MSR-FAN: Multi-scale residual feature-aware network for crowd counting

Haoyu Zhao1 | Weidong Min2,3 | Xin Wei3 | Qi Wang1 | Qiyan Fu1 | Zitai Wei1

1 School of Information Engineering, Nanchang University, Nanchang 330031, China
2 School of Software, Nanchang University, Nanchang 330047, China
3 Jiangxi Key Laboratory of Smart City, Nanchang 330047, China

Correspondence
Weidong Min, School of Software, Nanchang University, Nanchang 330047, China; Jiangxi Key Laboratory of Smart City, Nanchang 330047, China.
Email: minweidong@ncu.edu.cn

Funding information:
This work was supported by the National Natural Science Foundation of China (Grant Nos. 62076117 and 61762061), the Natural Science Foundation of Jiangxi Province, China (Grant No. 20161ACB20004) and Jiangxi Key Laboratory of Smart City (Grant No. 20192BCD40002).

1 INTRODUCTION
Crowd counting is an important task for many applications such as security systems and control systems. It aims to count how many people are there in a certain area. It can be described as the number of people or the density of an area [1–3]. But with the perspective changes problem and the scene scale variation, it is difficult to count the crowd number. It also is an emerging topic with the development of the deep learning and the computer vision. And there have been some excellent works to find a way to get better accuracy. So, this domain has its practical significance and hence attracted wide attention of people.

The methods to solve the crowd counting task are divided into two categories: the detection-based methods [4] and the map-based methods. The detection-based methods use detection algorithms to find all the people in the image and count them. The benefit is that it can locate the person. But for very dense scenes, such methods [5, 6] always tend to fail. The map-based methods extract image features and do regression operation to generate the density map and get the number of people. Such approaches [7, 8] can well overcome the crowd occlusion...
problem. However, the generalization ability of these models is poor. And some excellent works also try to solve the real street scenes, such as [9]. These methods have also driven the development of real-world data-driven crowd counting domain.

To alleviate the above problems, a multi-scale residual feature-aware network is proposed. This method combines the local and global feature using multiple receptive field sizes and learns the feature aware information on each image. It is an end-to-end approach that can generate high-quality density map and count the crowd number. Specially, the initial image is sent into the convolution layers of VGG to extract the image feature. To handle the perspective changes problem and enhance the initial image feature, the direction-based feature-enhanced network is employed. In the high-density crowd environment, the occlusion between people is relatively serious. The single-pixel feature cannot provide a reliable estimation of the number of people. So, this work proposes multi-scale residual block that encodes different scales information and reinforces the initial image feature. In addition, the feature-aware block is also used to learn the importance of the regional feature. The experiments based on benchmark datasets show that the proposed approach outperforms the existing methods.

The main contributions of this study are summarized as follows:

* This study proposes the MSR-FAN for crowd counting, which outperforms the existing state-of-the-art methods.
* The multi-scale residual block is proposed to detect the image features at different scales adaptively and encode multi-level contextual information into the features.
* The combination of the direction-based feature-enhanced network and the feature-aware block is used to solve the perspective changes problem and extract the difference between local features and image content for crowd counting.

The rest of this paper is organized as follows. Section 2 discusses related works. Section 3 introduces the whole method and the details of the proposed MSR-FAN for crowd counting. Section 4 shows the experiment results, and Section 5 gives the conclusions and future research directions.

## 2 RELATED WORK

Crowd counting is a rising research topic. Many excellent methods have been proposed for this difficult task. Generally, detect crowd counting approaches can be divided into two main categories: one is detection-based and the other is map-based.

### 2.1 Detection-based methods for crowd counting

In the early period, the object detection methods are used for crowd counting. The researchers used the detection-based method to detect all the people in the image. It needs the people in the image can be clearly detected. Wang et al. [10] proposed a new framework of adapting a pre-trained generic pedestrian detector to traffic scene. It can automatically select both confident positive and negative examples from the target scene. Zhao et al. [11] proposed a model-based approach that combined the Bayesian framework and a joint image likelihood for multiple humans based on the appearance of the humans. The visibility of the body is obtained by occlusion reasoning, and foreground separation. In addition, many people developed multi-class object detection methods to solve it. Barinova et al. [12] proposed a new approach with Hough transform for object detection. [13–16] introduced models for multi-class object detection that cast the problem as a structured prediction task.

