Received: 29 January 2021 | Revised: 7 March 2021 | Accepted: 28 March 2021

IET Image Processing

DOI: 10.1049/ipr2.12230

The Institution of Engineering and Technology WILEY

ORIGINAL RESEARCH PAPER

MFP-Net: Multi-scale feature pyramid network for crowd counting

Tao Lei[1,2] | Dong Zhang[1,2] | Risheng Wang[1,2] | Shuying Li[3] | Weijiang Zhang[4] | Asoke K. Nandi[5,6]

[1] Shaanxi Joint Laboratory of Artificial Intelligence, Shaanxi University of Science and Technology, Xi'an, China

[2] The School of Electronic Information and Artificial Intelligence, Shaanxi University of Science and Technology, Xi'an, P. R. China

[3] The School of Automation, Xi'an University of Posts and Telecommunications, Xi'an, China

[4] The School of Electronical and Control Engineering, Shaanxi University of Science and Technology, Xi'an, China

[5] The Department of Electronic and Electrical Engineering, Brunel University London, Uxbridge, UK

[6] School of Mechanical Engineering, Xi'an Jiaotong University, Xi'an, China

Correspondence
Shuying Li, The School of Automation, Xi'an University of Posts and Telecommunications, Xi'an 710121, China.
Email: angle_lisy@163.com

Funding information
National Natural Science Foundation of China, Grant/Award Numbers: 61871259, 61861024, 61701387; Natural Science Basic Research Program of Shaanxi, Grant/Award Number: 2021JC-47; Science and Technology Program of Shaanxi Province of China, Grant/Award Number: 2020NY-172

Abstract

Although deep learning has been widely used for dense crowd counting, it still faces two challenges. Firstly, the popular network models are sensitive to scale variance of human head, human occlusions, and complex background due to repeated utilization of vanilla convolution kernels. Secondly, the vanilla feature fusion often depends on summation or concatenation, which ignores the correlation of different features leading to information redundancy and low robustness to background noise. To address these issues, a multi-scale feature pyramid network (MFP-Net) for dense crowd counting is proposed in this paper. The proposed MFP-Net makes two contributions. Firstly, the feature pyramid fusion module is designed that adopts rich convolutions with different depths and scales, not only to expand the receptive field, but also to improve the inference speed of models by using parallel group convolution. Secondly, a feature attention-aware module is added in the feature fusion stage. The module can achieve local and global information fusion by capturing the importance of the spatial and channel domains to improve model robustness. The proposed MFP-Net is evaluated on five publicly available datasets, and experiments show that the MFP-Net not only provides better crowd counting results than comparative models, but also requires fewer parameters.

1 | INTRODUCTION

Dense crowd analysis is one of the most challenging tasks in video surveillance, traffic guidance, public safety prevention and control, and intelligent environment design. The task [1] focuses on crowd counting [2–10,64], crowd image segmentation [11–15], crowd detection and tracking [16–19], and crowd behaviour recognition and localization [20, 21]. Among them, crowd counting is a basic task in the field of crowd analysis and can be also applied to vehicle detection [22, 23], biometric counting [24–26] etc. Recently, a large number of approaches used for crowd counting have been reported, and these approaches can be roughly grouped into three categories: regression-based approaches, detection-based approaches, and density-map-estimation-based approaches.

Traditional approaches for crowd counting mainly depend on regression and detection technique. Detection-based methods are often used to calculate the number of people by detecting the head or appearance of pedestrians by means of dynamic frame detectors [27–31]. These methods are effective in sparse scenes, but they do not perform well in scenes with heavy human occlusions and complex backgrounds. Regression-based methods [32–36] are often used to construct regression models for crowd counting by learning the mapping relationship between shallow features of images and the number of crowds, such as Gaussian-mixture regression, linear regression etc. These regression-based methods can deal with dense crowd counting in complex scenes, but they seriously rely on the feature representation at shallow layers and ignore the spatial features of images, resulting in poor model generalization ability and accuracy.

With the rapid development of urbanization, crowd gathering activities are becoming more and more frequent. However, the early crowd counting models focus only on the overall representation of the situation, which is unsuitable for crowd analysis under some complex scenes. For the problem, the density map estimation is better than crowd counting since it estimates the number of people by integrating over the whole image and refines the distribution of local locations. As a result, the crowd counting task has evolved from simple crowd counting to density map estimation that can represent crowd distribution characteristic. Based on this idea, Lempitsky et al. [37] proposed a method based on density map estimation by learning a linear mapping between local features and density maps. However, the method easily suffers from some difficulties because the relationship between features and density maps is usually nonlinear. In order to reduce the difficulty of learning linear mapping relationship, Pham et al. [38] proposed a nonlinear mapping algorithm by utilizing spatial structure information, which applies random forest regression for semi-automatic training of the model. However, the algorithm provides a low accuracy for crowd counting under scenes with high density distribution due to the reliance on manually extracted low-level features.

Currently, deep learning is the most popular technique in computer vision due to the achievement of hierarchical feature representation. Based on the deep learning technique, Fu et al. [39] firstly applied general Convolutional Neural Network (CNN) to crowd counting by evaluating crowd density map. However, the general CNN suffers from the problem of loss of image spatial information caused by the employment of fully connected layer for image segmentation. To solve this problem, Fully Convolutional Neural Network (FCN) [40] uses convolution layer instead of fully connected layer to achieve end-to-end pixel-level classification. Since FCN achieves better image segmentation than traditional algorithm [41, 42], it becomes the most popular backbone for image segmentation tasks [43]. Currently, many researchers have applied FCN and improved FCNs to the field of dense crowd analysis, and these FCN-based methods are roughly classified into three categories in terms of model architecture and tasks, that is, multi-branch network for single-task, joined multi-branch networks for multi-task, cascade network for single-task.

Multi-branch networks usually use convolutional kernels with different sizes to capture multi-scale image information under different receptive fields. Based on this idea, Zhang et al. [5] designed a multi-column CNN(MCNN) consisting of three encoders, where 3×3, 5×5 and 7×7 convolutional kernels are used for these encoders, respectively. After that, multi-scale features of human head from three encoders are directly concatenated and fused. Finally, the predicted density maps are obtained by a decoder using 1×1 convolution. Although MCNN provides better feature representation than general CNNs due to the utilization of multi-scale features, when the networks go deeper, it will require more parameters and higher computational cost. To solve the problem, Zeng et al. [3] used the inception structure to extract multi-scale image features and used 1×1 convolution for dimensionality reduction, which to some extent reduces the number of parameters and computational cost of MCNN. Furthermore, Sam et al. [2] presented a switching CNN that includes three independent branching networks, where the input data is fed into one of the branching networks in terms of density rank of input data. More related works can be seen in [22, 44].

Since multi-task learning is helpful for improving the generalization ability of networks on the estimation of crowd density map, many researchers tried to design multi-branch networks for crowd density map estimation. Sindagi et al. [6] proposed a joining learning model used for two tasks, that is, density classification and density map estimation. By using high-level density prior knowledge, the proposed model can accelerate learning and improves prediction accuracy. To further improve prediction accuracy, Gao et al. [45] presented the Perspective Crowd Counting Network (PCC-Net) that consists of three parts: advanced density classification, density map estimation, and image semantic segmentation. For multi-task learning, semantic segmentation effectively distinguishes the background and foreground, and thus improves density map estimation and generalization ability of the network. Based on the idea, some researchers designed multi-task learning frameworks for pixel-level domain adaptation and density map estimation [8, 10, 11, 46]. These frameworks employ generative adversarial networks (GANs) to reduce the difference of scene changes between synthetic and real data, which further reduces the dependence of deep learning on real data. Wang et al. [11] proposed Spatial Fully Convolutional Network (SFCN) that uses dilated convolution and a spatial encoder to incorporate global contextual information leading to higher accuracy for crowd density map estimation. However, multi-task networks depend on richer label data that are uneasily obtained.

