Applied Intelligence
https://doi.org/10.1007/s10489-022-03639-5

Cascaded parallel crowd counting network with multi-resolution collaborative representation

Lei Lyu[1,2] · Run Han[1,2] · Ziming Chen[3,4]

Accepted: 14 April 2022

Abstract

Accurately estimating the size and density distribution of a crowd from images is of great importance to public safety and crowd management during the COVID-19 pandemic, but it is very challenging as it is affected by many complex factors, including perspective distortion and background noise information. In this paper, we propose a novel multi-resolution collaborative representation framework called the cascaded parallel network (CP-Net), consisting of three parallel scale-specific branches connected in a cascading mode. In the framework, the three cascaded multi-resolution branches efficiently capture multi-scale features through their specific receptive fields. Additionally, multi-level feature fusion and information filtering are performed continuously on each branch to resist noise interference and perspective distortion. Moreover, we design an information exchange module across independent branches to refine the features extracted by each specific branch and deal with perspective distortion by using complementary information of multiple resolutions. To further improve the robustness of the network to scale variance and generate high-quality density maps, we construct a multi-receptive field fusion module to aggregate multi-scale features more comprehensively. The performance of our proposed CP-Net is verified on the challenging counting datasets (UCF_CC_50, UCF-QNRF, Shanghai Tech A&B, and WorldExpo'10), and the experimental results demonstrate the superiority of the proposed method.

Keywords Crowd counting · Density map estimation · Cascaded multi-resolution CNN · Multi-scale fusion

1 Introduction

In large public places, such as sporting venues, train stations, business districts, and tourist attractions, thousands of people often gather in a fixed area [52]. In extremely dense crowds, there is so much shoving and jostling that people's movements are no longer entirely under their own control, which is liable to cause safety accidents, resulting in large casualties and social impacts [21]. To avoid this scenario, it is necessary to keep the crowd size within a reasonable scope by counting the flow of people in advance. In the early days, the task relied heavily on manpower. However, it greatly increases the human costs when the crowd size is large enough [4]. As a consequence, many scholars have attempted to apply computer vision techniques to crowd counting, aiming to automatically estimate crowd size in highly complicated unconstrained scenes.

Initially, the focus was on crowd counting using detection and regression [19]. Detection-based crowd counting methods utilize support vector machines (SVMs) and boosting for sparse crowds. In highly dense crowds, the individuals may occlude each other, seriously affecting counting accuracy. Gradually, regression-based methods that avoid solving the hard detection problem have become mainstream and have achieved great improvement. In particular, density regression-based methods can localize the crowd in density maps generated by pixel-wise regression. Then the crowd count is calculated as the integral of the density map [16].

Motivated by the recent successful use of convolutional neural networks (CNNs) in semantic segmentation [13, 27] and visual saliency [31, 50], CNN-based methods [8, 32,

✉ Lei Lyu
lvlei@sdnu.edu.cn

[1] School of Information Science and Engineering, Shandong Normal University, Jinan, 250358, China

[2] Shandong Provincial Key Laboratory for Distributed Computer Software Novel Technology, Jinan 250358, China

[3] Shandong Zhengzhong Information Technology Co., LTD, Jinan 250014, China

[4] Shandong Digital Applied Science Research Institute Co.,LTD, Jinan 250101, China

Published online: 19 May 2022

38, 41] have been introduced to address crowd counting in dense scenes. Despite many significant achievements, crowd counting is still limited by several factors, such as background clutter, heavy occlusions and perspective distortion. Among them, perspective distortion is the issue that has aroused the most concern of researchers in recent literature [10, 15], which embodied similar individuals in different relative locations that vary greatly in size [17] (shown in Fig. 1).

To address this problem, numerous methods [5, 10, 37, 51] have concentrated on remedying the scale variance by employing a multi-column architecture with various receptive fields. Although these approaches ease the scale issue to some extent, they are limited by several drawbacks. First, each branch usually extracts multi-scale features independently, resulting in discontinuous information extraction [24]. Second, due to the structural similarity of each parallel branch, the extracted features are nearly similar [20]. Furthermore, as the network deepens, the scope of the receptive field accumulates from shallow to deep, easily leading to the loss of spatial details. Similarly, background information also has a significant impact on the counting, as some human-shaped noise can easily be wrongly identified as positive.

Considering the above concerns, we present a novel multi-resolution collaborative representation framework called cascaded parallel network (CP-Net). In the cascaded parallel framework, the network starts from a simple front-end module for low-level information extraction and gradually adds multi-resolution parallel branches sensitive to multi-scale features to constitute the subsequent three feature extraction stages. At the back end of each stage, the parallel multi-resolution subnetworks mutually exchange information, which utilizes the complementary information of different-scale features to refine the scale-specific features. Moreover, the feature extraction block constituting each branch realizes the repeated fusion of features and the intentional suppression of specific channels, which can solve the continuous change in scale and the interference of background information. To more fully aggregate multi-scale features, a multi-receptive field fusion module is designed at the back end.

In summary, the contributions of our work are fourfold.

(1) We design a cascaded parallel network (CP-Net) with multi-resolution collaborative representation for crowd counting to better remedy the continuous scale variance and filter noise in unconstrained scenes;
(2) We construct a cross-branch information exchange module to mutually refine the scale-specific features by utilizing complementary information between multi-scale features;
(3) We construct a multi-receptive field fusion module at the back end of the network to further enhance the robustness to scale variance;
(4) Extensive experiments on four benchmarks demonstrate the superiority of the proposed CP-Net in crowd counting.

2 Related work

Crowd counting was regarded as a detection problem early, while the detection-based methods performed poorly in crowded scenes. Gradually, regression-based methods were used for better counting performance. Recently, density estimation methods based on CNNs have become mainstream. In this section, we briefly review the work most relevant to our work, which includes two aspects: multi-column-based methods and multi-scale fusion methods.

2.1 Multi-column architecture for crowd counting

Most multi-column methods are designed to capture multi-scale information employing columns of different receptive fields, and many are designed for multi-task learning. The pioneering work is the multi-column convolutional neural network (MCNN) [51], which aims at solving the scale problem by using three similar branches with different kernel sizes. Switch-CNN [33] uses three CNN regressors

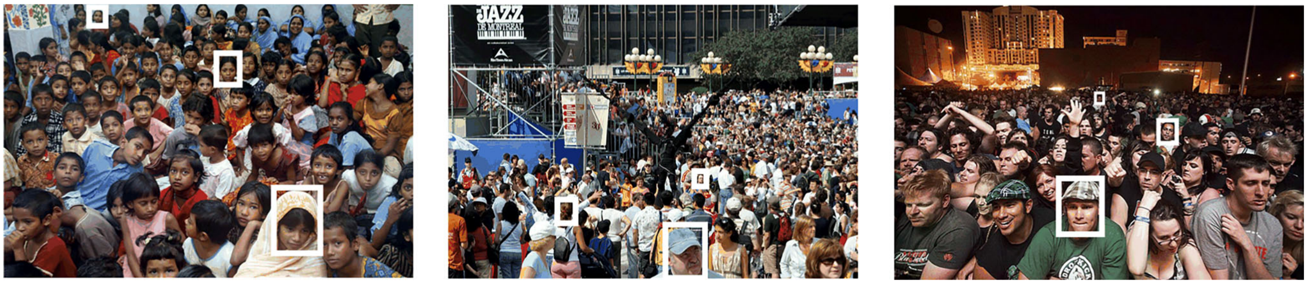


Fig. 1 One of the most challenging issues in crowd counting is perspective distortion, which is caused by the different distances from each person to the camera. In addition, such scenes are often accompanied by interfering complex backgrounds

similar to MCNN and trains a switch classifier to relay the image patch to the optimal regressor. CrowdNet [1] designs subnetworks of different depths to simultaneously extract features of different levels to achieve multi-scale feature fusion.

Furthermore, the contextual pyramid convolutional neural network (CP-CNN) [37] utilizes adversarial learning methods and combines global and local contextual information to produce high-quality density maps. The scale-aware attention network (SAAN) [10] adds a visual-attention mechanism to the CP-CNN that automatically selects between global images and local contextual information. The context-aware network (CANet) [24] employs four different average pooling branches to extract the context information of different receptive fields to improve network performance. MMNet [3] proposes a scale-aware framework that captures scale information through parallel filters of different sizes and supervises multi-scale fusion using multi-layer spatial information.

The attention scaling network (ASNet) [17] designs two subnetworks based on VGG16 [36] to handle the uneven distribution of crowds. The two subnetworks learn attention masks and scaling factors to assist in the generation of high-quality density maps. The perspective crowd counting network (PCC Net) [6] proposes a three-branch multi-task architecture in which the density map estimator, density classifier and fore/background segmentation cooperate to generate the final density map.