### 2.2 Map-based methods for crowd counting

With the development of deep learning methods [17, 18], and the crowd scenes become more crowded. Even the person in the image cannot be detected by object detection methods. So, researchers pay attention to map-based method, which can handle large scale and dense crowd scenes. Wu et al. [19] proposed a featured channel enhancement block-based method, which can enhance the positive characteristic channel. Xiong et al. [20] proposed a spatial divided-and-conquer network. They thought a dense region can always be divided until sub-region counts are within the previously observed closed set. Liu et al. [21] introduced a structured feature enhancement module based on conditional random fields to refine features mutually with a message passing mechanism. [22–28] proposed data-driven and adaptive methods that can understand highly congested scenes and perform well. Other works [29–33] used the multi-scale networks for image crowd counting, which are both accurate and cost-effective for practical applications. Cheng et al. [34] proposed a novel architecture called spatial awareness network to incorporate spatial context for crowd counting. Liu et al. [35] introduced an end-to-end trainable deep architecture that combines features obtained using multiple receptive field sizes and learns the importance of each feature. Besides these methods, some researchers also constructed some novel model and proposed new challenge for this task. Such as [36] first constructed a lightweight model, which is composed of an image feature encoder and a simple but effective decoder. Wang et al. [37] proposed a neuron linear transformation method for cross-domain crowd counting. It can transfer the source model to the target model.

Although with all the above methods can be applied to crowd counting, the perspective changes and the person scale variation are still the problems that need to be solved and better works should be proposed.

## 3 THE PROPOSED MSR-FAN FOR CROWD COUNTING

The goal of this work is to estimate the density map and count the crowd number by using a novel framework named
The proposed method mainly contains three submodules: the direction-based feature-enhanced network, the multi-scale residual block, and the feature-aware block. The general framework of the proposed method is shown in Figure 1, which can generate a high-quality crowd density map.

The initial image is sent into convolution layers to create the initial feature map. Such feature map contains the initial detailed information about the image. The direction-based feature-enhanced network is proposed to encode the perspective changes in four directions. This operation can enhance the initial image feature and effectively improve the quality of density maps for the densely crowded areas. The proposed multi-scale residual block can combine the multi-scale feature maps and fuse multiple channel information. The $3 \times 3$, $5 \times 5$ and $7 \times 7$ kernel size are used in this work. They are three normal sizes in many other convolutional neural networks, which have been proved they have good performance. In addition, the feature-aware block is employed to capture the differences between the features at a specific location and those in the neighbourhood. The whole approach is an end-to-end training process. The input crowd image can be converted to the density map directly and get the number of crowds.

3.1 The direction-based feature-enhanced network

To get the initial image feature, the image is sent into the first ten layers of a pre-trained VGG network. As Equation (1) shows, the image $I$ is dealt with $F_{VGG}$ to get the initial feature map $f_D$. The images are from nature crowd scenes and the most of the crowd areas are activated by the convolution operation, as the Figure 2 shows. It means that the CNN can catch the visual cues about the crowd, which will help count the number of the crowd. In fact, many other works also use such structure to get...
the image feature. But the feature got by VGG has limitation, [22] discussed that it just encodes the same receptive field over the entire image. So, the feature \( f_D \) needs to be processed further.

\[
f_D = F_{VGG} (I)
\]

To handle such limitation, this work also used direction-based feature-enhanced network to enhance the initial image feature, which is shown as Figure 3. The direction-based module is proposed in [31] to solve the perspective changes problem for crowd counting, which also is a useful tool. The network adopts \( f_D \) as input and contains four submodules: the down layer, the up layer, the right layer, and the left layer. Every submodule is a convolution layer with an ReLU activation function. The goal of such structure is to encode the perspective information in four directions.

The reason for the emergence of this network is that the convolution of each image pixel line should be different due to the perspective changes problem. However, the traditional convolution operation in each image part is the same, which cannot fully represent the local and global aggregation effect. Moreover, this is actually a set of density characteristics that has different aggregation effects for different rows. Therefore, such information needs to be reprocessed by the direction-based feature-enhanced network to extract more useful information that can represent the crowd density in the image.