Cascade networks can provide better feature representation since deeper convolutional layers are often deployed in cascade networks. Li et al. [47] presented a single-column deep network structure (CSRNet) that extracts multi-scale contextual information using dilated convolutional kernels of small-size. The CSRNet greatly improves the density map estimation and counting accuracy for crowd images. Jiang et al. [48] designed hierarchical decoders for different encoding stages and employed dense skip-connections to promote the fusion of multi-scale features, thus improving the quality of predicted density maps. Meanwhile, researchers developed relevant learning strategies based on CNNs to obtain prior knowledge before training, allowing the model to learn incrementally from easy to difficult [9, 49, 50]. In addition, for density map estimation, the removal of background noise is important. Since the attention mechanism allows the model to dynamically focus on the important positions in an image, it can enhance the feature

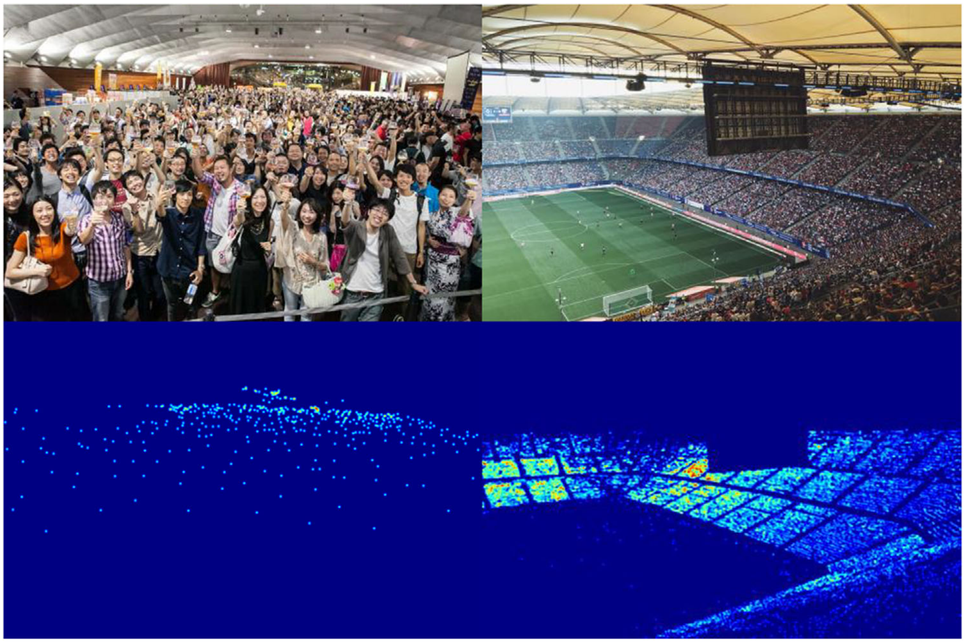


FIGURE 1 The first row is the samples from the NWPU-Crowd dataset [7] and the second row is the ground truth density maps

representation of networks. Based on the idea, researchers [51–55] often incorporated an attention module into networks to capture information of interesting regions while ignoring irrelevant information. Although these networks show strong robustness on background noise, they do not consider the relationship between feature channels and feature spatial. Compared with multi-branch networks mentioned above, the cascade networks often provide higher accuracy for the crowd density map estimation.

Although both cascade networks and multi-branch networks are successful for crowd counting and density map estimation under some simple scenes, they usually suffer from difficulties caused by complex scenes such as viewing angle of far distance, the scale variance, noise, human occlusions, complex background etc. Figure 1 shows some complex scenes. Despite cascading larger convolutional kernels for multiscale feature extraction can improve the prediction effect, it still causes some new problems, such as more parameters, higher computational cost, and more difficult training. Inspired by [56, 57], we propose a multiscale feature pyramid network (MFP-Net) for crowd counting and density map estimation. The main contributions of this paper are given as follows:

(1) For network encoding, we present a feature pyramid fusion module (FPFM). The FPFM employs multiple convolution kernels with different depths and dilation rates to perform group and parallel operations on the input feature maps, which can effectively capture multi-scale contextual information and obtain better feature representations.
(2) For feature fusion, we present a feature attention-aware module (FAAM). The FAAM can reduce the effect of background clutter, improves the robustness of the model by dynamically paying attention to interesting regions in images, and leans the visual correlation between spatial and channels.
(3) The proposed MFP-Net shows fast training and inference speed due to the utilization of parallel convolution, and provides better crowd counting and density map estimation due to the utilization of adaptive multi-scale feature fusion.

The remainder of the paper is organized as follows. Section 2 describes in detail the structure and advantages of the MFP-Net. Section 3 focuses on the ablation studies and the analysis of comparative experimental results. In Section 4, summary and discussion are presented.

2 | METHOD

In this paper, we propose a multi-scale feature pyramid network (MFP-Net) and apply it to the field of crowd counting and density map estimation. Figure 2 shows the architecture of MFP-Net that consists of feature extraction layer, feature pyramid fusion layer and feature attention-aware layer. Firstly, considering that VGG16 has a low training cost and high performance, we choose the first 10 layers of VGG16 as the feature extraction layer of MFP-Net. Secondly, the feature maps obtained by feature extraction layer are fed into the feature pyramid fusion module that adopts pyramid group convolution to efficiently perform multi-scale feature extraction and obtain different levels of fine-grained information from input images. Then, the obtained multi-scale feature maps are fed into the feature attention-aware module that can capture more meaningful features and achieve adaptive fusion of local and global information, which is helpful for improving feature representation and the robustness of the network on the tasks of crowd

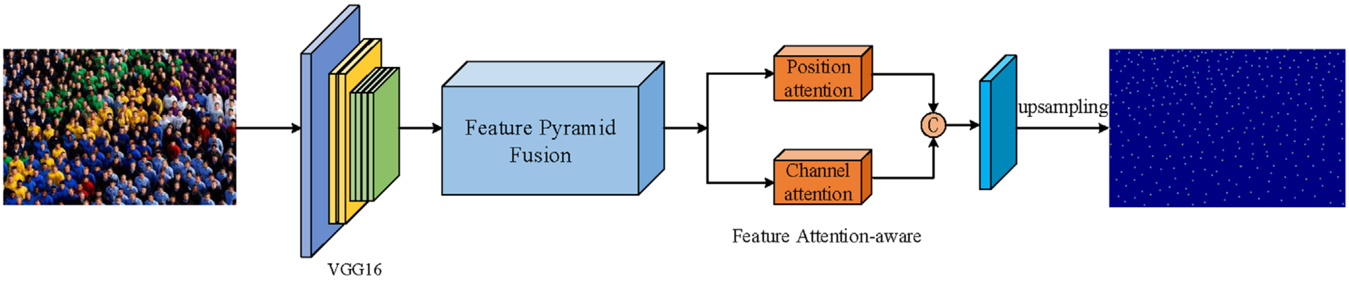


FIGURE 2 The architecture of the proposed MFP-Net. First, we use VGG16 as the backbone. Secondly, the feature pyramid fusion module is used to extract multi-scale information and then use the feature attention-aware module for feature fusion

