature maps are almost the same, and there is a large redundancy. And the network is complex and the training is time-consuming. Csnet, which also includes dilation convolution, has a “chessboard effect” because it uses the same dilation rate, that is, it cannot make full use of the feature map obtained by the front-end network.

In this paper, the convolution kernel is convoluted with the image through a convolution neural network. Through a series of convolution kernels, the features of the image are continuously extracted. Finally, the extracted high-level features are classified to generate a density map, and then the density map is summed to count the total number of people. However, it only makes simple operation of these features and cannot make good use of them. Therefore, this paper proposes a new method for counting dense crowds based on a new multi-scale attention mechanism. Its network structure is divided into backbone network, feature extraction network and feature fusion network. The feature extraction network is divided into two branches: feature branch and attention branch. Considering the diversity of data scale features, a new multi-scale module is added to both branches of this paper, and a res structure is added to the feature branch separately, so as to better obtain the population characteristics at different scales. The attention branch is used to constantly strengthen the head features in the dense crowd image, so that the density map of the head area is more obvious. In the feature fusion network, through the attention fusion module, attention features and image features are effectively fused to further improve the counting accuracy. Experiments on public data sets (Shanghai tech, ucf_cc_50, mall, UCSD) have obtained better parameter indicators than other methods.

II. RELATED WORK

A. POPULATION DENSITY ESTIMATION AND COUNTING OF MULTIVARIATE INFORMATION AGGREGATION

The network structure of efficient multi semantic spatial information aggregation and sheep density estimation is shown in Fig. 1, the multi-scale context information Ag aggregation part and the density map regression part. The multi information extraction part is used to optimize the high-level and low-level features; The multi-scale context information aggregation part is used to enrich the receptive field and realize the effective fusion of high-level and low-level features.

Generally, low-level features contain rich spatial details but lack semantic information, and high-level features contain rich semantic information but lack spatial information. Direct fusion of high-level and low-level features is a common means to obtain comprehensive expression of spatial and semantic information. However, this direct fusion method ignores the differences in semantic level and spatial level between high-level and low-level features, resulting in poor fusion effect and low feature utilization. Literature [18] proved that the feature fusion effect can be effectively enhanced by introducing more semantic information into low-level features or embedding more spatial information
into high-level features. Based on this, this paper proposes an efficient multi-semantic feature extraction network as shown in Figure 2, which is used to strengthen high-level and low-level features and lay the foundation for subsequent feature fusion. It mainly includes three parts: skeleton network vgg-19, multi-semantic supervision (MSS) and spatial embedding (SE). Vgg-16 network architecture provides a more suitable receptive field for most object detection and is easy to capture detailed information. It is one of the commonly used skeleton networks for sheep density estimation. However, for high-density images, the ability of vgg-16 to mine detailed features is slightly insufficient. Therefore, this paper selects vgg-19 network with similar network structure to vgg-16 but deeper network layers to obtain better initial characteristics. Experiments show that removing the full connection layer of vgg-19 network has little impact on the accuracy of sheep counting, and can effectively reduce network parameters. Therefore, this paper adopts vgg-19 with the full connection layer removed as the skeleton network to alleviate network redundancy while acquiring deep features. The low-level features contain more location and detail information, but have low semantics and more noise. In view of this problem, this paper proposes a multi-layer semantic supervision strategy MSS as shown in Figure 2 to deal with low-level features. Three semantic supervision modules (SS) are designed to be attached to the second, fourth and sixth layers of the vgg-19 backbone network to optimize the low-level features of the backbone network.

### B. MRVIFNet

The network structure of the crowd counting model MRVIFNet proposed in this paper, which integrates channels and spatial attention, is shown in Fig. 3. The whole adopts an easy end-to-end training codec architecture. Among them, the encoder uses the first 13 layers of vgg16 [1] network as the backbone, constructs a feature extraction network, and takes multiple semantic features at different depth levels to identify multi-scale people in the scene; While gradually recovering the spatial information, the decoder fully integrates the multi-scale information with the spatial context information to enhance the network representation ability.