The pyramid-dilated deep convolutional neural network (PDD-CNN) [42] proposes a pyramid dilated module, which extracts scale information through parallel convolution with different dilation rates. Similarly, the self-attention residual network (SARNet) [26] adopts a multi-scale convolutional module, which has a multi-branch structure and employs dilated convolution of different kernel sizes to extract differential scale information. Moreover, RGBT-CC [22] extracts optical information and thermal information using two modal-specific branches and aggregates the multi-modal information using a modal-shared branch.

2.2 Multi-scale fusion for crowd counting

These methods aim to better deal with scale variations using various multi-scale information fusion schemes. Some methods combine multi-scale features extracted from different depths of deep networks to solve the scale problem. Scale-adaptive CNN (SaCNN) [49] deploys a single-column CNN, which adapts the feature maps extracted from layers of different depths to the same sizes and then combines them to generate the final density map. Because SaCNN adapts to scale variations by fusing multi-layer features, it does not bring more parameters to the model. Based on VGG16, the congested scene recognition network (CSRnet) [20] first introduces dilated convolution in the counting field to obtain larger receptive fields. Later, DUBNet [29] proposes a scalable framework that uses a ResNet-based front-end network to extract features and a back-end network composed of dilated convolution to deliver a larger receptive field. Due to the shortcut structure in ResNet-50 [9], frequent multi-level feature fusion is realized in the front-end network.

Other methods deploy multi-branch architectures and use branches of different receptive fields to obtain multi-scale information. The scale-aware attention network (SAAN) [10] simultaneously exploits three subnetworks to extract multi-scale features and generate attention maps, and then generates feature maps by combining global and local attention. The shallow feature based dense attention network (SDANet) [28] extracts shallow features through a low-level feature extractor (LFE) and captures multi-scale information through dense connections of hierarchical features. Specifically, LFE is composed of dilated convolution kernels in different receptive fields, which is a variant of the inception block [39].

With the increasingly successful application of encoder-decoder networks in computer vision, the scale aggregation network (SANet) [2] also adopts an encoder-decoder architecture in which scale aggregation modules based on the inception block are designed to extract multi-scale features and a set of transposed convolutions are employed to restore the resolution of the feature map. The trellis encoder-decoder network (TEDnet) [16] performs hierarchical aggregation of features at multiple decoding stages and promotes multi-scale feature fusion through dense jumping connections of cross paths.

Moreover, some works, such as the perspective-aware convolutional neural network (PACNN) [34] and perspective-guided convolution network (PGCNet) [45], fuse the multi-scale features using the reconstructed perspective map of the scene. The reverse perspective network (RPNet) [46] distorts the image based on the estimated perspective map so that the people in the image have similar scales. However, due to image deformation, the estimated location of the crowd is not spatially accurate. Some works, such as L2SM [44], rescale the input image according to the predicted density level to deal with the scale problem. However, the density level is determined according to the number of people, which cannot accurately represent the crowd scale.

3 Our approach

In this paper, we propose a cascaded parallel network with multi-resolution collaborative representation for crowd counting (detailed in Fig. 2). Specifically, multiple resolution-specific

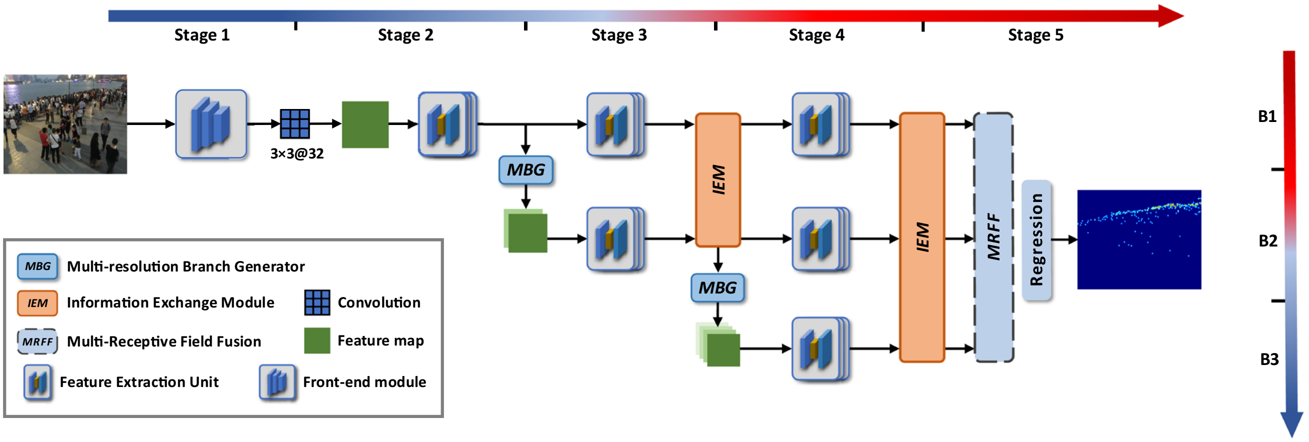


Fig. 2 The overall architecture of the Cascaded Parallel Network (CP-Net) with multi-resolution collaborative representation. B1, B2, and B3 represent branch1, branch2 and branch3, respectively. The proposed CP-Net can be divided into three parallel branches, which are connected by MBG and communicate through IEM. The parameters of convolution are represented as (kernel size) × (kernel size) @ (channel number)

branches, two information exchange modules and a multi-receptive field fusion module are incorporated to handle crowd counting in complex scenes. From the perspective of network architecture, CP-Net is a grid-shaped network that can be described in two dimensions: horizontal and vertical.

3.1 Overall of the CP-Net

From the horizontal, CP-Net consists of five stages from Stage 1 to Stage 5. Stage 1 contains a front-end module composed of the first seven layers of VGG16, which is used to extract low-level features of the input image. Each stage from Stage 2 to Stage 4 contains one or more feature extraction blocks composed of three Feature Extraction Units (FEUs). Specifically, due to the well-designed structure of FEU, feature extraction blocks can realize multi-level fusion and noise filtering while extracting high-level features. There are six feature extraction blocks in total, so eighteen multi-level fusions are conducted. Moreover, Stages 2 and 3 each contain a Multi-resolution Branch Generator (MBG) and an Information Exchange Module (IEM). As a cross-branch bridge, the IEM enables the features of each branch to receive complementary information from different resolutions. Stage 5 consists of two modules: the Multi-Receptive Field Fusion module (MRFF) and the regression layer, which aims to enhance the robustness to scale variance and generate high-quality density maps by fully aggregating multi-scale information.

From the vertical, the proposed CP-Net is divided into three parallel branches that are connected by MBGs. Specifically, branch 2 is generated by branch 1, and branch 3 is derived from branch 2. Branch 1 maintains the resolution of the input feature map, while branch 2 and branch 3 decrease successively in resolution and length. Each branch continuously performs multi-level feature fusion and noise filtering to reduce the adverse effect on counting. Furthermore, independent branches deliver complementary information through IEM to refine the scale-specific features. Ultimately, the three parallel branches are successively merged into the MRFF and regression layer, and high-quality density maps are output.

3.2 Feature extraction and branch generation

The details of the FEU are shown in Fig. 3. Each FEU contains a sequence of multiple operations for higher-level semantic information extraction and noise filtering, and a skip connection for multi-level information fusion. Specifically, the sequence consists of a 3 × 3 depthwise convolution (DWConv), an Information Filtering Module (IFM) and a 3 × 3 normal convolution. DWConv adopts channel sparse connections, which convolves each channel of the input feature map independently without interaction between channels. [7] proved that different channels in the feature map contain different feature information, so an information filtering module inspired by [11] intentionally follows the DWConv to suppress channels containing more noise information.