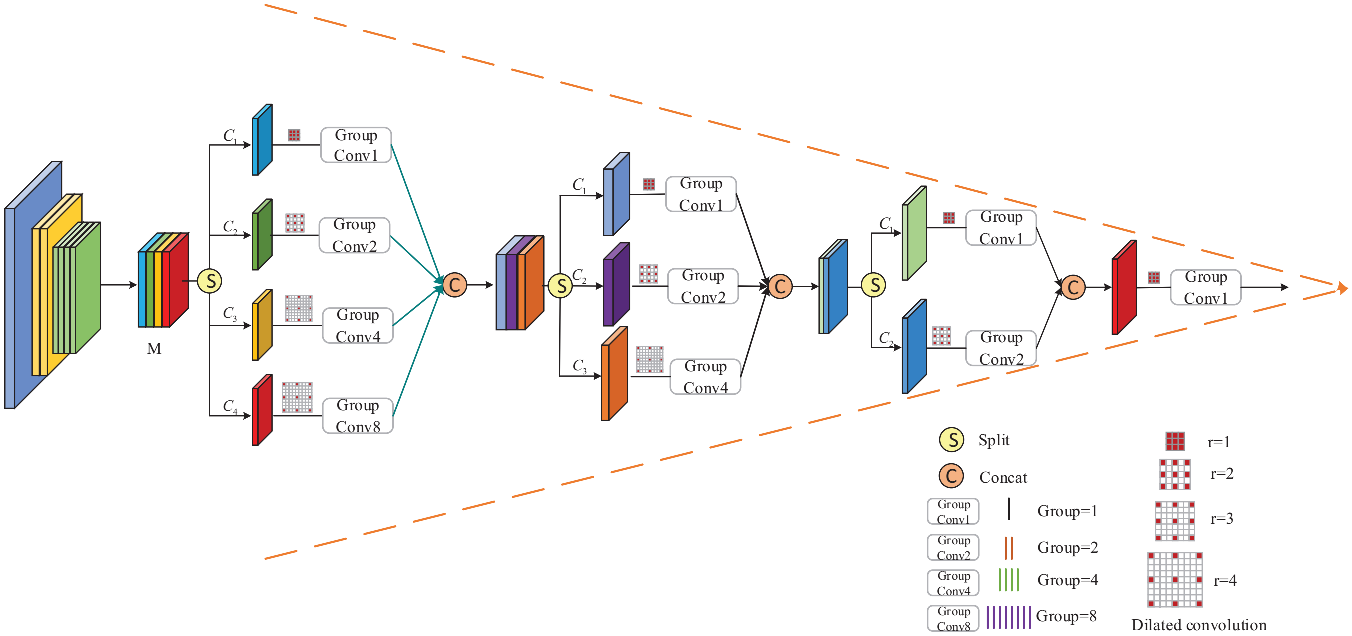


FIGURE 3 The architecture of the feature pyramid fusion module

counting and density map estimation. Finally, the feature maps are restored to the original image size by using bilinear interpolation without parameters. In fact, both the feature pyramid fusion module and the feature attention-aware module are used for feature fusion. However, the former aims to extract multi-scale information from crowed images, but the latter aims to capture the key information and suppress the influence of background noise.

2.1 | Feature pyramid fusion module

Dense crowd images often have a complex background and a large variation in the scale of crowd object proximity and distance. Thus, vanilla convolution has two drawbacks. One is vanilla convolution has a fixed receptive field and cannot efficiently extract multi-scale information in crowd images. The other is the stacking of different scale vanilla convolutional kernels causes the increase of the number of network parameters and the reduction of network robustness. To solve these two problems, we design the feature pyramid fusion module to capture the multi-scale information in crowd images. Roughly, we first split the feature maps into multiple blocks, and then perform 3 × 3 group convolution with different dilation rates on each block as shown in Figure 3. The feature pyramid fusion module performs different levels of filtering operations on the input feature maps by using group convolution, which captures different multi-scale contextual information in parallel computation. Moreover, the group convolution can reduce the computational cost and thus improves the inference speed of networks.

In Figure 3, the feature pyramid fusion module includes four pyramidal convolution layers. In each layer, the feature maps are first divided into proportional blocks. And then each block is performed group dilated-convolution. Specifically, the channel number of the input feature map is M. In the first layer, we divide feature maps into 4 blocks, and the channel number of each block is C_1, C_2, C_3, C_4, respectively, where $C_1 + C_2 + C_3 + C_4 = M$. All convolutional kernels are 3 × 3 with different dilation rates r(r=1,2,3,4), where the number of groups G increases by 2^n in pyramid-shape, for example, $G = (2^0, 2^1, 2^2, 2^3)$. For the second layer, we divide the feature maps into three blocks. Similarly, the size of all convolutional

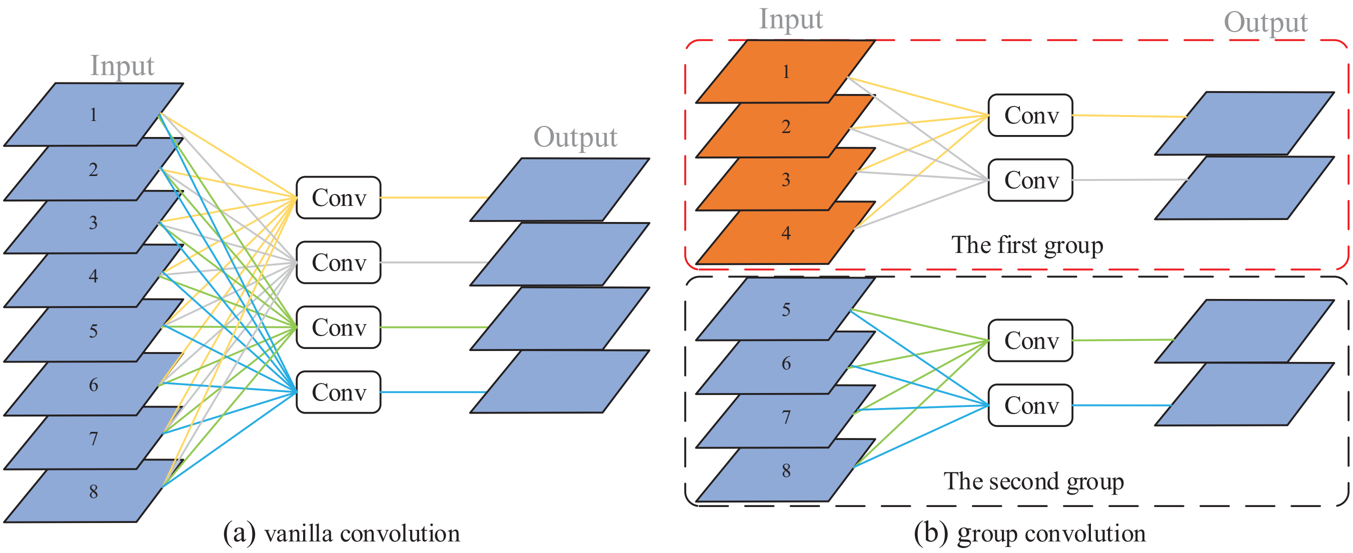


FIGURE 4 Comparison of vanilla convolution and group convolution

kernel is 3 × 3, and the dilation rate r and the number of groups G are incremented sequentially starting from 1. The third layer is divided into two blocks. For the last layer, we adopt vanilla convolution with $G = 1$. Here, we define the input feature maps as x and the output feature maps as y, then

$$y_i(x) = \begin{cases} PGconv(x, N_i, G_i, r_i), & i = 1 \\ PGconv(y_1(x), N_i, G_i, r_i), & i = 2 \\ \vdots & \\ PGconv(y_{L-1}(x), N_i, G_i, r_i), & i = L \end{cases}, \tag{1}$$

where$PGconv(x, N_i, G_i, r_i)$ is the pyramid group dilated-convolution,L is the number of layers,N_i is the number of blocks, r_i is the dilation rate, and G_i is the number of groups in each convolution operation. Note that L, N_i, r_i, and G_i are hyperparameters that can be adjusted according to different tasks. As shown in Figure 3, our presented feature fusion module is pyramid-shaped, and the number of groups as well as the dilation rate increases in a pyramidal pattern. In such a way that different levels of fine-grained features can be extracted, and multi-scale features can be clustered at the top of the pyramid. Compared with the vanilla convolution, the proposed pyramid group dilated-convolution can extract richer multi-scale image features due to the utilization of dilated convolution with different dilation rate. Compared with multi-scale convolution, the proposed method requires fewer parameters and lower computational cost due to the utilization of pyramid grouping. Therefore, the feature pyramid fusion module both extracts multi-scale information and simultaneously improves the inference speed of the network. Figure 4 shows the comparison of vanilla convolution and group convolution, where the number of groups is 2. Obviously, the computational cost is greatly reduced when we use group convolution instead of vanilla convolution. The computational cost of group convolution is denoted by F, and

$$F(G, K, C_{in}, C_{out}) = \frac{K^2 \times C_{in} \times C_{out} \times H \times W}{G}, \tag{2}$$

where G is the number of groups, K is the size of the convolution kernel, C_{in} and C_{out} are the number of input and output feature maps, respectively, H and W are the height and width of the feature maps. In fact, when performing group convolution, a larger value of G corresponds to a lower computational cost $F(G, K, C_{in}, C_{out})$. For the feature pyramid fusion module, the number of groups, blocks and dilation rates gradually decrease in a pyramidal way as the network depth increases, which aims to achieve a balance between feature representation and inference speed. Although group convolution is able to improve inference speed of networks, it leads to feature loss since feature maps are grouped for convolution. Consequently, the pyramid group dilated-convolution requires a low computational cost:

$$F(N, G, K, C_{in}, C_{out}) = \sum_{i=1}^{N} \left(\frac{K_i \times C_{in}^{i} \times C_{out}^{i} \times H \times W}{G_i} \right), \tag{3}$$

where N is the number of proportionally divided blocks. C_{in}^{i} and C_{out}^{i} are the number of input and output feature maps in the i^{th} group, respectively, and $C_{in}^{i} \in C_{in}, C_{out}^{i} \in C_{out}$. G_i is the number of groups in the i-th block, $G_i \in G$. Obviously, it can be concluded from (3) that the dilated convolution expands the receptive field without losing the resolution of the feature maps. Meanwhile the group parallel convolution can reduce computational cost.