This network selects four directions to do the convolution operation for extracting as much information hidden as possible due to perspective. In addition, this network also has other benefits. It does not change the input feature dimension and introduce the global information into the local feature map. It also encodes the crowd scenes efficiently. After encoding the direction-based image feature, the initial image feature is added with the output of direction-based module to enhance the initial feature. Then, the fused image feature is sent into the multi-scale residual block.

\subsection{3.2 The multi-scale residual block}

As shown in Figure 4, the multi-scale residual block (MSRB) is designed to extract the image features at different scales. It mainly includes the multi-scale convolution layer and residual block. The multi-scale convolution layer can detect the image features at different scales adaptively and encode multi-level contextual information into the features. This operation is due to the scale variation characteristic of crowd scenes. The multi-scale structure can do the convolution operation with different kernel sizes. Three common different size convolution kernels are employed in this work: \(3 \times 3, 5 \times 5\) and \(7 \times 7\). Many other classical neural networks have used these kernels in the different deep learning tasks and have proved that they can obtain good performance in the computer vision domain. To accelerate the training speed and the convergence process of the model, the residual block is used to reinforce the output feature of MSRB, which is also widely used in other mainstream networks.

The MSRB has two convolution layers which all consist of three bypasses. The input feature map is firstly sent into three bypasses with different convolution kernels, whose size is \( W_0 \times H_0 \times C_0 \). The ReLU activation function \( \sigma \) is followed by convolution operation as Equation (2) shows. The three bypasses will generate three outputs. Such structure with three kinds of convolution kernels can extract different scale image features and fuse them. The traditional convolution operation keeps the channel independent, but such independence will limit the network’s representation ability. So, the feature between different channels is then concatenated and sent to the next convolution layer.

\[
\sigma(x) = \max(0, x)
\]

The concatenation operation can combine multiple features. The feature extracted by multiple convolutions can be fused and the crowd information of the image can be represented better. This work concatenates the feature map on channel dimension. Due to the scale variation, every network bypass uses different kernel sizes to do the convolution operation. And different kernel sizes will extract multi-scale information.

Given the initial feature map \( f_0 \), when the feature map is operated by \( i \times i \) convolution kernel, the output feature map can be expressed as \( f_0^i \). As the Equation (3) shows, the \( f_0^i \) can be obtained by \( f_0 \).

\[
f_0^i = \sigma \left( w_{i_0}^i \times f_0 + b_i^i \right)
\]
EXPERIMENTS

The feature-aware block combines three kinds of the convolution kernel. The difference with the first layer will also help obtain more metric information with the combination will also make the network more efficient. The initial image feature and the multi-scale information. Such after the multi-scale convolution operation. It can combine shown as Equation (5) will also keep more initial information into the second layer of MSRB. The second layer also has three kinds of convolution kernels. The difference with the first layer is that the second layer only do once concatenation operation. After this, the output feature map size can be changed to \( W_1 \times H_1 \times C_1 \).

In order to keep the final feature map size as the same as the initial input feature due to the concatenation operation, the output feature after second concatenation operation is dealt with \( 1 \times 1 \) convolution kernel. Such \( 1 \times 1 \) kernel can reduce the scale of the feature map. The function of \( 1 \times 1 \) kernel is the same as pooling layer. But the pooling layer just adjusts the image size instead of reducing the image channel. And \( 1 \times 1 \) kernel will also help obtain more metric information with the combination of three kinds of the convolution kernel.

\[
\begin{align*}
F_k &= \begin{bmatrix} f_0^3, f_0^5, f_0^7 \end{bmatrix} \\
F_k' &= \begin{bmatrix} f_0^1, f_0^3, f_0^5 \end{bmatrix}
\end{align*}
\]

The combination of the direction-based feature-enhanced block are used to construct the MSR-FAN. This structure is often used to solve the problem of gradient disappearance. In this work, the most efficient way to use the multi-scale information is to concatenate it to the original input features. This operation adds the input feature to the output feature as \( f_0 \). In the multi-scale network structure, this operation shown as Equation (5) will also keep more initial information after the multi-scale convolution operation. It can combine the initial image feature and the multi-scale information. Such combination will also make the network more efficient.

\[
f_0 = f_0 + F_k'
\]