The IFM is implemented as follows. Global average pooling (GAP) is performed on the input feature map to transform the global information of each channel into the corresponding descriptor, followed by a sequence of operations {FC, ReLU, FC} to learn the interdependencies between channels, where FC indicates a fully connected layer. Then, a sigmoid function is employed as a gating mechanism to generate filtering factors corresponding to

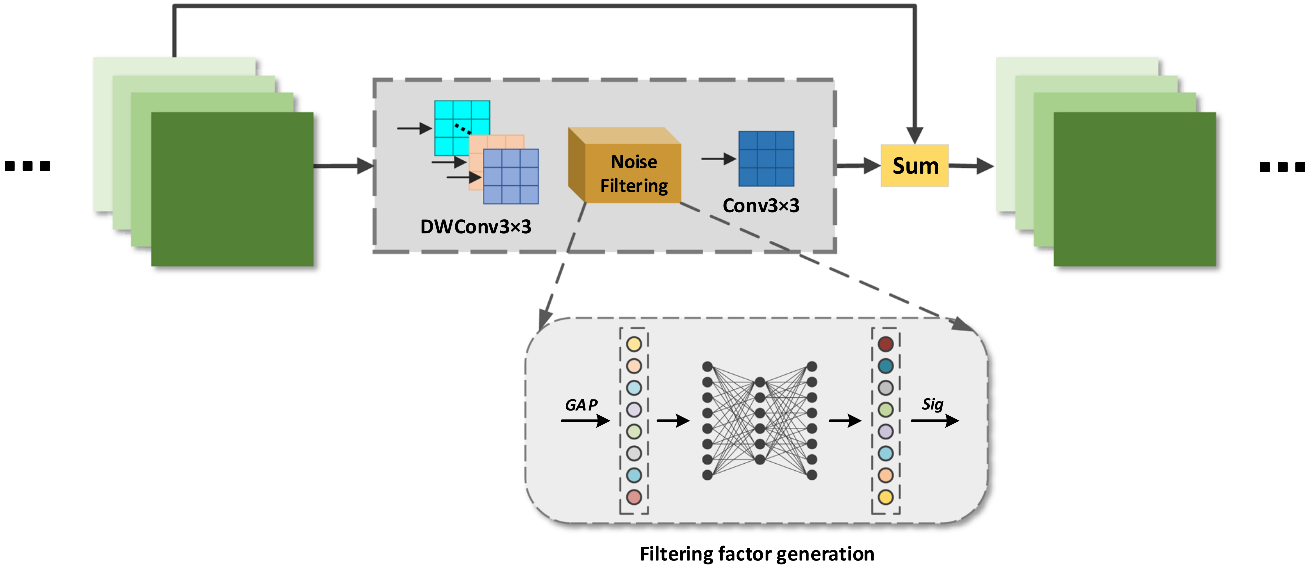


Fig. 3 Detailed configuration of the Feature Extraction Unit. The combination of DWConv and IFM can effectively filter the background noise, and the skip connection makes the network more adaptable to scale variations through multi-level fusion

each channel. Finally, selective information filtering can be realized by multiplying the filtering factor by the corresponding channel.

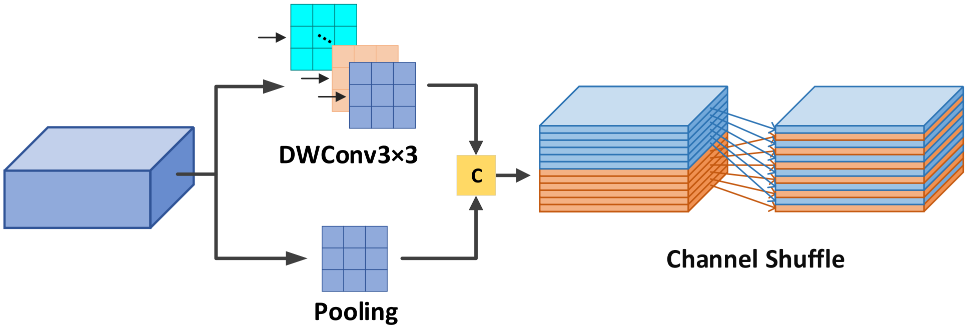
Then, a normal 3 × 3 convolution is employed to fuse the filtered independent channels and further extract higher-level semantic information. Subsequently, the output feature map is fused with the input feature map through an element-wise sum operation to achieve continuous multi-level feature fusion on a single branch.

Furthermore, we design the MBG to ensure that each branch has a different resolution and to minimize information loss. As shown in Fig. 2, the MBG is located at the front of branch 2 and branch 3 to generate and connect low-resolution branches. The structure of MBG is detailed in Fig. 4. We employ a depthwise convolution and max pooling to downsample the feature map, where the step of both is 2. Then, the two low-resolution feature maps are concatenated as the input of the new low-resolution branch. Since we perform two different downsampling operations on the original feature map, the number of channels in the new feature map is doubled compared with the original feature map, which can effectively alleviate the information loss caused by downsampling. Moreover, considering that the two parts of the low-resolution feature map are obtained independently, we shuffle the channels to promote the interaction of information between channels, which is conducive to the subsequent effective fusion between multi-level features.

3.3 Information exchange module

The features of different scales contain rich complementary information [23], which can be utilized to handle scale variations. To fully capture the complementarities between different scales, we design the Information Exchange Module (IEM), which performs information exchange across parallel branches rather than directly fusing features. In this process, the scale-specific feature map is supplemented by the different-scale feature maps of other resolution-specific branches, so that the features contain richer information and

Fig. 4 An illustration of the Multi-resolution Branch Generator. Two downsampling operations are adopted to double the channels of the newly generated low-resolution feature map. Furthermore, channel shuffling is conducted on the two spliced independent feature maps



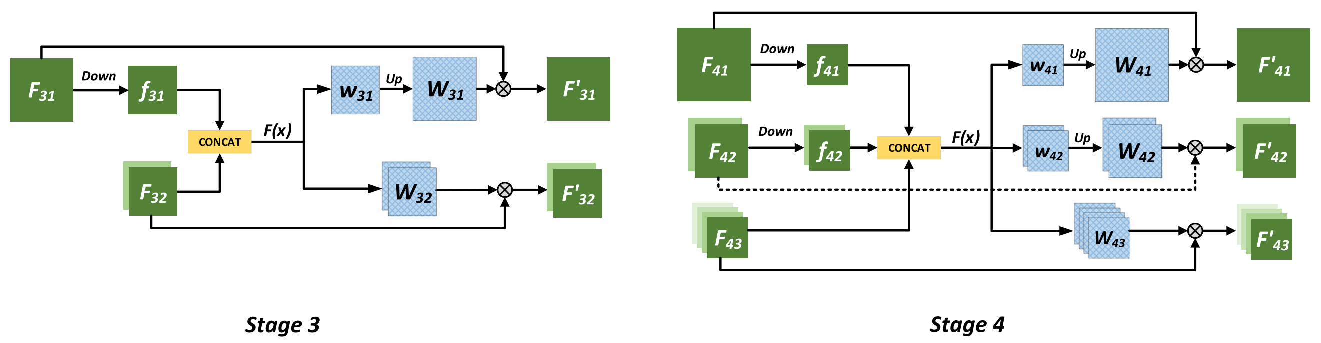


Fig. 5 The details of the Information Exchange Module. F_{SB} and W_{SB} represent the feature map and its corresponding weight map of Branch B in Stage S, respectively. f_{SB} and w_{SB} represent the feature map and its corresponding weight map after downsampling of F_{SB}. F'_{SB} represents the refined feature map after the information exchange

avoid excessive mixing of multi-scale features. We illustrate the IEM in Fig. 5.

In Stage 3, there are two parallel multi-resolution branches. First, F_{31} is subjected to 2 × 2 max pooling to align the size of F_{32}. Then, the aligned feature map is concatenated with F_{32}, and the weight map of each channel is generated through an information exchange function $F(x)$. Next, we perform an upsampling operation on the generated weight map to align the size of F_{31}. Subsequently, we multiply the feature map by the corresponding weight map.

Similarly, there are three parallel branches in Stage 4. Pool F_{41} and F_{42} to the size of F_{43}, and concatenate the three feature maps. Then, the weight maps generated by the function $F(x)$ are upsampled to the size of the corresponding input feature map, and matrix multiplication is performed. Finally, F'_{41}, F'_{42} and F'_{43} are taken as the inputs for Stage 5.

The information exchange function $F(x)$ is implemented as follows. The aligned feature maps concatenated together are followed by a sequence of 1 × 1 convolution, ReLU function, and 1×1 convolution, where 1×1 convolutions are employed to gather information from different channels and resolutions and the ReLU function is employed to eliminate negative values. Then, after a sigmoid function, the weight maps are obtained. The formal expression of this process is as follows.

$$\{W_1, \ldots, W_n\} = S\left(C_2\left(R\left(C_1\left(\{F_1, \ldots, F_n\}\right)\right)\right)\right) \tag{1}$$

where $\{F_1, \ldots, F_n\}$ and $\{W_1, \ldots, W_n\}$ represent n groups of aligned multi-scale feature maps and their corresponding weight maps, respectively. C_1 and C_2 indicate two 1 × 1 convolutional layers. S and R represent the sigmoid function and ReLU function, respectively.