In conclusion, our presented feature pyramid fusion module not only provides better feature representation caused by multi-scale feature extraction, but also achieves fast inference speed and requires lower computational cost due to the employment of pyramid group convolution.

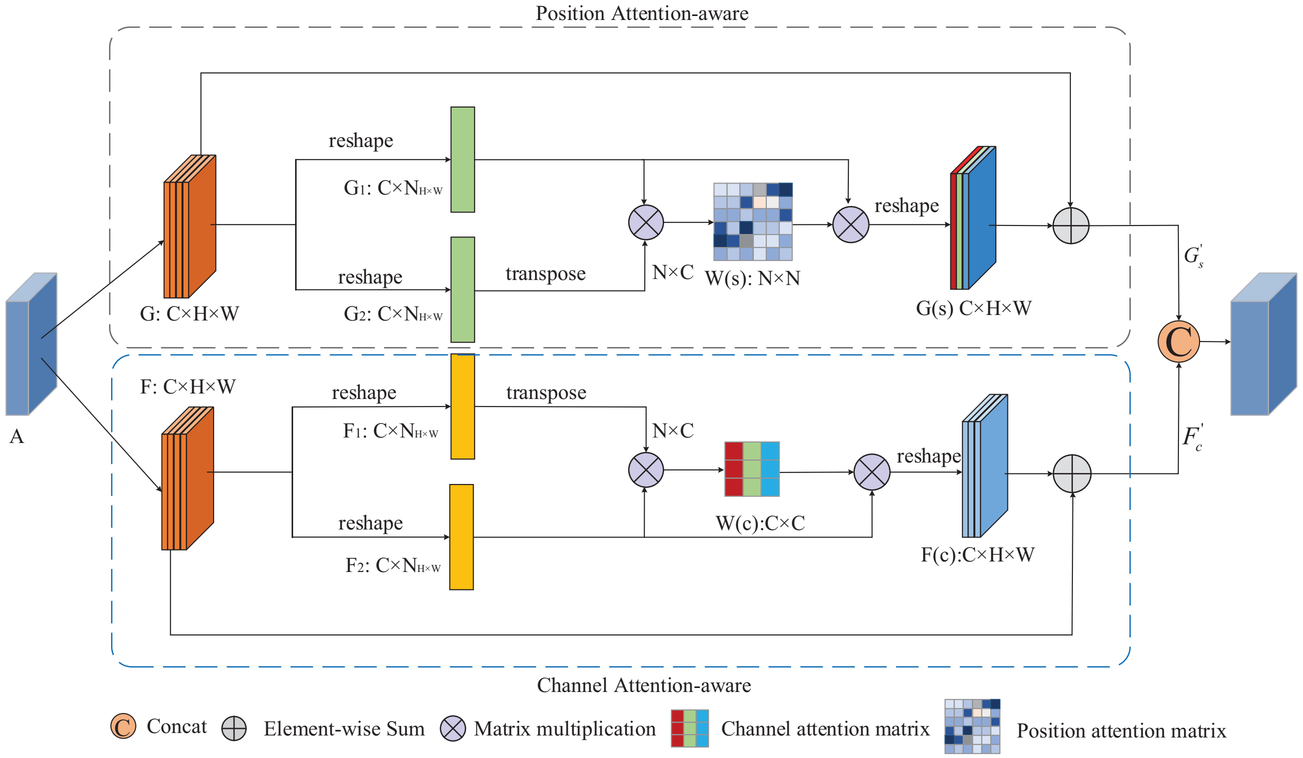


FIGURE 5 The structure of feature attention-aware module

2.2 | Feature attention-aware module

For the task of crowd density estimation, congested scenes have disturbing factors such as scale, perspective, occlusion, brightness etc. Previous works extract features as equally important and fuse them indiscriminately, resulting in poor robustness. However, the cluttered background is often identified as crowd during the actual density map estimation that affects the accuracy of the model prediction. Therefore, it is a challenge to effectively identify confusable features. The attention mechanism can enhance the robustness of feature representation by focusing on important regions and key channels of feature maps. Liu et al. [51] adopted attention mechanism to obtain local location information in feature maps yet ignores the global correlation between feature channels. Sindagi et al. [58] injected foreground and background segmentation information into the counting network through an attention mechanism but did not consider the relationship between each location in the image. For these problems, we present a feature attention-aware module that integrates both position and channel attention mechanism. The position attention-aware focuses on the spatial correlation between each position in the image to capture the global correlation of images, which is helpful for mapping the distribution pattern of crowd density. The channel attention-aware focuses on the correlation between different channels in feature maps and emphasizes the interdependence between feature maps by assigning different weights. Due to the semantic relevance of spatial and channels, we use both channel and position attention and then adaptively fuse the obtained feature maps.

Figure 5 illustrates the structure of feature attention-aware module. The structure that contains two types of attention modules that explore local and global contextual information by constructing associations between features to improve feature representation for crowd counting and density estimation. On the one hand, the position attention-aware module encodes a wider range of contextual information than convolution operation, thus enhancing the representation of local features. On the other hand, the channel attention-aware module reduces the effect of useless feature maps caused by background noise. The feature map $G \in \mathbb{R}^{C\times H\times W}$output from the feature pyramid fusion module is fed into a convolution layer to obtain two feature maps (G_1, G_2) that are reshaped to $\mathbb{R}^{C\times N}$, where C is the number of channels of feature maps, $H \times W$ denotes the spatial dimension, $N = H \times W$ denotes the number of image pixel points. We then perform a matrix multiplication between the transpose of G_2 and G_1, that is,$(G_2^{C\times N})^T \times G_1^{C\times N}$. Finally, the obtained results are fed into a softmax layer to obtain the spatial correlation matrix $W(s) \in \mathbb{R}^{N\times N}$:

$$w_s^{ij} = \frac{exp(G_2^j \cdot G_1^i)}{\sum_i^N exp(G_2^j \cdot G_1^i)}, \tag{4}$$

where w_s^{ij} denotes the correlation measure between the j^{th} position and thei^{th} position in a feature map. we perform a matrix multiplication between G_1 and $W(s)$, that is, $G_1 \times W(s)$ and then reshape the result to $\mathbb{R}^{C\times H\times W}$ to obtain the position attention-aware maps $G(s)$. Next, we multiply the obtained result by a spatial scale parameter μ and perform an element-wise summation with G to obtain the result $G_s' \in \mathbb{R}^{C\times H\times W}$:

$$G_s' = \mu(G_1^{C\times N} \times ((G_2^{C\times N})^T \times G_1^{C\times N})) + G^{C\times H\times W}, \quad (5)$$

where the spatial scale parameter μ is a parameter learned gradually from 0.

In addition, each channel mapping of high-level features can be viewed as a class-specific response with different semantic interrelated. By exploiting the interdependencies between channel mappings, we can emphasize the interdependent feature mappings and improve the semantic-specific feature representation. Therefore, we present a channel attention-aware module to explicitly model the interdependencies between channels. We first feed $F \in \mathbb{R}^{C\times H\times W}$ that is the output of the feature pyramid fusion module into a convolution layer to get two feature maps F_1 and F_2, $\{F_1, F_2\} \in \mathbb{R}^{C\times H\times W}$. Secondly, we reshape them to $\mathbb{R}^{C\times N}$, where $N = H \times W$ denotes the number of pixels in an image. Further, we perform the multiplication of the matrix $F_2^{C\times N} \times (F_1^{C\times N})^T$. The global correlation matrix $W(c) \in \mathbb{R}^{C\times C}$ is then calculated via a softmax slayer, and it is defined as:

$$w_c^{ij} = \frac{exp(F_2^j \cdot F_1^i)}{\sum_i^C exp(F_2^j \cdot F_1^i)}, \quad (6)$$

where w_c^{ij} denotes the value of the weight of thej^{th} channel on the i^{th} channel. Then we perform the matrix multiplication $W(c) \times F_2$ and reshape it to $\mathbb{R}^{C\times H\times W}$. As a result, the channel attention feature map $F(c)$ is obtained. After that, we multiply the obtained result by a channel scale parameter ρ and perform an element-wise summation with F to obtain the result $F_c' \in \mathbb{R}^{C\times H\times W}$:

$$F_c' = \rho((F_2^{C\times N} \times (F_1^{C\times N})^T) \times F_2^{C\times N}) + F^{C\times H\times W}, \quad (7)$$

where ρ is the channel scale parameter learned gradually from 0. In (7), we can see that a feature map is obtained from the summation of the original features and the attention features. Such an approach preserves more fine-grained information and helps models to enhance the feature representation between channels.