In addition, the residual block is also used in MSRB. This structure is often used to solve the problem of gradient disappearance. In this work, the most efficient way to use the multi-scale information is to concatenate it to the original input features. This operation adds the input feature to the output feature as \( f_0 \). In the multi-scale network structure, this operation shown as Equation (5) will also keep more initial information after the multi-scale convolution operation. It can combine the initial image feature and the multi-scale information. Such combination will also make the network more efficient.

\[
S_{ij} = U_{ij} - f_0
\]

Equation (7) denotes the difference between \( U_0 \) and \( f_0 \). It can make the salient information more outstanding and can understand the difference between multiple channels better. Then, \( S_{ij} \) is concatenated with the input feature \( f_0 \) to enhance the representation of the image feature, as Equation (8) shows.

\[
C_O = \left[ S_{ij}, f_0 \right]
\]

Finally, the feature \( C_O \) passes to a decoder consisting of several dilated convolutions that produces the density map.

The following details reflect the structure of the proposed MSR-FAN:

- The combination of the direction-based feature-enhanced network, the multi-scale residual block, and the feature-aware block are used to construct the MSR-FAN.
- The multi-scale residual block can extract the image features at different scales and encode multi-level contextual information into the features.
- The feature-aware block can fuse the different channel’s features and extract the difference between local features and content features. It can improve the quality of the generated density map.

3.3 The feature-aware block

After MSRB, the output feature \( f_0 \) is sent into the feature-aware block, as Figure 5 shows. Some work [35] also used such feature aware module, which can extract multi-scale features hidden in the different channels of the feature map. The whole procedure can be described as Equation (6):

\[
C_O = \left[f_o, U\left(\Phi_1\left(\text{Avg}\left(f_o, C_\alpha\right)\right) - f_o\right)\right]
\]

With the use of average pooling, the feature map \( f_0 \) is averaged into \( K(j) \times K(j) \) blocks with Avg function. The size of \( f_0 \) is \( W_j' \times H_j' \times C_\alpha \) (weight, height, channel). The average pooling layer operates on the \( C_\alpha \) channel. It will generate \( K(j) \times K(j) \) blocks.

Such combination of the different channels can avoid the independence of each feature channel, which will limit the representation power. Ref. [35] used four different scales here, with corresponding block size \( K(j) \). But this work uses the MSRB before, which has extracted the scale information. So, the size of \( K(j) \) is decided by the \( f_0 \). Then, the averaged blocks are dealt with a convolutional network \( \Phi_1 \) with kernel size 1. This operation aims to fuse features between different channels without changing the channel dimensions. And function \( U \) represents bilinear interpolation to up-sample the array of feature. It aims to recover the size of \( U_0 \).

The 1 \( \times \) 1 convolution kernel can fuse the different channel features. To capture the feature-aware information hiding in the different channels, this paper defines the feature-aware feature as \( S_{ij} \), which is shown as Equation (7).

\[
S_{ij} = U_{ij} - f_0
\]

The structure of the feature-aware block

FIGURE 5

The feature-aware block can fuse the different channel’s features. This operation adds the input feature to the output feature as \( f_0 \). In the multi-scale network structure, this operation shown as Equation (5) will also keep more initial information after the multi-scale convolution operation. It can combine the initial image feature and the multi-scale information. Such combination will also make the network more efficient.

4 EXPERIMENTS

The MSR-FAN is implemented under the Windows 10 and Pytorch 1.2.0 experimental environment. The hardware environments are Inter Xeon E-2136 3.3 GHz and Quadro P5000.
This work uses the Adam optimizer with a learning rate of 0.00001 and L2Loss loss function.

4.1 Evaluation metrics

This work uses the common standard evaluation metrics to test the methods: mean absolute error (MAE) and the root mean squared error (RMSE) [45]. They are defined as follows, Equations (9) and (10).

\[
\text{MAE} = \frac{1}{N} \sum_{i=1}^{N} |y_i - \hat{y}_i| \quad (9)
\]

\[
\text{RMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2} \quad (10)
\]

The parameter \( N \) represents the total number of test images, \( y_i \) is the ground-truth number of people inside the whole \( i \)th image and \( \hat{y}_i \) is the estimated number of people.