Open and integrate the channel and spatial attention module, focus on the foreground crowd area, suppress weak correlation background features, and generate high-quality and high-resolution density map for crowd counting.

The codec includes two parts, in which the coder can extract pedestrian features of different scales. In order to extract multi-level and more representative depth features, and to facilitate network construction and training, this part selects the first 13 layers of vgg16 network that have been pre-trained as the backbone network of the encoder. These features extracted at different depths can capture pedestrian information at different scales. As the network depth increases, the resolution of the feature map gradually decreases and the dimension gradually increases. The decoder is mainly used to gradually recover the image...
spatial feature information and focus on the foreground crowd area. The multi-level depth features recovered by decoding are fused with the corresponding layer features output by the encoder at each stage to minimize the feature loss caused by convolution, downsampling and other operations, and further integrate spatial context information. After fusion, channels and spatial attention are added to the features to highlight the foreground crowd area and suppress the weight of the features in the weak correlation background area. The decoder’s fusion of feature maps at different stages mainly involves channel splicing of two feature maps. After feature fusion, the resolution of the new feature map remains unchanged, and the channel is the sum of the two. The network parameter configuration is shown in Table 1.

### C. SHARED FEATURE EXTRACTION MODULE

The multi view crowd counting method [19] often sets independent but isomorphic feature extraction modules for pictures from different perspectives, which not only increases the overall complexity of the model, but also introduces too many redundant features. However, the pictures from different perspectives are not independent of each other, but closely related. It is assumed that the cameras in the scene are fixed and have been registered [20], and the video frames from each camera have been synchronized. Multiple views as the same group of inputs are only projections of the same scene at the same time on different camera poses. Therefore, vbs-cnn makes pictures from different perspectives share the same feature extraction module, and improves feature utilization through tight coupling between different perspectives. As shown in Figure 4, inspired by sasnet [18], this paper designs a feature extraction module with certain structural complexity. In this module, different middle layers of the backbone network are used to extract features of multiple scales of the input picture. In this paper, three scales are used to represent the people far, medium and close to the camera. The downsampling span between different scales is 2. The backbone network may be decomposed into a plurality of coding blocks and decoding blocks. The input picture is first encoded by coding blocks of different scales to form an internal representation \{V1, V2, V3\}, and then decoded by the corresponding decoding block to form a feature \{V1, V2, V3\} containing information of respective scales. The internal structures of the encoding block and the decoding block are shown in Fig. 4. Figure 4 shows the number of channels of the output characteristic diagram at each stage. This feature extraction module is different from sasnet [21].

### TABLE 1. Network parameters.

| Encoder          | Decoder          |
|------------------|------------------|
| Conv1_1(3-64-1)  | Upsampling       |
| Conv1-2(3-64-1)  | Concat           |
| max-pooling      | CBAM Module      |
| Conv2-1(3-128-1) | Conv6-1(1-256-1) |
| Conv2-2(3-128-1) | Conv6-2(3-256-1) |
| Max-pooling      | Upsampling       |
| Conv3-1(3-256-1) | Concat           |
| Conv3-2(3-256-1) | CBAM Module      |
| Conv3-3(3-256-1) | Conv7-1(1-128-1) |
| Max-pooling      | Conv7-2(3-128-1) |
| Conv4-1(3-512-1) | Upsampling       |
| Conv4-2(3-512-1) | Concat           |
| Conv4-3(3-512-1) | CBAM Module      |
| Max-pooling      | Conv8-1(1-64-1)  |
| Conv5-1(3-512-1) | Conv8-2(3-64-1)  |
| Conv5-2(3-512-1) | Upsampling       |
| Conv5-3(3-512-1) | Conv9-1(3-32-1)  |
|                   | Conv9-2(3-32-1)  |
|                   | Conv10-1(1-1-1)  |
in two main points: 1) it emphasizes that coding blocks and decoding blocks exist as independent modules in the pipeline, and different modules are connected by input and output. 2) Sasnet [18] uses the first several convolution layers and pooling layers of vgg16 as the backbone network, and this module uses fewer convolution cores, which greatly reduces the number of learnable parameters of the whole model.