The final feature maps via the feature attention-aware module are defined as:

$$B^{2C\times H\times W} = F_c' \oplus G_s', \quad (8)$$

where F_c' denotes the finally obtained channel correlation feature maps and G_s' denotes the spatial correlation feature maps, $\{F_c', G_s'\} \in \mathbb{R}^{C\times H\times W}$, $\oplus$ denotes concatenate operation. Furthermore, we perform dimensionality reduction using 1 × 1 convolution to achieve adaptive information fusion at different scales.

In summary, position attention captures the contextual relationships of global position in space, while channel attention models the global correlations between feature maps. The presented feature attention-aware module uses both channel and position attention, and employs an adaptive fusion of feature maps to reflect the semantic dependencies of channels and positions leading to the improvement of model robustness.

3 | EXPERIMENT

To evaluate the effectiveness of the proposed MFP-Net, five state-of-the-art network models MCNN [5], CSRNet [47], SFCN [11] and SFCN+ [11] are considered as comparative approaches in our experiments. Besides, all comparative networks and the proposed MFP-Net are performed on five popular datasets ShanghaiTech [5], NWPU-Crowd [7], UCF_CC_50 [59], UCF-QRNF [60] and GCC [8]. In this section, we illustrate the evaluation metrics and experimental details. Then we perform the ablation studies on the ShanghaiTech dataset. The experimental results of the proposed MFP-Net on other datasets are reported at last.

3.1 | Evaluation metrics

For crowd density estimation, two popular evaluation metrics are mean absolute error (MAE) and mean squared error (MSE), that is,

$$MAE = \frac{1}{N}\sum_{i=1}^{N} |Y_i - \hat{Y}_i|, \quad (9)$$

$$MSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N} |Y_i - \hat{Y}_i|^2}, \quad (10)$$

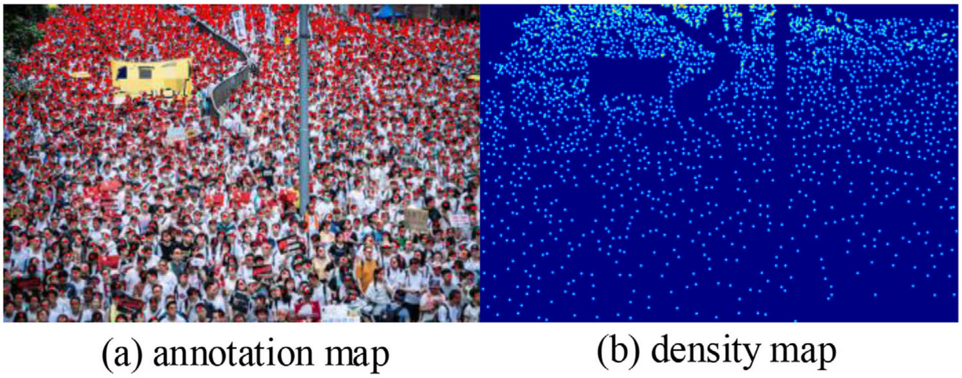
where N is the number of samples from the testing set, Y_i denotes the groundtruth crowd count, and $\hat{Y}_i$ is the predicted crowd count in the i^{th} test image. The predicted count is obtained by performing summation over the crowd density map output from a model. For MAE and MSE, if the values of MAE and MSE are smaller, then the test sample is closer to the groundtruth. To further evaluate the quality of the estimated density maps, the Peak Signal-to-Noise Ratio (PSNR) and the Structural SIMilarity (SSIM) are also used in our experiments.

3.2 | Experimental setup

To measure the error between the estimated density map and the groundtruth, we adopt pixel-wise mean square error (MSE) loss as the objective function. The optimization of the model parameters θ is defined as follows:

$$Loss(\theta) = \frac{1}{2B}\sum_{i=1}^{B} ||F_i^{GT} - \hat{F}_i^{PRE}||_2^2, \quad (11)$$

where B is the batch size, and F_i^{GT} is the groundtruth density map of the test image.



(a) annotation map

(b) density map

FIGURE 6 An annotation map and the corresponding groundtruth density map. (a) shows the head position marked by red points. (b) shows the density map

We adopt the same strategy as previous work [7, 36 47, 52 54] to generate groundtruth density maps. Specifically, each image is associated with a series of 2D points that are the positions of human heads in a crowd scene. We use Gaussian kernels to blur each head annotation. In our experiments, considering the spatial distribution of all images in each dataset, the density map F_i^{GT} generated by the Gaussian convolution with fixed size is defined as:

$$F_i^{GT} = \sum_{x_i \in P} \delta(x - x_i) \times G_{\sigma^2}(x), \qquad (12)$$

where x is the pixel position in the image, and x_i denotes the position of the$i$-th human head in the annotated map δ. $G_{\sigma^2}(x)$ is the Gaussian kernel and σ represents the standard deviation of Gaussian distribution. Respectively, we set the Gaussian kernel size to 15 and σ to 4 for all datasets for fair comparison. Figure 6 shows an annotation map and its density map.

The proposed MFP-Net is an end-to-end training framework. We fine-tune the first 10 layers of VGG16 using a pre-trained model and initialize the other layers with a Gaussian function with a standard deviation of 0.01. To increase the diversity of the data, we take a 0.5 probability of level flipping for data augmentation during the training. The MFP-Net is optimized using the Adam algorithm with a learning rate $lr = 1 \times 10^{-5}$, and it is implemented on an desktop with NVIDIA GTX2080 Ti GPU and the PyTorch 1.6.0 framework.

3.3 | Ablation studies on the ShanghaiTech dataset

The ShanghaiTech dataset [5] contains 1198 images where a total of 330,165 human heads are marked. The dataset is divided into two parts A and B. The Part A includes 482 crowded scene images with different resolutions, where 300 and 182 images are used for training and testing, respectively. For training conveniently, we randomly crop these images to some small-sized images of size 200 × 200. The Part B includes 716 sparse scene images, where 400 and 316 images are used for training and testing, respectively. All the image size is 1024 × 768 in the Part B.

TABLE 1 Ablation studies on ShanghaiTech dataset, the best values are bolded

	Part A		Part B	
Method	MAE	MSE	MAE	MSE
MFP-Net(A)	90.2	160.0	30.4	50.4
MFP-Net(B)	75.6	125.4	12.2	17.6
MFP-Net(C)	70.5	118.8	11.0	15.1
MFP-Net	**65.5**	**112.5**	**8.7**	**13.8**

In this paper, two contributions are highlighted, one is that the feature pyramid fusion module (FPFM) is used for feature integration after the feature extraction layer; The other is that the feature attention-aware module (FAAM) is integrated into the proposed MFP-Net to perform feature selection and avoid the influence of background noise. To demonstrate the effectiveness of the two contributions, we conducted comprehensive experiments on ShanghaiTech dataset. In Table 1, MFP-Net(A) means that vanilla convolution with a convolutional kernel of size 3 × 3 is used in FPFM, MFP-Net(B) means that dilated convolution with fixed dilation rate r=2 is used in FPFM, MFP-Net(C) means that the FAAM module is removed. MFP-Net is our proposed method where the dilation rate $r = (1, 2, 3, 4)$ in FPFM.