Each weight value on the weight map receives information from different scales and different channels. By multiplying the weight map and feature map of each scale, the cross-scale and cross-channel information exchange is realized. The strategy of multi-scale feature communication adopted in this module allows the continuous exchange of information across the parallel branches, making the features extracted from each branch more comprehensive. Thus, the information exchange module can refine the resolution-specific features and obtain rich spatial information.

3.4 Multi-receptive field fusion module

Through the information exchange module, we can acquire multi-resolution feature maps rich in more complete spatial and semantic information. To further enhance the robustness to scale variance, a Multi-Receptive Field Fusion module (MRFF) is designed to aggregate the refined multi-scale features, which is illustrated in Fig. 6.

We adopt bilinear interpolation to upsample the low-resolution feature maps to align the feature maps. In this process, some redundant information will inevitably be induced in the feature map. In view of this situation, we employ three dilated convolutions with different dilation rates to filter out redundant information and refine the feature maps. Dilated convolution enlarges the receptive field without increasing the number of parameters while maintaining the resolution. The size of the convolution kernel determines the size of the receptive field by convolving with the input feature map. The larger the kernel, the larger the receptive field, but it will also bring more parameters, which will increase the computational burden. Dilated convolution can solve this issue well and can be converted from a normal convolution.

$$k' = k + (k - 1) \times (d - 1) \tag{2}$$

where k' and k are the sizes of the dilated convolution and normal convolution, respectively, and d denotes the dilation rate. If $d = 1$, the dilated convolution is equivalent to the normal convolution.

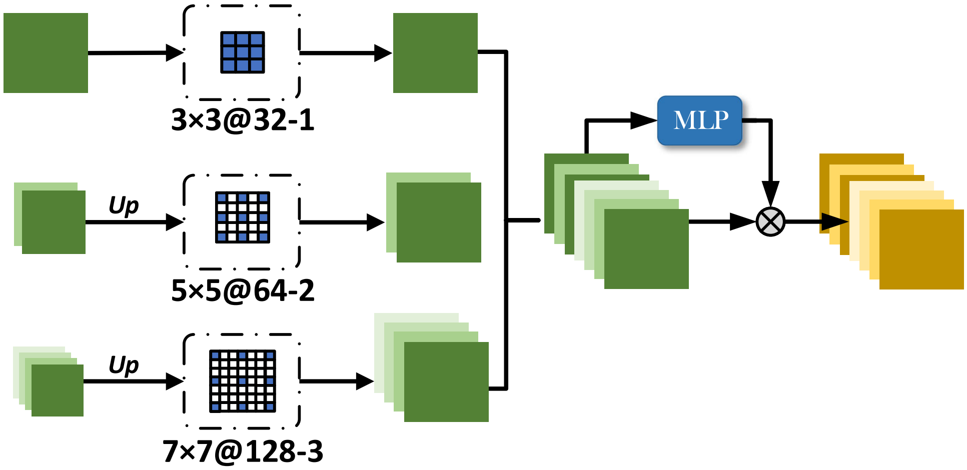


Fig. 6 The details of the Multi-Receptive Field Fusion module. “Up” indicates upsample and the parameters of convolution are represented as (kernel size) × (kernel size) @ (channel number)-(dilation rate)

Moreover, instead of simple concatenation or weighted average fusion, we assign each channel a specific weight based on its importance in all channels and adaptively fuse the filtered multi-scale features by multiplying the weights and the corresponding channels. Here, the weights are learned by a multi-layer perceptron (MLP) with three fully connected layers (FC). The first FC is employed for dimension reduction to reduce the computational burden, and the last FC is employed for dimension increment to adapt to the input channels. Before being input into the MLP, the feature map performs a GAP operation on each channel, and the weights output from the MLP are activated by a Sigmoid function. Then, the feature map is fed into the regression layer to generate the final density map.

3.5 Loss function

Counting methods based on density estimation commonly use Euclidean distance as the loss function to optimize the pixel-wise error:

$$L(\Theta) = \frac{1}{2b}\sum_{i=1}^{b}\left\| F\left(I^i;\Theta\right) - G^i\right\|_2^2 \tag{3}$$

where b is the batch size, I^i denotes the input image, $F\left(I^i;\Theta\right)$ is the density map generated by CP-Net with parameters Θ, and G^i is the ground truth. However, the Euclidean loss ignores the correlation between pixels. To obtain a higher-quality density map, the Structural Similarity in Image (SSIM) [43] is introduced into the loss function to further measure the local consistency of the density map. The final combined loss function is defined as:

$$L = L_E + \lambda\left(1 - \frac{1}{N}\sum_x \mathrm{SSIM}(x)\right) \tag{4}$$

where L_E is the Euclidean distance, N is the total count of pixels in the density map, and the parameter λ is set to 0.001.

4 Experiments

In this section, we first briefly describe the evaluation metrics and the implementation details of our experiment. Then, we summarize the counting datasets used in the experiments and present the comparison results with state-of-the-art methods. Finally, an ablation study is performed to comprehensively analyze the effectiveness of the proposed CP-Net.

4.1 Evaluation Metrics

To evaluate the counting performance of CP-Net, the widely used Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) are adopted as evaluation metrics, which are formulated as follows:

$$Mae = \frac{1}{N}\sum_{i=1}^{N}\left|y_i - \hat{y}_i\right| \tag{5}$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left|y_i - \hat{y}_i\right|^2} \tag{6}$$

where N denotes the count of test images, y_i and $\hat{y}_i$ denote the predicted and ground-truth counts of the i-th test image, respectively.

Moreover, to measure the gap between the predicted and ground-truth density maps, the Structural Similarity in Image (SSIM) and Peak Signal-to-Noise Ratio (PSNR) are adopted in the experiment.

4.2 Implementation details

Network settings The front-end module consists of the first seven convolutional layers of the VGG16 with pretrained parameters. The number of channels in the three branches is set to 32, 64, and 128. The regression layer is composed of two 3 × 3 convolutional layers with 128 and 64 channels, and a 1 × 1 convolutional layer for outputting the final density map.

Data augmentation To augment the training data, we crop fixed-size image patches at different locations in each image. Specifically, the clipping sizes of Shanghai Tech Part B and WorldExpo'10 are 512 × 512 and 512 × 672, respectively, while those of Shanghai Tech Part A, UCF_CC_50 and UCF-QNRF are 256 × 256. Notably, the complete image is still fed into our CP-Net during the testing phase. Moreover, the horizontal flip strategy is also adopted to double the training data. To generate ground-truth density maps, we employ a normalized Gaussian kernel to process each point annotation in the original dataset. The ground-truth count of each image can be obtained by summing the pixel values in the ground-truth density map.

Training details Our framework is deployed using PyTorch [30] and conducted on NVIDIA RTX 3090 GPUs. During the training phase, CP-Net is trained using the Adam optimizer [18]. Except for the batch size of WorldExpo'10, which is set to 8, the batch size of other datasets, such as Shanghai Tech A&B, is set to 4. During the testing phase, the batch size of datasets with different image resolutions is set to 1. To prevent the model from skipping the optimal solution during training, we adopt the learning rate decay to obtain a lower learning rate as the training batch increases. Specifically, we set the initial learning rate and the decay rate to 1e-4 and 0.995, respectively, and reduce the learning rate from epoch 1.

4.3 Comparison with the state-of-the-art methods

We compare our CP-Net with a number of state-of-the-art methods in terms of both counting accuracy and density map quality on the four most commonly used and challenging public crowd counting datasets, including UCF_CC_50 [14], Shanghai Tech A&B [51], UCF-QNRF [15], and WorldExpo'10 [48]. This subsection begins with a brief summary of each counting dataset, followed by the comparison results with the SOTA.

4.3.1 Comparison of counting accuracy

Shanghai Tech A&B includes 1,198 challenging images with 330,165 annotations. The dataset is divided into Part A and Part B, and the image styles of the two parts are very different. The images of Part A are randomly selected from the internet, and their resolutions and crowd densities vary dramatically. Part B comprises street-view images from fixed cameras, where the crowds are relatively sparse compared with Part A. Both Part A and Part B exhibit large variations in crowd scale, making this dataset one of the most widely used in crowd counting.

On this dataset, we compare our CP-Net with the state-of-the-art methods, and Table 1 reports the detailed comparison results. On part A, our CP-Net achieved considerably excellent performance. Specifically, CP-Net obtains the best MAE of 58.5, which is improved by 3.8% compared with the existing best method and achieves the best RMSE of 95.4, which is improved by 1.5%. Although the SOTA methods have achieved few errors on Part B, our CP-Net still improves the counting performance, achieving the optimal values of MAE and RMSE: 6.7 and 10.6, respectively. The crowd density and style of the two subdatasets differ greatly, but CP-Net achieves excellent performance in both, indicating the strong adaptability of our model to different scenes.