III. NETWORK IMPROVEMENT

A. FEATURE EXTRACTION NETWORK AND MODULE

In the network model proposed in this paper, the feature extraction network adopts a new attention mechanism. It includes two branches: feature branch and attention branch. The feature branch is used to extract the crowd distribution features in the image; The attention branch is used to accurately estimate the head position, modify the obtained crowd density map, and obtain a high-quality crowd density estimation. The feature branch includes the basic feature extraction module (as shown in Fig. 5), the new multi-scale module and the auxiliary structure. The basic feature extraction module is mainly used to restore low-resolution features to high-resolution features, and provide richer spatial distribution information for density map estimation of dense population counting. Attention branch includes basic attention module and new multi-scale module. In this paper, the structure of the basic attention module is the same as that of the basic feature module, and its role is to restore low resolution features to high-resolution features, which is conducive to the accurate positioning of the head position. For the feature extraction network, a new multi-scale module is proposed to improve the output characteristics of the two branches and improve the computational efficiency. With the increasing depth of neural network, the volume of network parameters becomes larger and larger, and the weights of a large number of parameters tend to zero, which leads to high redundancy and waste of computing resources. One way to solve this problem is to introduce a sparse filter. Thus Szegedy proposed the initiation structure. The classic inception is composed of 1 × 1, 3 × 3, 5 × 5 convolution layers and one pooling layer (as shown in Fig. 6). The size of convolution kernel directly determines the perceptual ability to different targets. In this paper, considering the change range of the size of people in the dense crowd image, in order to extract the large-scale crowd features in the image, we replace the pooling layer in the incidence structure with 7 × 7 convolution layer (as shown in Fig. 7). At the same time, considering that in order to improve the network computing efficiency, we further add the above 5 × 5 convolution layers, replaced by 2 cascaded 3 × 3 convolution, 7 × 7 replace the convolution layer with 3 cascaded 3 × 3. Convolution. The receptive field range will not change before and after replacement [17]. Therefore, we propose a new multi-scale module (as shown in Fig. 8). The new multi-scale module enhances the concentration of the population density features in the feature branch, further expands the receptive field, and makes the mapping area of the pixel points on the output feature map of each layer on the input picture larger. At the same time, the new multi-scale module can also enhance the head position information in the attention branch.

The function of the feature fusion module is to apply the output features of the attention branch to the output features of the feature branch, and achieve the fusion of the two through multiplication to obtain a higher quality population density map. Among them, the attention fusion module plays a key role, and its structure is shown in Figure 5.

Among the high-level features, the rich and abstract feature information puts forward higher requirements for the feature recognition ability of the network. In the attention fusion module, through matrix transformation and its combination, the feature dimension or element position changes, that is, the channel information changes, so as to achieve feature reorganization. These recombined features can further enrich the feature description of the dense population density map and improve the network recognition ability. In this paper, attention is defined as:

$$X \in R^{C \times C} : \quad x_{ji} = \frac{\exp(A_i \cdot A_j)}{\sum_{i=1}^{C} \exp(A_i \cdot A_j)} \quad (1)$$

where, the number of C channels; A represents the characteristic value of the density map; Measure the influence of the i-th channel on the jth channel. The output characteristics are:

$$E \in R^{C \times H \times W} : \quad E_j = \beta \sum_{i=1}^{C} (x_{ji}A_i) + A_j \quad (2)$$

wherein, $\beta$ Represents the scale coefficient, initialized to 0, and gradually learns to allocate to a larger weight. The resulting feature e of each channel is the weighted sum of all channel features and original features.

B. LOSS FUNCTION

The loss function consists of two parts. First, the Euclidean distance LC is used to measure the loss of pixel level between
FIGURE 5. Network structure proposed in this paper.

FIGURE 6. Basic feature extraction module, which is also used as basic attention module in this paper.