As shown in Table 1, The result of MFP-Net is MAE of 65.5 and MSE of 112.5, which is 24.7-point and 47.5-point improvement over the MFP-Net(A) on ShanghaiTech part A, respectively. On Part B, MFP-Net also achieves the best results: MAE of 8.7 and MSE of 13.8, which is 2.3-point and 1.3-point improvement over MFP-Net(C), respectively. The experimental results of MFP-Net outperformed MFP-Net(A) and MFP-Net(B), which demonstrates the effectiveness of the FPFM module. The experimental results of MFP-Net are better than that of MFP-Net(C), which demonstrates the effectiveness of the FAAM module. As can be seen that FPFM utilizes pyramid group convolution with variable dilation rate for multi-scale feature extraction, which shows better adaptability for crowd images. And FAAM effectively fuses local and global contextual information to enhance the feature representation of our model and thus improve the prediction accuracy. We compare the

TABLE 2 Comparison of MFP-Net with other methods on ShanghaiTech dataset, "None" means no pre-training, the best values are bolded

		Part A		Part B	
Method	Pre-training	MAE	MSE	MAE	MSE
MCNN	None	110.9	170.4	26.3	41.6
CSRNet	ImgNt	70.0	116.1	10.8	16.8
SFCN	ImgNt	70.5	117.0	11.4	17.2
SFCN+	ImgNt	68.1	113.3	9.1	15.4
MFP-Net	ImgNt	**65.5**	**112.5**	**8.7**	**13.8**

TABLE 3 Experimental results of different methods on NWPU-Crowd validation set, the best values are bolded

		NWPU-Crowd			
Method	Pre-training	MAE	MSE	PSNR	SSIM
MCNN	None	217.1	698.6	28.61	0.876
CSRNet	ImgNt	103.0	**433.8**	29.89	0.891
SFCN	ImgNt	106.2	615.1	29.95	0.929
SFCN+	ImgNt	95.0	587.4	30.57	0.950
MFP-Net	ImgNt	**90.3**	458.0	**30.61**	**0.955**

proposed MFP-Net with other state-of-the-art methods including MCNN [5], CSRNet [47] and SFCN [11] on ShanghaiTech dataset. SFCN uses VGG [61] as the backbone and SFCN+ uses ResNet101 [62] as the backbone. Note that we use a pre-trained model based on ImageNet Database [63] for parameter initialization. The experimental data are shown in Table 2.

In Table 2, the proposed MFP-Net achieves a 2.6-point improvement in MAE and a 0.8-point improvement in MSE over SFCN+ [11] on Part A. MFP-Net also provides better results for MSE of 8.7 and MAE of 13.8 for sparser scenes in Part B. It can be seen that our model has a better generalization ability to scenes of different scales, because FPFM with dilated convolution of variable dilation rate and FAAM with attention of two channels are able to extract and sense the features extracted from different receptive fields adaptively to achieve better results.

3.4 | Main comparisons

We compare our method with state-of-the-art methods on NWPU-Crowd dataset comprehensively. The NWPU-Crowd dataset [7] is the largest crowd counting and localization dataset available, with a total of 5109 images and 2,133,238 labeled instances, with the number of people in each image ranging from 0 to 20,033. For practical application to improve the generalization ability of the model, this dataset introduces 351 negative samples (namely nobody scenes), each of which has similar texture features as the crowded scenes. Secondly, due to the large variation of lighting and scene, the appearance of human head in each image varies greatly. In the data preprocessing stage, we first resize the high-resolution images to 2048-px scale with the original aspect ratio. We randomly cropped all images to smaller images of size 576 × 768 and flipped them horizontally to perform data augmentation during the training. To evaluate the quality of our method to generate density maps, we use two criteria PSNR and SSIM, and the results of our experiments on the validation set are shown in Table 3.

Table 3 shows that the value of MAE for MFP-Net is 90.3, with an improvement of 4.7-point over SFCN+. The value of MSE provided by MFP-Net is not the best because the crowd density and distribution varied significantly, and the nobody scenes are easy to confuse in the NWPU-Crowd dataset. Our proposed MFP-Net has the best PSNR of 30.61 and SSIM of 0.955, because its FAAM uses attention mechanism to reduce the effect of background noise by fully integrating the multi-scale contextual information in the crowd image. It is clear that SFCN+ provides higher values of PSNR and SSIM than both CSRNet and MCNN. The main reason is that SFCN+ employs a spatial encoder structure to enrich feature maps, but CSRNet and MCNN don't discriminate features in fusion stage.

To demonstrate the effectiveness of our method for density map estimation, we select five representative samples (e.g. scenes with nobody, severe occlusion, high density, and poor lighting conditions) and visualize the prediction results.

Figure 7 shows the comparative results of estimated density maps on the NWPU-Crowd dataset using different methods. The first column is a negative sample whose texture information is similar to that of the dense crowd. Since CSRNet directly fuses the extracted features without differentiation, resulting in poor prediction result as shown in the image at the third row and the first column. SFCN+ uses a spatial encoder structure to encoder the context information, which is effective for noise suppression leading to better result than CSRNet. The proposed MFP-Net uses FAAM to be aware of contextual multi-scale information, which can suppress background noise and improves the generalization ability of the model. It thus provides better result than SFCN+. In the second column, since CSRNet, SFCN and SFCN+ ignore some heavily occluded locations in images, they obtain poor prediction results. In contrast, MFP-Net sufficiently exploits the contextual relationships in crowd images, it thus achieves better prediction than CSR-Net and SFCN+. In the case of extremely poor lighting conditions, our model still achieves fine prediction results as shown in the third and fifth columns of Figure 7, which shows that our model has strong robustness. The fourth column is a highly congested scene and the proposed MFP-Net has a better accuracy of crowd counting than SFCN+.

It is clear that the proposed MFP-Net can provide good prediction results for various complex scenarios, since FPFM uses pyramid group dilated-convolution to capture multi-scale features and FAAM uses attention mechanism to fully integrate local and global information to improve feature representation.

In order to show the superiority of MFP-Net, Table 4 shows the comparison of the number of parameters and computational cost for different networks. In this experiment, the input data is an image of size 3 × 576 × 768. According to the experiments,

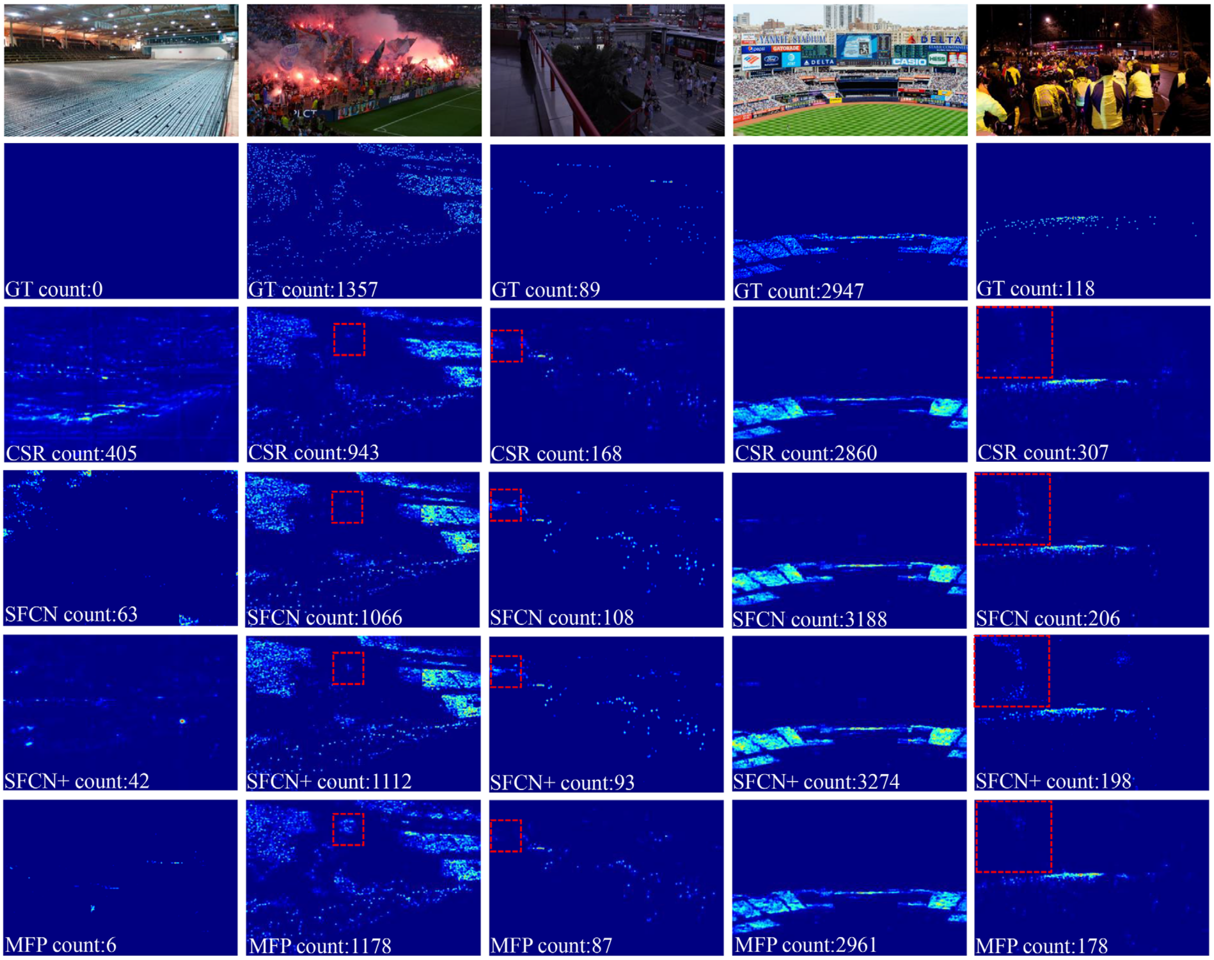


FIGURE 7 Five groups of visualization results on the NWPU-Crowd validation set. The first row is the original images, and the second row indicates the groundtruth density maps. The last four rows show the density maps and counts predicted by CSRNet, SFCN, SFCN+, and our method MFP-Net, respectively. "GT count" indicates the real counts in the image