UCF_CC_50 includes 50 gray images that cover various crowd densities in various scenarios, such as concerts, marathons, and stadiums. This dataset is appropriate for testing the training performance of the model with a small batch dataset. We train our CP-Net on the UCF_CC_50 dataset and compare it with nine S-O-T-A methods. According to the criteria developed in [14], a 5-fold cross-validation is conducted on UCF_CC_50. The experimental data of all the methods are summarized in Table 2. Our CP-Net achieves the best MAE of 198.2 and the best RMSE of 283.9, reducing the error by 7.2 and 13.4, respectively, compared with the second-best. The excellent results prove the strong ability of our network to deal with extremely dense scenes.

UCF-QNRF includes 1,535 challenging images with approximately 1.25 million annotations. This dataset contains images of crowd scenes under various perspectives and density levels. The average resolution of the images in the

Table 1 Error comparison on the Shanghai Tech A&B dataset. Best performance is bolded

Methods	Part A		Part B	
	MAE	RMSE	MAE	RMSE
MCNN [51]	110.2	173.2	26.4	41.3
CSRNet [20]	68.2	115.0	10.6	16.0
CFF [35]	65.2	109.4	7.2	12.2
TEDnet [16]	64.2	109.1	8.2	12.8
CANet [24]	62.3	100.0	7.8	12.2
DUBNet [29]	64.6	106.8	7.7	12.5
SDANet [28]	63.6	101.8	7.8	10.2
RPNet [46]	61.2	96.9	8.1	11.6
MMNet [3]	60.8	99.0	7.6	11.7
AMRNet [25]	61.6	98.4	7.0	11.0
SARNet [26]	64.4	100.2	8.4	13.4
PDD-CNN [42]	64.7	99.1	8.8	14.3
CP-Net (Ours)	**58.5**	**95.4**	**6.7**	**10.6**

dataset is larger relative to most other datasets, resulting in a wide range in the head size of each person. On the UCF-QNRF dataset, our CP-Net achieves the best MAE of 91.2 and the best RMSE of 156.6, which are improved by 2.8% and 1.5%, respectively, compared with the suboptimal ones (listed in Table 2). The drastic density and scale variations make this dataset extremely challenging, while our CP-Net still achieves satisfactory results, which further proves the strong generalization ability and universality of our method.

WorldExpo’10 consists of 3,980 images with approximately two hundred thousand annotations randomly selected from 1,132 videos taken by 108 cameras during the 2010 Shanghai World Expo. The training set contains 3,380 images taken from 103 cameras in 103 scenes, and the testing set contains 600 images taken from 5 cameras in 5 different scenes. The large number of scenarios included in the dataset makes it a challenging dataset, and it is often used to verify the cross-scenario counting capability of the method. As Table 3 lists, CP-Net outperforms other state-of-the-art methods on the WorldExpo’10 dataset. More specifically, our CP-Net obtains the lowest counting errors on three scenes and the optimal average counting error, demonstrating the powerful cross-scene counting performance of our CP-Net.

4.3.2 Comparison of density map quality

We compare our CP-Net with nine state-of-the-art methods to verify the performance of our method in generating high-quality density maps, and the quantitative comparison results are detailed in Table 4. It can be seen that our CP-Net outperforms the nine methods in both PSNR and SSIM, which shows that our CP-Net is state-of-the-art.

Table 2 Error comparison on the UCF_CC_50 and UCF-QNRF datasets. Best performance is bolded

Methods	UCF_CC_50		**UCF-QNRF**	
MAE	RMSE	MAE	RMSE	
MCNN [51]	377.6	509.1	277	426
CSRNet [20]	266.1	397.5	120.3	208.5
CFF [35]	-	-	93.8	**146.5**
TEDnet [16]	249.4	354.5	113	188
DUBNet [29]	243.8	329.3	105.6	180.5
SDANet [28]	227.6	316.4	-	-
MMNet [3]	209.7	309.7	104	178
AMSNet [12]	208.4	297.3	101.8	163.2
SARNet [26]	242.3	320.4	-	-
PDD-CNN [42]	205.4	311.7	115.3	190.2
CP-Net (Ours)	**198.2**	**283.9**	**91.2**	156.6

Table 3 Mean absolute error comparison on the WorldExpo’10 dataset. Best performance is bolded

Methods	WorldExpo’10					
	S1	S2	S3	S4	S5	Avg
MCNN [51]	3.4	20.6	12.9	13.0	8.1	11.6
CSRNet [20]	2.9	11.5	8.6	16.6	3.4	8.6
SANet [2]	2.6	13.2	9.0	13.3	3.0	8.2
ADMG [40]	4.0	18.1	**7.2**	12.3	5.7	9.5
TEDNet [16]	2.3	10.1	11.3	13.8	**2.6**	8.0
PCC Net [6]	**1.9**	18.3	10.5	13.4	3.4	9.5
RPNet [46]	2.4	10.2	9.7	11.5	3.8	8.2
SARNet [26]	2.5	10.8	8.6	15.2	3.5	7.6
CP-Net (Ours)	2.5	**9.9**	7.9	**11.0**	**2.6**	**6.8**

MCNN [51], CSRNet [20], CFF [35], and DSSINet [23] are four representative methods for crowd counting based on density estimation, which also focus on generating high-quality density maps. Qualitatively, some density maps predicted by the above S-O-T-A methods are visualized in Fig. 7 for a more intuitive comparison. Obviously, compared with other methods, the density maps generated by our CP-Net are clearer and more accurate, which is quite close to the ground truth. The density maps estimated by other methods contain more noise and are fuzzier. This shows that our CP-Net can better resist noise interference and generate high-quality density maps.

Moreover, we also select some representative density maps from other datasets for visualization in Fig. 8. It can be clearly seen in the visualized images that our CP-Net shows excellent density estimation performance in both extremely congested scenes with thousands of people and sparse scenes with only a few people. In summary, CP-Net

Table 4 Comparison of density map quality on Shanghai Tech Part A. Best performance is bolded

Methods	Part A	
	SSIM	PSNR
MCNN [51]	0.52	21.40
CP-CNN [37]	0.72	21.72
Switch-CNN [33]	0.67	21.91
PCC Net [6]	0.74	22.78
CSRNet [20]	0.76	23.79
SANet [2]	0.78	23.36
ANF [47]	0.78	24.10
CFF [35]	0.78	25.40
TEDnet [16]	0.83	25.88
CP-Net (Ours)	**0.87**	**27.90**