4.2 Comparisons with state-of-the-art

This work compared the state-of-the-art methods on four benchmark datasets: ShanghaiTech [38], UCF_CC_50 [39], UCSD [40] and Mall [41]. ShanghaiTech dataset contains 1198 annotated images with 330,165 people. It consists of two parts: part_A with 482 images and part_B with 716 images. The two parts are divided into training data and testing data. UCF_CC_50 dataset only contains 50 images. The number of each image varies a lot, which makes it become a challenging image set. UCSD dataset has 2000 frames captured by surveillance cameras. Each image contains less than 50 persons. But the initial image’s size is small, which is difficult to generate the high-quality density map. The mall dataset contains 2000 annotated images. The researchers took pictures of pedestrians in the shopping mall. The experiment results show that the proposed MSR-FAN can get a satisfactory result on these four datasets.

Table 1 shows the comparison results of MSR-FAN with other methods on ShanghaiTech dataset. '-' means that the results are not given by author in the corresponding paper.

| Method         | Part_A MAE | Part_A RMSE | Part_B MAE | Part_B RMSE |
|----------------|------------|-------------|------------|-------------|
| Count-DANet [7]  | 85.9       | 130.3       | –          | –           |
| FCE-MCNN [19]   | 104.2      | 157.8       | 19.1       | 31.4        |
| FCE-CMTL [19]   | 83.6       | 122.3       | 16.7       | 25.8        |
| DSSINet [21]    | 60.6       | 96.0        | 6.85       | 10.34       |
| CSRNet [22]     | 68.2       | 115.0       | 10.6       | 16.0        |
| Yang [23]       | 61.2       | 96.9        | 8.1        | 11.6        |
| Wan [25]        | 63.1       | 96.2        | 8.72       | 13.56       |
| Bayesian+ [26]  | 62.8       | 101.8       | 7.7        | 12.7        |
| MBTTBF-SCFB [30]| 60.2       | 94.1        | 8.0        | 15.5        |
| PCC Net [31]    | 73.5       | 124.0       | 11.0       | 19.0        |
| CANNet [35]     | 62.3       | 100.0       | 7.8        | 12.2        |
| Ours            | 59.9       | 94.6        | 7.6        | 11.3        |

Table 1 Comparison of MSR-FAN with other methods on ShanghaiTech dataset. '-' means that the results are not given by author in the corresponding paper.

Table 2 shows the comparison results of MSR-FAN with other methods on UCF_CC_50 dataset. It is a challenging dataset with a various number of people in the image. The existing methods such as [21, 30, 34, 35] have achieved good performance. And the multi-scale model this paper proposed also tests on this dataset and obtains good performance. It gets the 210.7 MAE and 239.6 RMSE, which is better than other state-of-the-art methods. It can also be found that the MSR-FAN and CANNet [35] are close to each other in MAE values. The feature-aware module can extract the difference between local features and context features. It can be inferred that the direction-based feature-enhanced network and the MSRB can pay more attention to the scale information, which is useful to solve the perspective changes and the size various.

Table 2 Comparison of MSR-FAN with other methods on UCF_CC_50 dataset.

| Method     | MAE  | RMSE |
|------------|------|------|
| FCE-MCNN [19]     | 312.8 | 405.2 |
| FCE-CMTL [19]     | 287.7 | 382.2 |
| DSSINet [21]      | 216.9 | 302.4 |
| CSRNet [22]       | 266.1 | 397.5 |
| Wan [25]          | 355.0 | 560.2 |
| Bayesian+ [26]    | 229.3 | 308.2 |
| MBTTBF-SCFB [30]  | 231.3 | 300.9 |
| SANet+SPANet [34] | 232.6 | 311.7 |
| CANNet [35]       | 212.2 | 243.7 |
| Ours             | 210.7 | 239.6 |

Table 2 Comparison of MSR-FAN with other methods on UCF_CC_50 dataset.
TABLE 3 Comparison of MSR-FAN with other methods on UCSD dataset

| Method                | MAE   | RMSE |
|-----------------------|-------|------|
| MSHD [8]              | 2.10  | 6.5  |
| CSRNet [22]           | 1.07  | 1.35 |
| Yang [23]             | 1.32  | 1.23 |
| SANet+SPANet [34]     | 1.00  | 1.28 |
| Ours                  | 1.22  | 1.36 |

TABLE 4 Comparison of MSR-FAN with other methods on mall dataset

| Method                | MAE   | RMSE |
|-----------------------|-------|------|
| Count-DANet [7]       | 2.1   | 2.7  |
| MSHD [8]              | 2.9   | 14.4 |
| SAMC-Net [42]         | 2.3   | 2.9  |
| RPF [43]              | 3.2   | 15.5 |
| Li [44]               | 2.5   | 3.1  |
| Ours                  | 2.0   | 2.4  |

of ours is close to [34]. But the accuracy this work obtained is sufficient for practical applications in this dataset.