FIGURE 7. Traditional concept structure.

the predicted density map and the real density map.

\[ L_C = \frac{1}{N} \sum_{i=0}^{N} ||\hat{D}_i - D_i||^2_2 \]  

(3)

where \( n \) is the number of pixels in the scene density map, \( \hat{D}_i \) shows the \( i \)th pixel in the predicted density map, and \( D_i \) represents the \( i \)th pixel in the real density map. Euclidean distance can measure the difference between independent pixels in two pictures, but it ignores the correlation between pixels. Therefore, the local consistency loss \( L_C \) is added to the loss function.

\[ L_E = 1 - \frac{1}{N} \sum_{i=0}^{N} SSIM(\hat{D}_i, D_i) \]  

(4)

The structure similarity index \( SSIM \) [23] can be used to measure the local consistency between two positions \( X \) and \( Y \) in a picture. The calculation formula is:

\[ SSIM(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \]  

(5)

\( \mu_x, \mu_y \) are the mean values of local regions centered on \( x \) and \( y \) respectively, \( \sigma_x^2, \sigma_y^2 \) are the variances of the local regions centered on \( x \) and \( y \), respectively, and \( \sigma_{xy} \) are the corresponding covariances. \( C_1 \) and \( C_2 \) are two constants, which are used to control the stability of division calculation. Follow the settings in [24]. In this paper, we define a pixel position as the center and a size of \( 1 \times 1 \). The area of 1 is the local area of interest, and the normalized Gaussian kernel with a standard deviation of 1.5 is used to calculate the mean and variance in this area. The final loss function is weighted by \( L_C \) and \( L_E \)

\[ \text{loss} = L_C + \lambda L_E \]  

(6)
and every pixel in the receptive field is used; When the cavity rate is 3, the receptive field is $7 \times 7$. As shown in FIG. 10 (c), at this time, there are still 9 pixels used, and the utilization rate drops to 18%.

Using vgg-16 for feature extraction makes the resolution of the feature map gradually decline, and for pixel intensive tasks, too small a resolution will lose a lot of detail information. Therefore, this paper uses the first ten layers of vgg-16 convolution and takes the result after the third pooling operation as the back-end input. Using hole convolution can obtain a larger receptive field, but when stacking multiple identical hole convolutions, some pixels will always be ignored, that is, some information will be completely lost, which is also very unfavorable for pixel intensive prediction tasks. Therefore, in order to make up for this part of completely lost pixels, this paper proposes a multi-scale residual feature perception module (MSR). The module structure is shown in Fig.11. This module is composed of three hole convolutions with different hole rates and two ordinary convolutions. First, through a $3 \times 3$ convolution reduces the channel dimension and reduces the subsequent calculation. Then, three parallel hole convolutions are used to extract multi-scale features, with hole rates of 1, 2 and 3 respectively. The convolution with a hole rate of 1 can capture every detail of the image, and the convolution with a hole rate of 2 and 3 can obtain a larger receptive field and obtain multi-scale features. Then, by means of identity mapping, the extracted three multi-scale features are added to the input to form a residual structure, so that the pixel information loss caused by the hole is supplemented and the fusion features are obtained. After each hole convolution, the linear rectifying unit (relu) activation layer is connected, and finally the fused features are passed through a $3 \times 3 \times 512$ convolution for further fusion.

C. RESIDUAL FEATURE PERCEPTION OF MULTISCALE HOLE CONVOLUTION

The distribution of sheep in the pasture is usually uneven, so the density of sheep in different areas on the image is different. At the same time, due to the camera’s shooting angle, in reality, sheep with the same body size are closer to the camera and farther from the camera, so they need different sizes of receptive fields to capture features of different scales. However, increasing the convolution kernel will lead to an increase in the amount of network parameters, and the application of the intelligent counting function in the ranch usually requires running the program on the edge device. The memory of the edge device is limited, and the model with large parameters is not applicable. Therefore, this paper uses hollow convolution to solve the problem of the increase in network parameters.