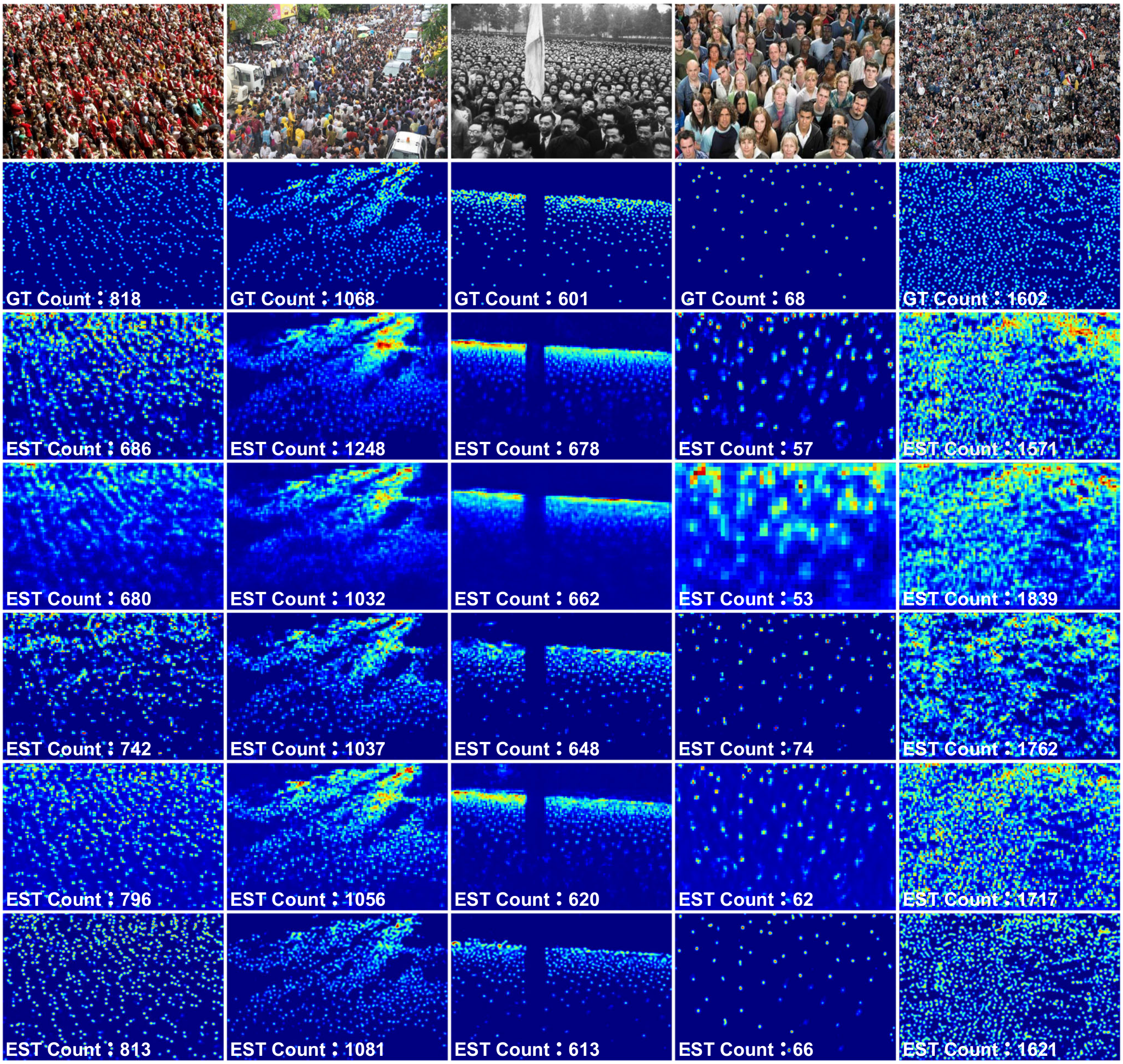


Fig. 7 Comparison of results from the proposed CP-Net along with other state-of-the-art methods. Top Row: Sample images from Shanghai Tech Part A. Second Row: Ground truth. Third Row: The results of MCNN. Fourth Row: The results of CSRNet. Fifth Row: The results of CFF. Sixth Row: The results of DSSINet. Last Row: The results of our CP-Net

can not only accurately count crowds, but also accurately predict the crowd distribution in various scenes.

4.4 Ablation study

In this section, we perform a comprehensive ablation study on the Shanghai Tech Part A dataset to demonstrate the impact of different modules and settings on CP-Net performance. For convenience, we use "CP-Net" and "STA" to represent our complete method and Shanghai Tech Part A, respectively. STA covers various scenes, densities and perspectives, which is quite suitable for ablation studies. The comparison results are detailed in Tables 5 and 6, with the best performance highlighted in bold.

4.4.1 The effect of the cascaded multi-resolution collaborative structure

Since we introduce three cascaded parallel branches into the backbone, the impact of multi-resolution branches on the counting performance is first explored. We build a "Baseline" by removing the three branches generated from

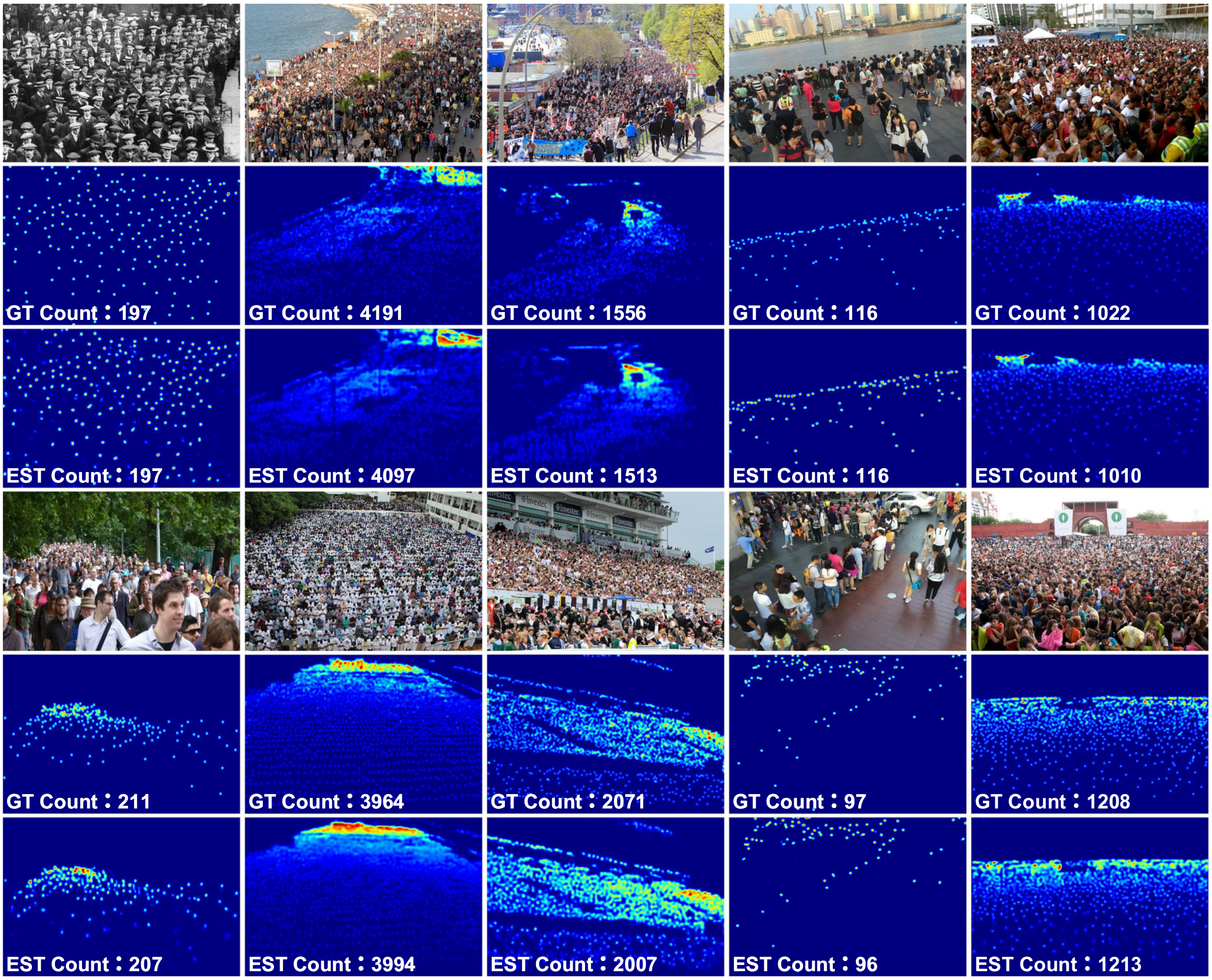


Fig. 8 Visualization results of CP-Net on different datasets. The images cover a variety of perspectives, scenes and densities. “EST Count” and “GT Count” correspond to the estimated count and the ground-truth count, respectively

Stage 1 to Stage 3 and then add branch 1, branch 1, branch 2, and branch 1, branch 2, branch 3 to build “CP-B1”, “CP-B1&2”, and “CP-B1&2&3”, respectively. It is worth noting that to control variables, there is no cross-branch information exchange in the three networks and only one regression layer in Stage 5. Then, we explore the effect of the channel number of branches with different resolutions on network performance. Specifically, we employ the

Table 5 Ablation study of different network structure settings on STA dataset

Methods	STA	
	MAE	RMSE
Baseline	83.0	140.3
CP-B1	70.1	115.4
CP-B1&2	65.2	106.6
CP-B1&2&3	62.9	99.8
CP-c32	60.8	96.6
CP-Net	**58.5**	**95.4**

Table 6 Ablation study of the proposed modules on the STA dataset

Methods	STA	
	MAE	RMSE
CP-rIF	62.9	103.2
CP-rSK	62.1	101.7
CP-rIE	61.6	100.2
CP-rMF1	66.5	108.6
CP-rMF2	59.7	99.0
CP-Net	**58.5**	**95.4**

DWConv with step 2 instead of MBG for downsampling and set the number of channels for each branch to 32 (represented as “CP-c32”).

As listed in Table 5, with the continuous addition of multi-resolution branches, the counting performance of CP-Net continues to improve. Especially after introducing branch 1 into the Baseline, the network makes significant improvements in counting performance. With the gradual addition of branch 2 and branch 3, the counting performance of the network is also continuously enhanced. Each branch maintains a specific-sized resolution, ensuring efficient extraction of multi-scale features, and the complementarity between features is also used to refine each other. Furthermore, CP-c32 is weaker than CP-Net in counting performance, indicating that our channel number setting is effective and can retain richer information.