Table 4 shows the comparison results of MSR-FAN with other methods on the mall dataset. The mall dataset is similar to UCSD, which contains single scene and a small number of pedestrians. On this dataset, this work gets the best result compared with [7, 8, 42, 43, 44]. It shows that MSR-FAN has good generalization ability for general datasets. The results of other methods all get a good performance. This result indicates that this dataset is not challenging.

In addition to making comparisons on these four benchmark datasets, this paper also visualizes the experimental results on ShanghaiTech part_A dataset. As the Figure 6 shows, the MSR-FAN and the MSR-FAN without direction-based feature-enhanced network are tested. The five images are chosen and each image contains abundant person, which is representative of crowd scenes. The row 3 shows the estimated results of the MSR-FAN. The estimated density map can reflect the density of the crowd as the ground-truth (row 4). To show the usefulness of the introduction of direction-based feature-enhanced network, this work also removes this network from MSR-FAN to test the model performance. From the row 2, it can be found that the MSR-FAN without the network loses some detailed information compared with ground-truth.

Figures 7 and 8 show the exact results of MSR-FAN on ShanghaiTech part_A and part_B. The blue points represent the real number of crowd people. The red points represent the estimated number of the crowd. The tested images are selected randomly from the dataset. Every point is an image. From the results, it can be found that the proposed model performs well on ShanghaiTech. The red line and the blue line fit well. The minimum error is zero. The trends of the two curves are very similar, which means that the error in the model is very small. The more similar the lines are, the better the model performs.

4.3 Ablation studies for the proposed MSR-FAN

To understand the validity of different sub-modules, this work does some ablation studies on ShanghaiTech dataset. The different feature extraction feed-forward network and the MSR-FAN without direction-based feature-enhanced network are considered to be tested. This work considers the VGG and the ResNet as the different feature extraction network, due to their good representational capacity of extracting features. The following configurations are the detailed information:

a. The MSR-FAN with different feature extraction network: denoted as the MSR-FAN (VGG-MSR-FAN) and the ResNet-MSR-FAN.

b. The MSR-FAN without direction-based feature-enhanced network: denoted as the MSR-FAN-no-DFN and the ResNet-MSR-FAN-no-DFN.

From Table 5, it can be found that the replacement of ResNet cannot get better accuracy and even lead to lower MEA and RMSE. This work considers that the deep structure of ResNet will lose more useful and significant information when facing the perspective problem and the very dense scenes. The configuration (b) confirms the effectiveness of the use of the DB-FEN, which encodes the input feature in four different directions. The ResNet feed-forward network also performs not well here, the MAE on part_A just reaches to 78.6, which is far away VGG-MSR-FAN (MSR-FAN). So, the combination of the direction-based feature-enhanced network, MSRB, and the feature-aware block can get satisfactory result.

5 CONCLUSION

This paper proposes a multi-scale residual feature-aware network that combines multi-scale feature using multiple receptive field sizes and learns the feature aware information of each
To represent the regional feature better, this work proposes multi-scale residual block to get the multi-scale information. Besides, the feature-aware block is also employed to learn the importance of the regional feature. Experiment results based on benchmark datasets show that the proposed approach outperforms the existing methods.

In the future, this work will continue to improve the multi-scale structure and combine new technology to promote the quality of the generated density map.

ACKNOWLEDGMENTS
This work was supported by the National Natural Science Foundation of China (Grant No. 62076117 and No. 61762061), the Natural Science Foundation of Jiangxi Province, China (Grant No. 20161ACB20004) and Jiangxi Key Laboratory of Smart City (Grant No. 20192BCD40002).