Hole convolution was proposed by Yu et al. [25] to solve the problem of image segmentation. Different from ordinary convolution, void convolution introduces a void rate hyper parameter. By controlling the number of holes inserted in the convolution kernel, this parameter increases the receptive field while maintaining the image resolution and not increasing the parameter amount. However, hole convolution also has disadvantages, which will lead to grid effect. Because hole convolution expands the receptive field by inserting holes in ordinary convolution, the convolution kernel is discontinuous, and its sampling features are discrete and lack of correlation. The larger the hole rate, the more features will be ignored. As shown in Fig. 10 (a), when the void rate is 1, the receptive field is $3 \times 3$. At this time, it is normal convolution, and every pixel in the receptive field is used; When the cavity rate is 2, the receptive field is $5 \times 5$. As shown in Fig. 10 (b), only 9 pixels are used, with a utilization rate of 36%; When the cavity rate is 3, the receptive field is $7 \times 7$. As shown in FIG. 10 (c), at this time, there are still 9 pixels used, and the utilization rate drops to 18%.

Using vgg-16 for feature extraction makes the resolution of the feature map gradually decline, and for pixel intensive tasks, too small a resolution will lose a lot of detail information. Therefore, this paper uses the first ten layers of vgg-16 convolution and takes the result after the third pooling operation as the back-end input. Using hole convolution can obtain a larger receptive field, but when stacking multiple identical hole convolutions, some pixels will always be ignored, that is, some information will be completely lost, which is also very unfavorable for pixel intensive prediction tasks. Therefore, in order to make up for this part of completely lost pixels, this paper proposes a multi-scale residual feature perception module (MSR). The module structure is shown in Fig.11. This module is composed of three hole convolutions with different hole rates and two ordinary convolutions. First, through a $3 \times 3$ convolution reduces the channel dimension and reduces the subsequent calculation. Then, three parallel hole convolutions are used to extract multi-scale features, with hole rates of 1, 2 and 3 respectively. The convolution with a hole rate of 1 can capture every detail of the image, and the convolution with a hole rate of 2 and 3 can obtain a larger receptive field and obtain multi-scale features. Then, by means of identity mapping, the extracted three multi-scale features are added to the input to form a residual structure, so that the pixel information loss caused by the hole is supplemented and the fusion features are obtained. After each hole convolution, the linear rectifying unit (relu) activation layer is connected, and finally the fused features are passed through a $3 \times 3 \times 512$ convolution for further fusion.

D. GLOBAL MULTI-SCALE CONTEXT INFORMATION AGGREGATION NETWORK

Capturing rich contextual relationships helps the network understand complex scenes, and is an effective way to alleviate the problems of variable perspectives, occlusion and scale transformation in crowd density estimation. However, in high-density crowd images, with the increase of the number of targets, the correlation between images increases dramatically. While improving the regression accuracy, context information inevitably increases the amount of model calculation, which restricts the actual application deployment and landing of the algorithm. Therefore, how to efficiently aggregate multi-scale context information is the core of high-density population density estimation. This paper proposes a global multi-scale context information aggregation network as shown in FIG. 13. Through two lightweight hole spatial pyramid pooling (s-aspp) modules, the context information of different scales of low-level and high-level features is gradually captured and merged, so as to enhance the global expression of features on the premise of ensuring limited computing costs. For the convenience of expression, the two
s-aspp modules are marked as s1-aspp and s2-aspp respectively. S-aspp module document [25] proposes the idea of hole spatial pyramid pooling (ASPP), which samples the feature map in parallel through hole convolution at different sampling rates, expands the receptive field, and obtains context information at different scales. However, ASPP has a large number of channels in the feature mapping process, resulting in high computation and model redundancy. Based on this, this paper designs a lightweight hole space pyramid pooled s-aspp [26] structure, taking s1-aspp as an example: first, through four point-by-point convolution layers with a core size of 1, the channel dimension of the high-level features obtained from the multivariate information extraction network is reduced, and the channel information interaction is performed; Secondly, the concept like structure is adopted, and the step convolution is adopted (to reduce the redundancy of the model, enrich the receptive field of the feature map with the expansion rate of 1, 6, 12, and 18, and capture more context information); finally, the fusion operation is performed on the processed feature map to enhance the global expression of features.