4.4.2 The effect of the feature extraction unit

To explore the influence of different configurations of feature extraction units on the feature extraction ability of the network, other conditions being consistent, we remove the information filtering module and the skip connection in the FEU. For simplicity, we denote the two networks formed by the unit that removed the filtering module and the unit that removed the skip connection as “CP-rIF” and “CP-rSK”, respectively. As shown in Table 6, the counting performance of the two networks is significantly reduced compared with CP-Net. During the three feature extraction stages, the skip connections in each independent branch allow abundant details to be retained and continuous multi-level fusions to be conducted. Moreover, the introduction of the filtering module can control the information expression of channels containing different objects and reduce the negative impact of noise on network performance, which can be proven by the visualized density maps in Fig. 9. As shown in the visualization results, CP-rIF without IFM mistakenly recognizes head-shaped objects as heads, while CP-Net more accurately distinguishes the foreground and background. It is worth noting that in the ground truth, a point annotation is mistakenly labeled in a place where there is no one (framed in a white box), while our CP-Net avoids this error.

4.4.3 The effect of the information exchange module

To evaluate the effect of the IEM, we remove the IEM from the backend of Stage 3 and Stage 4 (represented as “CP-rIE”). Therefore, the three branches extract features independently without performing information communication across branches. The quantitative results are shown in Table 6. The experimental data of CP-rIE (MAE is 61.6 and RMSE is 100.2) are significantly worse than those of CP-Net (MAE is 58.5 and RMSE is 95.4) on STA, indicating that the correlation between parallel branches established by IEM is conducive to improving the counting performance. Furthermore, we compare the visualized results to visually show the effect of IEM on density map quality. As shown in Fig. 10, after introducing the IEM, the quality of the density map significantly improved, especially in extremely dense areas. The IEM breaks the independence between branches and enables each specific feature map to obtain complementary information from other resolutions.

4.4.4 The effect of the multi-receptive field fusion module

To validate the advantage of MRFF, other conditions being equal, we remove the fusion part of the MRFF and only output the feature map extracted by branch 1 (represented as “CP-rMF1”). Moreover, we also remove the MLP in the MRFF to explore the effectiveness of adaptive fusion (represented as “CP-rMF2”). As shown in Table 6, the estimation errors of CP-rMF1 are significantly higher than those of CP-Net on STA, with the MRFF reducing the error by 8.0. Similarly, the counting performance of CP-Net with MLP is better than that of CP-rMF2 without MLP.

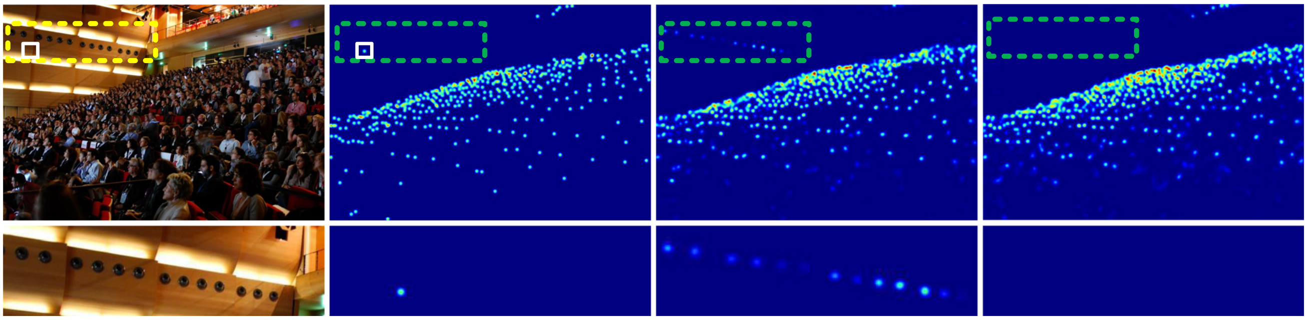


Fig. 9 From left to right, the first and second columns are the original image and the corresponding ground-truth density map, respectively. The third and fourth columns are the density maps estimated by CP-rIF and CP-Net, respectively. The image patch in the second row corresponds to the details highlighted in the first row

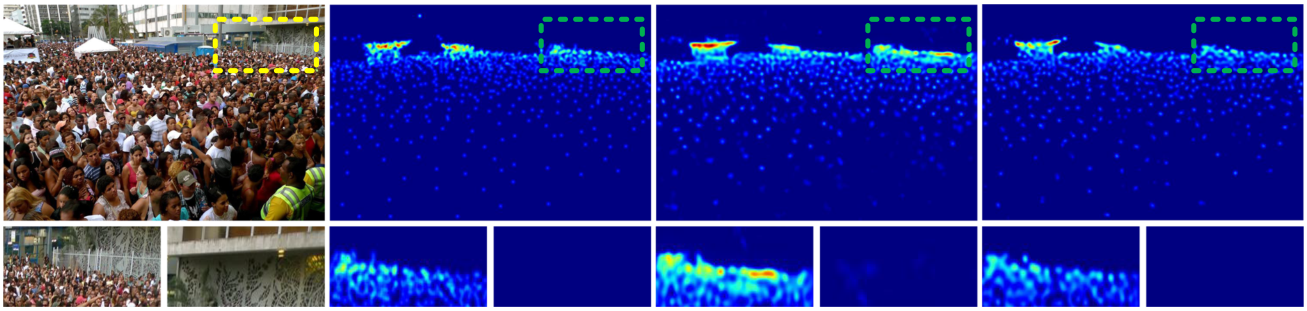


Fig. 10 From left to right, the original image, the corresponding ground-truth density map, and the density maps estimated by CP-rIE and CP-Net are shown successively. The two image patches at the bottom are enlarged versions of the details highlighted in the image

The comparison results indicate that the MRFF can further improve the robustness of the network and greatly improve the counting accuracy by aggregating multi-scale features.

5 Conclusion

In this paper, we propose a novel multi-resolution collaborative architecture called CP-Net for accurate crowd counting and high-quality density map generation. The proposed CP-Net gradually cascades low-resolution branches in parallel to ensure the acquisition of multi-scale features. To refine the resolution-specific features, we introduce an information exchange module for cross-branch information communication. Furthermore, we construct a multi-receptive field fusion module at the back end of the network to fully aggregate multi-scale features and make the network more robust to scale variations. Extensive experiments on four benchmark datasets have shown that CP-Net is state-of-the-art in terms of density map quality and crowd counting accuracy.

In the future, we will focus on a loss function more suitable for extremely congested scenes and further research on crowd localization and crowd semantic segmentation via high-quality density maps. In addition, we will extend CP-Net to other applications, such as cell counting, animal counting, and vehicle counting.

Acknowledgements This work was supported in part by the National Natural Science Foundation of China under Grant 61976127, in part by the Shandong Provincial Natural Science Foundation under Grants ZR2021LZL012, ZR2019MF071, in part by the Jinan Science and technology innovation development Foundation under Grant 202126003.

References

1. Boominathan L, Kruthiventi SS, Babu RV (2016) Crowdnet: a deep convolutional network for dense crowd counting. In: Proceedings of the 24th ACM international conference on Multimedia (ACM MM), pp 640–644
2. Cao X, Wang Z, Zhao Y, Su F (2018) Scale aggregation network for accurate and efficient crowd counting. In: Proceedings of the european conference on computer vision (ECCV), pp 734–750
3. Dong L, Zhang H, Ji Y, Ding Y (2020) Crowd counting by using multi-level density-based spatial information: a multi-scale CNN framework. Inf Sci 528:79–91
4. Fan Z, Zhang H, Zhang Z, Lu G, Zhang Y, Wang Y (2022) A survey of crowd counting and density estimation based on convolutional neural network. Neurocomputing 472:224–251
5. Fan Z, Zhu Y, Song Y, Liu Z (2020) Generating high quality crowd density map based on perceptual loss. Appl Intell 50(4):1073–1085
6. Gao J, Wang Q, Li X (2019) PCC Net: Perspective crowd counting via spatial convolutional network. IEEE Trans Circuits Syst Video Technol 30(10):3486–3498
7. Gao J, Wang Q, Yuan Y (2019) SCAR: spatial-/channel-wise attention regression networks for crowd counting. Neurocomputing 363:1–8
8. Gu L, Pang C, Zheng Y, Lyu C, Lyu L (2021) Context-Aware Pyramid attention network for crowd counting. Appl Intell, 1–17
9. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), pp 770–778
10. Hossain M, Hosseinzadeh M, Chanda O, Wang Y (2019) Crowd counting using Scale-Aware attention networks. In: Proceedings of the IEEE Winter conference on applications of computer vision (WACV), pp 1280–1288
11. Hu J, Shen L, Sun G (2018) Squeeze-and-Excitation Networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), pp 7132–7141
12. Hu Y, Jiang X, Liu X, Zhang B, Han J, Cao X, Doermann D (2020) NAS-Count: Counting-by-Density with Neural Architecture Search. In: Proceedings of the European conference on computer vision (ECCV), pp 747–766
13. Huang Z, Wang X, Huang L, Huang C, Wei Y, Liu W (2019) CCNEt: Criss-Cross Attention for Semantic Segmentation. In: Proceedings of the IEEE international conference on computer vision (ICCV), pp 603–612
14. Idrees H, Saleemi I, Seibert C, Shah M (2013) Multi-source Multi-scale Counting in Extremely Dense Crowd Images. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), pp 2547–2554
15. Idrees H, Tayyab M, Athrey K, Zhang D, Al-Maadeed S, Rajpoot N, Shah M (2018) Composition loss for counting, density map estimation and localization in dense crowds. In: Proceedings of the European conference on computer vision (ECCV), pp 532–546
16. Jiang X, Xiao Z, Zhang B, Zhen X, Cao X, Doermann D, Shao L (2019) Crowd counting and density estimation by trellis Encoder-Decoder networks. In: Proceedings of the IEEE conference on