ORCID
Haoyu Zhao  https://orcid.org/0000-0003-3832-6439
Weidong Min  https://orcid.org/0000-0003-2526-2181
Xin Wei  https://orcid.org/0000-0003-0976-3381
REFERENCES

1. Liu, Y., et al.: Crowd counting via cross-stage refinement networks. IEEE Trans. Image Process. 29, 6800–6812 (2020)

2. Idrees, H., Soomro, K., Shah, M.: Detecting humans in dense crowds using locally-consistent scale prior and global occlusion reasoning. IEEE Trans. Pattern Anal. Mach. Intell. 37(10), 1986–1998 (2015)

3. Huang, S., et al.: Body structure aware deep crowd counting. IEEE Trans Image Process. 27(3), 1049–1059 (2018)

4. Li, M., et al.: Estimating the number of people in crowded scenes by mid based foreground segmentation and head-shoulder detection. In: International Conference on Pattern Recognition (ICPR). Florida, pp. 1–4 (2008)

5. Lin, Z., Davis, L.S.: Shape-based human detection and segmentation via hierarchical part-template matching. IEEE Trans. Pattern Anal. Mach. Intell. 32(4), 604–618 (2010)

6. Yang, H., et al.: Driver yawning detection based on subtle facial action recognition. IEEE Trans. Multimedia 23, 572–583 (2021)

7. Zhang, S., Li, H., Kong, W.: Object counting method based on dual attention network. IET Image Process. 14(8), 1621–1627 (2020)

8. Ma, T., Ji, Q., Li, N.: Scene invariant crowd counting using multi-scales head detection in video surveillance. IET Image Process. 12(12), 2258–2263 (2018)

9. Wang, Q., et al.: Pixel-wise crowd understanding via synthetic data. Int. J. Comput. Vision 129, 225–245 (2021)

10. Wang, M., Wang, X.: Automatic adaptation of a generic pedestrian detector to a specific traffic scene. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Colorado Springs, pp. 3401–3408 (2011)

11. Zhao, T., Nevatia, R., Wu, B.: Segmentation and tracking of multiple humans in crowded environments. IEEE Trans. Pattern Anal. Mach. Intell. 30(7), 1198–1211 (2008)

12. Bartošová, O., Lempitsky, V., Kohli, P.: On the detection of multiple object instances using Hough transforms. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR). California, pp. 2233–2240 (2010)

13. Desai, C., Ramanan, D., Fowlkes, C.: Discriminative models for multi-class object layout. In: Proceedings of IEEE International Conference on Computer Vision (ICCV). Kyoto, Japan, pp. 229–236 (2009)

14. Descombes, X., Minlos, R., Zhizhina, E.: Object extraction using a stochastic birth-and-death dynamics in continuum. J. Math. Imaging Vision 33(3), 347–359 (2009)

15. Leibe, B., Leonardis, A., Schiele, B.: Robust object detection with interleaved categorization and segmentation. Int. J. Comput. Vision 77(1–3), 259–289 (2008)

16. Wu, B., Nevatia, R., Li, Y.: Segmentation of multiple, partially occluded objects by grouping, merging, assigning past detection responses. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Anchorage, Alaska, pp. 1–8 (2008)

17. Wang, Q., et al.: Discriminative fine-grained network for vehicle re-identification using two-stage re-ranking. Sci. China Inf. Sci. 63(11), 1–12 (2020)

18. Min, W., et al.: A new approach to track multiple vehicles with the combination of robust detection and two classifiers. IEEE Trans. Intell. Transp. Syst. 19(1), 174–186 (2018)

19. Wu, X., et al.: Feature channel enhancement for crowd counting. IET Image Process. 14(11), 2376–2382 (2020)

20. Xiong, H., et al.: From open set to closed set counting objects by spatial divide-and-conquer. In: Proceedings of IEEE International Conference on Computer Vision (ICCV). Seoul, Korea, pp. 8361–8370 (2019)

21. Liu, L., et al.: Crowd counting with deep structured scale integration network. In: Proceedings of IEEE International Conference on Computer Vision (ICCV). Venice, Italy, pp. 1774–1783 (2020)

22. Li, Y., Zhang, X., Chen, D.: CSRNet: Dilated convolutional neural networks for understanding the highly congested scenes. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Utah, USA, pp. 1091–1100 (2018)

23. Yang, Y., et al.: Reverse perspective network for perspective-aware object counting. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Washington, pp. 4373–4382 (2020)