IV. EXPERIMENT AND ANALYSIS

A. DATA COLLECTION AND EXPERIMENTAL ANALYSIS

Shanghaitech [5] is a large sheep counting data set, with a total of 1198 images and 330165 heads. According to the data source and scene sparsity, it can be divided into part_A and part_B these two parts, part_A [27] was randomly collected from the Internet, and sheep were densely distributed. There were 300 images as the training set and 182 images as the test set; And part_B [28] according to some surveillance videos collected from Shanghai, sheep are sparsely distributed, with 400 images as the training set and 316 images as the test set. The experimental results of this dataset are shown in Table 2.

Compared with the existing algorithms, MRVIFNet in part_A means that both AE and RMSE have reached the optimal value, while in part_B, the performance is second only to sfanet [29]. The change trend of the total loss is shown in FIG. 14. At the beginning of training, the loss is large due to the high degree of randomness. However, with the continuous iterative training of the model, the loss shows an obvious downward trend and tends to be stable; Part_A fluctuates within the overall controllable range. Part_B is basically stable after 400 iterations.

Ucf-qnrf [30] is a very challenging data set with rich scenes and disorderly population distribution. A total of 1535 images were labeled, including 1201 images in the training set and 334 images in the test set. The total number of labeled images reached 1251642. Table 3 shows the experimental results of various population counting algorithms on the ucf-qnrf dataset. It can be seen from the table that the two performance indicators MAE and RMSE of MRVIFNet network are optimal, which proves that MRVIFNet model has good performance in cross scene counting. The training loss curve of MRVIFNet is shown in Figure 15. The fluctuation of the
first 500 iterations is large, and it gradually tends to be stable after 500 iterations.

The experiment compares the performance of the SMCAN network and the mainstream counting model. Table 4 shows the experimental results of various models on Shanghai Tech dataset. It can be seen from the table that from the quantitative data and comparison of the experimental results of SMCAN and classical crowd counting methods, compared with the use of similarity measurement loss function and attention module alone, SMCAN has achieved better results in Shanghai Tech dataset, and its MAE and RMSE of Part A are both better than CFF model. The experimental data on NWPU also fully illustrates the excellent prediction effect of SMCAN.

Figure 16 shows the test set, ground live density map and estimated density map, estimated population and real population for Part A and Part B including test images. It can be seen from the ground real density map and estimated density map, whether it is dense scenes or relatively sparse scenes, and the estimated number of people is very close to the real number of people.

TABLE 3. The performance comparison of different methods on UCF_QNRF.

| Method    | Part A | Part B |
|-----------|--------|--------|
| MCNN[5]   | 277    | 426    |
| CMTL[30]  | 252    | 514    |
| IA-DCCN[31] | 125    | 186    |
| RANet[32] | 111    | 190    |
| DUBNet[28] | 105.6  | 180.5  |
| MRVIFNet  | 104.8  | 175.4  |