computer vision and pattern recognition (CVPR), pp 6133–6142

17. Jiang X, Zhang L, Xu M, Zhang T, Lv P, Zhou B, Yang X, Pang Y (2020) Attention scaling for crowd counting. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), pp 4706–4715
18. Kingma DP, Ba J (2015) Adam: a method for stochastic optimization. In: 3Rd international conference on learning representations (ICLR), pp 273–297
19. Li B, Huang H, Zhang A, Liu P, Liu C (2021) Approaches on crowd counting and density estimation: a review. Pattern Anal Applic 24(3):853–874
20. Li Y, Zhang X, Chen D (2018) CSRNEt: dilated convolutional neural networks for understanding the highly congested scenes. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), pp 1091–1100
21. Liu H, Xu B, Lu D, Zhang G (2018) A path planning approach for crowd evacuation in buildings based on improved artificial bee colony algorithm. Appl Soft Comput 68:360–376
22. Liu L, Chen J, Wu H, Li G, Li C, Lin L (2021) Cross-Modal Collaborative Representation Learning and a Large-Scale RGBt Benchmark for Crowd Counting. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), pp 4823–4833
23. Liu L, Qiu Z, Li G, Liu S, Ouyang W, Lin L (2019) Crowd counting with deep structured scale integration network. In: Proceedings of the IEEE international conference on computer vision (ICCV), pp 1774–1783
24. Liu W, Salzmann M, Fua P (2019) Context-Aware Crowd counting. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), pp 5099–5108
25. Liu X, Yang J, Ding W, Wang T, Wang Z, Xiong J (2020) Adaptive mixture regression network with local counting map for crowd counting. In: Proceedings of the European conference on computer vision (ECCV), pp 241–257
26. Liu YB, Jia RS, Liu QM, Zhang XL, Sun HM (2021) Crowd counting method based on the Self-Attention residual network. Appl Intell 51(1):427–440
27. Liu Z, Qi X, Fu CW (2021) One thing one click: a self-training approach for weakly supervised 3d semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), pp 1726–1736
28. Miao Y, Lin Z, Ding G, Han J (2020) Shallow feature based dense attention network for crowd counting. In: Proceedings of the AAAI conference on artificial intelligence (AAAI), pp 11765–11772
29. Oh MH, Olsen P, Ramamurthy KN (2020) Crowd counting with decomposed uncertainty. In: Proceedings of the AAAI conference on artificial intelligence (AAAI), pp 11799–11806
30. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, Killeen T, Lin Z, Gimelshein N, Antiga L et al (2019) Pytorch: an imperative style, high-performance deep learning library. Advances in Neural Information Processing Systems 32:8026–8037
31. Qin X, Zhang Z, Huang C, Gao C, Dehghan M, Jagersand M (2019) BASNEt: boundary-aware salient object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), pp 7479–7489
32. Rong L, Li C (2021) Coarse- and fine-grained attention network with background-aware loss for crowd density map estimation. In: Proceedings of the IEEE Winter conference on applications of computer vision (WACV), pp 3675–3684
33. Sam DB, Surya S, Babu RV (2017) Switching convolutional neural network for crowd counting. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), pp 4031–4039
34. Shi M, Yang Z, Xu C, Chen Q (2019) Revisiting perspective information for efficient crowd counting. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), pp 7279–7288
35. Shi Z, Mettes P, Snoek CG (2019) Counting with focus for free. In: Proceedings of the IEEe international conference on computer vision (ICCV), pp 4200–4209
36. Simonyan K, Zisserman A (2015) Very deep convolutional networks for Large-Scale image recognition. In: 3Rd international conference on learning representations (ICLR), pp 1–14
37. Sindagi VA, Patel VM (2017) Generating high-quality crowd density maps using contextual pyramid CNNs. In: Proceedings of the IEEE international conference on computer vision (ICCV), pp 1861–1870
38. Song Q, Wang C, Wang Y, Tai Y, Wang C, Li J, Wu J, Ma J (2021) To choose or to fuse? scale selection for crowd counting. In: Proceedings of the AAAI conference on artificial intelligence (AAAI), pp 2576–2583
39. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A (2015) Going deeper with convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), pp 1–9
40. Wan J, Chan A (2019) Adaptive density map generation for crowd counting. In: Proceedings of the IEEE international conference on computer vision (ICCV), pp 1130–1139
41. Wan J, Luo W, Wu B, Chan AB, Liu W (2019) Residual regression with semantic prior for crowd counting. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), pp 4036–4045
42. Wang W, Liu Q, Wang W (2021) Pyramid-Dilated Deep convolutional neural network for crowd counting. Appl Intell, 1–13
43. Wang Z, Bovik AC, Sheikh HR, Simoncelli EP (2004) Image quality assessment: From error visibility to structural similarity. IEEE Transactions on Image Processing 13(4):600–612
44. Xu C, Qiu K, Fu J, Bai S, Xu Y, Bai X (2019) Learn to scale: Generating multipolar normalized density maps for crowd counting. In: Proceedings of the IEEE international conference on computer vision (ICCV), pp 8382–8390
45. Yan Z, Yuan Y, Zuo W, Tan X, Wang Y, Wen S, Ding E (2019) Perspective-Guided Convolution networks for crowd counting. In: Proceedings of the IEEE international conference on computer vision (ICCV), pp 952–961
46. Yang Y, Li G, Wu Z, Su L, Huang Q, Sebe N (2020) Reverse perspective network for Perspective-Aware object counting. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), pp 4374–4383
47. Zhang A, Yue L, Shen J, Zhu F, Zhen X, Cao X, Shao L (2019) Attentional neural fields for crowd counting. In: Proceedings of the IEEE international conference on computer vision (ICCV), pp 5714–5723
48. Zhang C, Kang K, Li H, Wang X, Xie R, Yang X (2016) Data-Driven Crowd understanding: a baseline for a Large-Scale crowd dataset. IEEE Transactions on Multimedia 18(6):1048–1061
49. Zhang L, Shi M, Chen Q (2018) Crowd counting via Scale-Adaptive convolutional neural network. In: Proceedings of the IEEE Winter conference on applications of computer vision (WACV), pp 1113–1121

50. Zhang Q, Cong R, Li C, Cheng MM, Fang Y, Cao X, Zhao Y, Kwong S (2020) Dense attention fluid network for salient object detection in optical remote sensing images. IEEE Trans Image Process 30:1305–1317
51. Zhang Y, Zhou D, Chen S, Gao S, Ma Y (2016) Single-Image Crowd counting via Multi-Column convolutional neural network. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), pp 589–597
52. Zhu L, Li C, Yang Z, Yuan K, Wang S (2020) Crowd density estimation based on classification activation map and patch density level. Neural Comput & Applic 32(9):5105–5116

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Lei Lyu is an associate professor of School of Information Science and Engineering, Shandong Normal University, Jinan, China. He received a Ph.D. degree in computer application technology from University of Chinese Academy of Sciences in 2013. His current research interests include artificial intelligence and virtual reality.

Run Han was born in Shandong, China, in 1997. He received the B. S. degree from Shandong Normal University, China, in 2020. He is currently pursuing a master's degree in School of Information Science and Engineering, Shandong Normal University. His research interest includes computer vision and deep learning.

Ziming Chen was born in Shandong, China, in 1995. He received a B.S. degree and the M.S. degree from Shandong Normal University in 2018 and 2021, respectively. His research interest includes computer vision and deep learning.