TABLE 4 Comparison of the efficiency of different networks, the best values are bolded

Model	operations (GFLOPs)	parameters (M)	storage usage (MB)
CSRNet	182.82	16.26	62.05
SFCN	183.83	16.33	62.34
SFCN+	273.42	38.59	147.75
MFP-Net	**128.55**	**8.41**	**32.10**

the computational cost of our model is 128.55 GFLOPs, and the number of parameters is 8.41M, which is lower than CSRNet, SFCN and SFCN+ since our model adopts group convolution that is faster than the vanilla convolution. Note that MCNN is missed in Table 4 since the network does not use a backbone.

3.5 | Comparison in other datasets

To further validate the generalization ability of our model, we further conducted experiments on three popular datasets in this section.

The UCF_CC_50 dataset [59] has a limited sample of 50 images, which is extremely congested crowd counting dataset. To make a fair comparison, we follow the 5-fold cross-validation method in [59]. The images are randomly divided into 5 groups and each of them includes 10 images. In the experiment, four groups are used for training and one group is used for testing, so that there are 5 ways of training and testing, and we show their average values.

The UCF-QRNF dataset [60] has a large span of crowd density, where 1,201 and 334 images are used for training and testing, respectively. During the training, we randomly crop the images into 224 × 224 patches and take a horizontal flip with 0.5 probability for data augmentation.

The GTA5 Crowd Counting Dataset (GCC) [8] is a large-scale synthetic dataset (Synthetic Data) with different scenes and variable environmental conditions, which consists of 15,212 images with a resolution of 1080 × 1920. We randomly divided this dataset into two groups, the training set (75%), and the test set (25%).

Table 5 shows the experimental results for the three datasets. In Table 5, MFP-Net obtains MAE/MSE of 112.2/190.7 on the UCF-QRNF dataset containing a variety of scenes,

TABLE 5 Experimental results of different methods on several mainstream datasets, the best values are bolded

Method	UCF_CC_50		UCF-QRNF		GCC (RS)	
	MAE	MSE	MAE	MSE	MAE	MSE
MCNN	376.6	508.0	276.5	441.2	101.0	216.5
CSRNet	265.5	394.2	121.3	208.0	38.5	86.6
SFCN	266.4	396.7	135.1	239.8	36.1	81.0
SFCN+	245.3	**375.8**	114.5	193.6	28.8	71.2
MFP-Net	**240.8**	384.4	**112.0**	**190.7**	**28.2**	**70.1**

and MAE/MSE of 28.2/70.1 on the largest dataset GCC. However, the prediction results on UCF_CC_50 dataset are not optimal due to imbalanced samples and small number of images in the dataset. The experiments show that our method MFP-Net performs well on most of crowd scenes, but it does not perform well for some datasets due to a small number of training samples.

4 | SUMMARY AND DISCUSSION

In this paper, we have proposed a multi-scale feature pyramid network (MFP-Net) and applied it to the task of crowd density estimation. MFP-Net is different from current models due to the introduction of a feature pyramid fusion module and a feature attention-aware module. The feature pyramid fusion module can effectively extract different levels of fine-grained information in images by using dilated convolution with variable dilation rate while ensuring that the resolution of the feature map is not degraded, and the training efficiency is further improved due to the employment of parallel group convolution. The feature attention-aware module can improve the robustness of MFP-Net by adaptively extracting local and global contextual information and focusing on important spatial locations and channels. According to the experiment results, MFP-Net has clear advantages for crowd density estimation and crowd counting in highly congested and noisy environments. In future work, we will investigate how to adaptively adjust the dilation rate and the number of groups according to the complexity of the data to reduce the computational cost and improve the prediction accuracy of the model.

ACKNOWLEDGEMENTS

This work was supported in part by the National Natural Science Foundation of China under Grant nos. 61871259, 61861024, and 61701387, and in part by Natural Science Basic Research Program of Shaanxi (No. 2021JC-47), Science and Technology Program of Shaanxi Province of China (No. 2020NY-172).

PERMISSION TO REPRODUCE MATERIALS FROM OTHER SOURCES

None

ORCID

Tao Lei https://orcid.org/0000-0002-2104-9298
Asoke K. Nandi https://orcid.org/0000-0001-6248-2875

REFERENCES

1. Gao, G., et al.: CNN-based density estimation and crowd counting: A survey. arXiv preprint arXiv:2003.12783 (2020)
2. Sam, D.B., et al.: Switching convolutional neural network for crowd counting. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4031–4039. IEEE, Piscataway (2017)
3. Zeng, L., et al.: Multi-scale convolutional neural networks for crowd counting. In: 2017 IEEE International Conference on Image Processing (ICIP), pp. 465–469. IEEE, Piscataway (2017)
4. Zhang, C., et al.: Cross-scene crowd counting via deep convolutional neural networks. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 833–841. IEEE, Piscataway (2015)
5. Zhang, Y., et al.: Single-image crowd counting via multi-column convolutional neural network. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 589–597. IEEE, Piscataway (2016)
6. Sindagi, V.A., Patel, V.M.: CNN-based cascaded multi-task learning of high-level prior and density estimation for crowd counting. In: 2017 IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), pp. 1–6. IEEE, Piscataway (2017)
7. Wang, Q., et al.: NWPU-crowd: A large-scale benchmark for crowd counting and localization. IEEE Trans. Pattern Anal. Mach. Intell. (2020). https://doi.org/10.1109/TPAMI.2020.3013269
8. Wang, Q., et al.: Learning from synthetic data for crowd counting in the wild. In: 2019 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 8198–8207. IEEE, Piscataway (2019)
9. Wang, Q., et al.: Density-aware curriculum learning for crowd counting. IEEE Trans. Cybern. (2020)
10. Gao, J., Wang, Q.: Feature-aware adaptation and density alignment for crowd counting in video surveillance. IEEE Trans. Cybern. (2020)
11. Wang, Q., et al.: Pixel-wise crowd understanding via synthetic data. Int. J. Comput. Vis. 129, 225–245 (2021)
12. Kang, K., Wang, X.: Fully convolutional neural networks for crowd segmentation. arXiv preprint arXiv:1411.4464 (2014)
13. Zhao, T., Nevatia, R.: Bayesian human segmentation in crowded situations. In: 2003 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. II–459. IEEE, Piscataway (2003)
14. Dong, L., et al.: Fast crowd segmentation using shape indexing. 2007 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1–8. IEEE, Piscataway (2007)
15. Ali, S., Shah, M.: A lagrangian particle dynamics approach for crowd flow segmentation and stability analysis. In: 2007 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1–6. IEEE, Piscataway (2007)
16. Yuan, Y., et al.: Tracking as a whole: Multi-target tracking by modeling group behavior with sequential detection. IEEE trans. Intell. Transp. Syst. 18(12), 3339–3349 (2017)
17. Li, X., et al.: A multiview-based parameter free framework for group detection. In: AAAI Conference on Artificial Intelligence. AAAI Press, Palo Alto (2017)
18. Wang, Q., et al.: Detecting coherent groups in crowd scenes by multiview clustering. IEEE Trans. Pattern Anal. Mach. Intell. 42(1), 46–58 (2018)
19. Liu, W., et al.: Data fusion based two-stage cascade framework for multimodality face anti-spoofing. IEEE Trans. Cogn. Develop. Syst. (2021)
20. Mehran, R., et al.: Abnormal crowd behavior detection using social force model. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 935–942. IEEE, Piscataway (2009)
21. Popoola, O.P., Wang, K.: Video-based abnormal human behavior recognition–A review. IEEE Trans. Syst. Man Cybern. Syst. 42(6), 865–878 (2012)
22. Onoro-Rubio, D., López-Sastre, R.J.: Towards perspective-free object counting with deep learning. In: European Conference on Computer Vision (ECCV), pp. 615–629. Springer, Berlin (2016)
23. Zhang, S., et al.: FCN-rLSTM: Deep spatio-temporal neural networks for vehicle counting in city cameras. In: 2017 IEEE International Conference on Computer Vision (ICCV), pp. 3667–3676. IEEE, Piscataway (2017)
24. Arteta, C., et al.: Counting in the wild. In: European Conference on Computer Vision (ECCV), pp. 483–498. Springer, Berlin (2016)