24. Shi, Z., et al.: Crowd counting with deep negative correlation learning. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Utah, USA, pp. 5382–5390 (2018)

25. Wang, J., et al.: Residual regression with semantic prior for crowd counting. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Washington, pp. 4031–4040 (2020)

26. Ma, Z., et al.: Bayesian loss for crowd count estimation with point supervision. In: Proceedings of IEEE International Conference on Computer Vision (ICCV). Venice, Italy, pp. 6141–6150 (2020)

27. Babu Sam, D., et al.: Locate, Size and Count: Accurately Resolving People in Dense Crowds via Detection, IEEE Trans. Pattern Anal. Mach. Intell. 1–1 (2019). https://doi.org/10.1109/TPAMI.2020.2974830

28. Xiong, H., Liu, H., Liu, C., et al.: From Open Set to Closed Set: Supervised Spatial Divide-and-Conquer for Object Counting. Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Seoul, Korea, pp. 8361–8370 (2020)

29. Zeng, L., et al.: Multi-scale convolutional neural networks for crowd counting. In: Proceedings of IEEE International Conference on Image Processing (ICIP). Beijing, China, pp. 465–469 (2017)

30. Sindagi, V.A., Patel, V.M.: Multi-level bottom-top and top-bottom feature fusion for crowd counting. In: Proceedings of IEEE International Conference on Computer Vision (ICCV). Seoul, Korea, pp. 1002–1012 (2019)

31. Gao, J., Wang, Q., Li, X.: PCC Net: Perspective crowd counting via spatial convolutional network. IEEE Trans. Circuits Syst. Video Technol. 30(10), 3486–3498 (2019)

32. Liu, X., Yang, J., Ding, W.: Adaptive mixture regression network with local counting map for crowd counting. In: European Conference on Computer Vision (ECCV). Glasgow, England, pp. 241–257 (2020)

33. Wang, Q., et al.: NWPU-Crowd: A large-scale benchmark for crowd counting and localization. IEEE Trans. Pattern Anal. Mach. Intell. 99, 1 (2020)

34. Cheng, Z., et al.: Learning spatial awareness to improve crowd counting. In: Proceedings of IEEE International Conference on Computer Vision (ICCV). Seoul, Korea, pp. 6151–6160 (2019)

35. Liu, W., Salzmann, M., Fua, P.: Context-aware crowd counting. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, CA, pp. 5094–5103 (2019)

36. Wang, Q., et al.: Density-aware curriculum learning for crowd counting. IEEE Trans. Cybern. 99, 1–13 (2020)

37. Wang, Q., et al.: Neural Linear Transformation: Modeling the Domain Shift for Crowd Counting. IEEE Trans. Neural Netw. Learn. Syst. 1–13 (2020). https://doi.org/10.1109/TNNLS.2020.3033428

38. Zhang, Y., et al.: Single-image crowd counting via multi-column convolutional neural network. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, pp. 589–597 (2016)

39. Idrees, H., et al.: Multi-source multi-scale counting in extremely dense crowd images. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Portland, Oregon, pp. 2547–2554 (2013)

40. Zhang, C., et al.: Cross-scene crowd counting via deep convolutional neural networks. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston, pp. 833–841 (2015)

41. Chen, K., et al.: Feature mining for localised crowd counting. In: Proceedings of BMVC. Bristol, UK, pp. 1–3 (2012)

42. Li, H., Zhang, S.H., Kong, W.H.: Crowd counting using a self-attention multi-scale cascaded network. IET Comput. Vis. 13(6), 556–561 (2019)
43. Xu, B., Qiu, G.: Crowd density estimation based on rich features and random projection forest. In: IEEE Winter Conference on Applications of Computer Vision (WACV), Nevada, pp. 1–8 (2016)

44. Li, H., Zhang, S.H., Kong, W.H.: Bilateral counting network for single-image object counting. Vision Comput. 36(8), 1693–1704 (2019)

45. Siva, P., et al.: Real-time, embedded scene invariant crowd counting using scale-normalized histogram of moving gradients (HoMG). In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, pp. 885–892 (2016)

How to cite this article: Zhao, H., et al.: MSR-FAN: Multi-scale residual feature-aware network for crowd counting, IET Image Processing, 1–10 (2021) https://doi.org/10.1049/ipt2.12175.