B. MRVIFNET PERFORMANCE EVALUATION IN SHEEP DATASET

The counting results of target detection methods yolov4 and yolov5 [31], density estimation methods MCNN [32], csrnet [33], dsnet [34] and kdmg [13] are compared on the sheep data set. The target detection method detects the target on the image and counts the number of detected target frames. Mae and RMSE indexes were used to evaluate the accuracy of counting and the robustness of the model. On the sheep data set, the evaluation results of Mae and RMSE are shown in Table 5. It can be seen that compared with other methods, the Mae and RMSE of MRVIFNet in this paper are the lowest. Compared with the suboptimal method dsnet, the MAE decreases by about 34.6% and the RMSE decreases by about 43.9%. From the experimental results of different methods, it can be observed that yolov4 and yolov5 target detection methods count by counting the detected sheep target frames. Due to the serious overlap of sheep targets and the large number of small targets, the method of target detection in this scenario will produce a lot of missed detection; Compared with target detection methods, methods based on density estimation such as csrnet, net and kdmg have better counting effect, and these algorithms are more suitable for the situation of serious occlusion of sheep data; MCNN [35] learns features of different sizes through three different convolution channels, but due to the limitation of branching structure, the information in each column is not shared, and the actual effect also proves that this method does not adapt well to the multi-scale changes of sheep; Csrnet uses hole convolution at the back end of the network to keep the resolution of features unchanged and extract deeper features. This method can extract more detailed features, but it still cannot adapt to the problem of large scale changes; Dsnet uses closely connected hole convolution to maintain the continuous transmission of information and the extraction of multi-scale features, but too dense connections easily lead to data overfitting. Because the proportion of training data in the order of 50 ∼ 100 in the training data is very small, the prediction accuracy of the network for this part of data drops linearly; Kdmg method uses the results of network prediction to generate supervision information, but uses less constraints, resulting in the generated supervision information is not particularly suitable for sheep data sets; The MRVIFNet in this paper superimposes multiple multi-scale residual feature perception modules, extracts multi-scale features by using
hole convolution with three different hole rates, adapts to the scale change of sheep, and transmits the shallow feature information to the deep layer by using residual connection, which weakens the grid effect brought by hole convolution and makes up for the lost pixel feature information. Therefore, it can achieve the best performance in Mae and RMSE values. This experiment proves that MRVIFNet can accurately count the number of sheep in the sheep farm, and the overall average error is about 1.7.

V. CONCLUSION
Aiming at the problem of uneven distribution of sheep and large scale change, this paper proposes a multi-scale residual sheep density estimation network, which uses hole convolution with different hole rates to obtain multi-scale features, and then adds residual structure to weaken the grid effect caused by the use of hole convolution. Experiments show that this method has better effect, more accurate counting and better robustness than the traditional crowd density estimation method. And for the csnet back-end network, this paper conducts a comparative experiment between the expanded convolution structure and the ordinary convolution structure. It is found that although the expanded convolution has a slightly stronger feature extraction ability than the ordinary convolution, the training is more time-consuming. In this algorithm, the common convolutional network structure is used to shorten the time required for training the model; Adopt deeper network to enhance the feature extraction ability of the network; The dense connection structure is added in the network to enhance the transmission of feature and gradient information in the network. The new multi-scale module introduced can make the spatial concentration of deep-seated features higher, expand the receptive field, and obtain a higher quality population density map. In the deep features, the attention fusion module is used to improve the feature discrimination, so as to improve the performance of the network. The experimental results prove the effectiveness of this method. In future work, we will consider how to use deformable convolution and other facets to more accurately focus on the sheep area, so as to further improve the accuracy of crowd counting.

REFERENCES
[1] Y. Xuguang, Z. Peng, and L. Dan, “Detection and counting of nursery seedlings based on YOLOv3 neural network,” J. Forestry Eng., vol. 7, no. 3, pp. 174–179, 2022, doi: 10.13360/j.issn.2096-1359.202105018.
[2] L. Shushu, “Research and application of population counting method based on convolutional neural network,” Anhui Univ. Archit., Hefei, China, Tech. Rep. 4, 2022, doi: 10.27784/d.cnki.gahjz.2022.000323.
[3] S. Mingkuan, C. Hongbin, Z. Jinli, and C. Yong, “Construction of artificial neural network model based on four conventional infection related biomarkers for early prediction of bloodstream infection,” Mod. Immunol., vol. 42, no. 2, pp. 97–103, 2022.
[4] S. D. Khan, Y. Salih, B. Zafar, and A. Noorwali, “A deep-fusion network for crowd counting in high-density crowded scenes,” Int. J. Comput. Intell. Syst., vol. 14, no. 1, pp. 1–12, Dec. 2021.
[5] S. D. Khan and S. Basalamah, “Sparse to dense scale prediction for crowd counting in high density crowds,” Arabian J. Sci. Eng., vol. 46, no. 4, pp. 3051–3065, Apr. 2021.
[6] S. Basalamah, S. D. Khan, and H. Ullah, “Scale driven convolutional neural network model for people counting and localization in crowd scenes,” IEEE Access, vol. 7, pp. 71576–71584, 2019.
[7] S. D. Khan and S. Basalamah, “Scale and density invariant head detection deep model for crowd counting in pedestrian crowds,” Vis. Comput., vol. 37, no. 8, pp. 2127–2137, Aug. 2021.
[8] G. Huaping, W. Rui, W. Jing, S. Yange, L. Jian, and L. Meng, “Dual convolutional neural network for population counting,” J. Xinyang Normal Univ., vol. 34, no. 4, pp. 650–654, 2021.
[9] D. Yongxing, M. Xiaowei, Q. Ling, and L. Baoshan, “Sheep counting based on VDNet convolutional neural network,” Laser Technol., vol. 45, no. 5, pp. 675–680, 2021.