25. Sirinukunwattana, K., et al.: Locality sensitive deep learning for detection and classification of nuclei in routine colon cancer histology images. IEEE Trans. Med. Imaging 35(5), 1196–1206 (2016)
26. Lempitsky, V., Zisserman, A.: Learning to count objects in images. In: Advances in Neural Information Processing Systems (NIPS), pp. 1324–1332. Curran Associates, Inc., Red Hook (2010)
27. Topkaya, I.S., et al.: Counting people by clustering person detector outputs. In: 2014 IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), pp. 313–318. IEEE, Piscataway (2014)
28. Felzenszwalb, P.F., et al.: Object detection with discriminatively trained part-based models. IEEE Trans. Pattern Anal. Mach. Intell. 32(9), 1627–1645 (2009)
29. Wu, B., Nevatia, R.: Detection of multiple, partially occluded humans in a single image by Bayesian combination of edgelet part detectors. In: 2005 IEEE International Conference on Computer Vision (ICCV), pp. 90–97. IEEE, Piscataway (2005)
30. Wang, M., Wang, X.: Automatic adaptation of a generic pedestrian detector to a specific traffic scene. In: 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3401–3408. IEEE, Piscataway (2011)
31. Dollar, P., et al.: Pedestrian detection: An evaluation of the state of the art. IEEE Trans. Pattern Anal. Mach. Intell. 34(4), 743–761 (2011)
32. Chen, K., et al.: Feature mining for localised crowd counting. In: Proceedings of the British Machine Vision Conference (BMVC), p. 3. Springer, London (2012)
33. Change Loy, C., et al.: From semi-supervised to transfer counting of crowds. In: 2013 IEEE International Conference on Computer Vision (ICCV), pp. 2256–2263. IEEE, Piscataway (2013)
34. Chan, A.B., Vasconcelos, N.: Bayesian poisson regression for crowd counting. In: 2009 IEEE International Conference on Computer Vision (ICCV), pp. 545–551. IEEE, Piscataway (2009)
35. Ryan, D., et al.: Crowd counting using multiple local features. In: 2009 Digital Image Computing: Techniques and Applications, pp. 81–88. IEEE, Los Alamitos (2009)
36. Chan, A.B., et al.: Privacy preserving crowd monitoring: Counting people without people models or tracking. In: 2008 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1–7. IEEE, Piscataway (2008)
37. Lempitsky, V., Zisserman, A.: Learning to count objects in images. Advances in Neural Information Processing Systems (NIPS), pp. 1324–1332. Curran Associates, Red Hook (2010)
38. Pham, V.Q., et al.: Count forest: Co-voting uncertain number of targets using random forest for crowd density estimation. In: 2015 IEEE International Conference on Computer Vision (ICCV), pp. 3253–3261. IEEE, Piscataway (2015)
39. Fu, M., et al.: Fast crowd density estimation with convolutional neural networks. Eng. Appl. Artif. Intell. 43, 81–88 (2015)
40. Long, J., et al.: Fully convolutional networks for semantic segmentation. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3431–3440. IEEE, Piscataway (2015)
41. Lei, T., et al.: Adaptive morphological reconstruction for seeded image segmentation. IEEE Trans. on Image Process 28(11), 5510–5523 (2019)
42. Lei, T., et al.: Superpixel-based fast fuzzy C-means clustering for color image segmentation. IEEE Trans. on Fuzzy Syst. 27(9), 1753–1766 (2019)
43. Lei, T., et al.: DefED-Net: Deformable encoder-decoder network for liver and liver tumor segmentation. IEEE Trans. on Radia. and Plasma Med. Sci. (2021)
44. Sindagi, V.A., Patel, V.M.: Generating high-quality crowd density maps using contextual pyramid CNNs. In: 2017 IEEE International Conference on Computer Vision (ICCV), pp. 1861–1870. IEEE, Piscataway (2017)
45. Gao, J., et al.: PCC net: Perspective crowd counting via spatial convolutional network. IEEE Trans. Circuits. Syst. Video Technol. 30(10), 3486–3498 (2019)
46. Wang, Q., et al.: Neuron linear transformation: Modeling the domain shift for crowd counting. IEEE Trans. Neural Netw. Learn.Syst. (2020), 1–13. https://doi.org/10.1109/TCYB.2020.3033428
47. Li, Y., et al.: Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes. In: 2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1091–1100. IEEE, Piscataway (2018)
48. Jiang, X., et al.: Crowd counting and density estimation by trellis encoder-decoder networks. In: 2019 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6133–6142. IEEE, Piscataway (2019)
49. Shi, Z., et al.: Crowd counting with deep negative correlation learning. In: 2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5382–5390. IEEE, Piscataway (2018)
50. Liu, W., et al.: Context-aware crowd counting. In: 2019 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5099–5108. IEEE, Piscataway (2019)
51. Liu, N., et al.: Adcrowdnet: An attention-injective deformable convolutional network for crowd understanding. In: 2019 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3225–3234. IEEE, Piscataway (2019)
52. Guo, D., et al.: Dadnet: Dilated-attention-deformable convnet for crowd counting. In: ACM International Conference on Multimedia, pp. 1823–1832. ACM, New York (2019)
53. ossain, M., et al.: Crowd counting using scale-aware attention networks. In: 2019 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 1280–1288. IEEE, Piscataway (2019)
54. Wei, B., et al.: MSPNET: Multi-supervised parallel network for crowd counting. In: 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2418–2422. IEEE, Piscataway (2020)
55. Gao, J., et al.: SCAR: Spatial-/channel-wise attention regression networks for crowd counting. Neurocomputing 363, 1–8 (2019)
56. Fu, J., et al.: Dual attention network for scene segmentation. In: 2019 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3146–3154. IEEE, Piscataway (2019)
57. Duta, I.C., et al.: Pyramidal convolution: Rethinking convolutional neural networks for visual recognition. arXiv preprint arXiv:2006.11538 (2020)
58. Sindagi, V.A., Patel, V.M.: Inverse attention guided deep crowd counting network. In: 2019 IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), pp. 1–8. IEEE, Piscataway (2019)
59. Idrees, H., et al.: Multi-source multi-scale counting in extremely dense crowd images. In: 2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2547–2554. IEEE, Piscataway (2013)
60. Idrees, H., et al.: Composition loss for counting, density map estimation and localization in dense crowds. In: European Conference on Computer Vision (ECCV), pp. 532–546. Springer, Berlin (2018)
61. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
62. He, K., et al.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778. IEEE, Piscataway (2016)
63. Deng, J., et al.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 248–255. IEEE, Piscataway (2009)
64. Liu, X., et al.: Leveraging unlabeled data for crowd counting by learning to rank. In: 2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7661–7669. IEEE, Piscataway (2018)
65. Cao, X., et al.: Scale aggregation network for accurate and efficient crowd counting. In: European Conference on Computer Vision (ECCV), pp. 734–750. Springer, Berlin (2018)
66. Nam, H., Han, B.: Learning multi-domain convolutional neural networks for visual tracking. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4293–4302. IEEE, Piscataway (2016)
67. Sindagi, V., Yasarla, R., Patel, V. M.: Jhu-crowd++:Large-scale crowd counting dataset and a benchmark method. IEEE Trans. Pattern Anal. Mach. Intell. (2020). https://doi.org/10.1109/TPAMI.2020.3035969

How to cite this article: Lei, T., et al.: MFP-Net: Multi-scale feature pyramid network for crowd counting. IET Image Process. 2021;15:3522–3533. https://doi.org/10.1049/ipr2.12230