[10] X. Feiyu and Z. Xiwei, “Research on population counting algorithm based on convolutional neural network,” Comput. Technol. Develop., vol. 31, no. 7, pp. 42–46, 2021.

[11] C. Lu, “Video dense group behavior recognition based on convolutional neural network,” Nanjing Univ. Posts Telecommun., Nanjing, China, Tech. Rep. 6, 2020.

[12] F. Songwei, L. Xiangyu, and Z. Ping, “Study on automatic counting of shrimp fry based on improved convolutional neural network,” Fishery Modernization, vol. 47, no. 6, pp. 35–41, 2020.

[13] W. Qyuan, W. Xiaodong, and Z. Lianjun, “Crowd counting technology based on single column feature fusion convolutional neural network,” Wireless Commun. Technol., vol. 29, no. 4, pp. 30–34, 2020.

[14] C. X. min, Y. S. Han, and Z. Rui, “Population counting model based on multi-task learning and convolutional neural network from coarse to fine,” Comput. Sci., vol. 47, no. S2, pp. 183–187 and 208, 2020.

[15] B. Wenzia, Z. Xin, H. Gensheng, L. Linsheng, L. Dong, and L. Ze, “Estimation and counting of wheat ear density in field based on deep convolution neural network,” J. Agricult. Eng., vol. 36, no. 21, pp. 186–193 and 323, 2020.

[16] H. Jiehao, L. Xiaoguang, W. Junde, and C. Xiaoyue, “A new population counting method based on convolutional neural network,” Frontiers Inf. Technol. Electron. Eng., vol. 21, no. 8, pp. 1150–1161, 2020.

[17] L. Jingang and Z. Li, “Dense population counting model based on multi-layer BP neural network,” Comput. Sci., vol. 47, no. S2, pp. 183–187 and 208, 2020.

[18] 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, Feb. 2022.

[19] C. Siqin, “Population counting based on full convolutional neural network,” Electron. Des. Eng., vol. 26, no. 2, pp. 75–79, 2018.

JIANMING XU was born in Wugang, Hunan, China, in 1977. He received the master’s degree from Hunan University, China. Now, he works at Shaoyang University. His research interests include information processing and algorithm research.

WEICHUN LIU received the M.S. degree in communication and information system from Zhengzhou University, in 2007, and the Ph.D. degree in information and communication engineering from the College of Electronic Science, NUDT, in 2020. Since Fall 2020, he has been an Docent with Shaoyang University. His research interests include computer vision, machine learning, and computer graphics.

YANG QIN was born in Huanggang, Hubei, China, in 1997. He is currently pursuing the degree with the School of Mechanical and Energy Engineering, Shaoyang University. His research interests include machine vision and algorithm research.

GUANRONG XU was born in Wugang, Hunan, China, in 1997. He is currently pursuing the degree with the School of Mechanical and Energy Engineering, Shaoyang University. His research interests include machine vision and algorithm research.