International Journal of Innovative Computing, Information and Control
Volume **18**, Number **4**, August **2022**

ICIC International ©2022 ISSN 1349-4198
pp. **1147–1162**

LEVERAGE MULTI-SCALE DILATED CONVOLUTIONAL NEURAL NETWORK WITH GLOBAL ATTENTION FEATURE FUSION FOR CROWD COUNTING

Meilei Lv[1], Kuncai Zhang[2], Xiaoyun Zheng[3], Wei Yang[4] and Zhe-Ming Lu[2,*]

[1]College of Electrical and Information Engineering
Quzhou University
No. 78, North Jiuhua Road, Quzhou 324000, P. R. China
meilei_lv@163.com

[2]School of Aeronautics and Astronautics
Zhejiang University
No. 38, Zheda Road, Hangzhou 310027, P. R. China
nampl@zju.edu.cn; *Corresponding author: zheminglu@zju.edu.cn

[3]State Grid Wenzhou Power Supply Co., Ltd.
No. 1314, Jinxiu Road, Wenzhou 325000, P. R. China
ZXY-1024@163.com

[4]State Grid Quzhou Power Supply Co., Ltd.
No. 6, Xinhe Road, Quzhou 324000, P. R. China
easteryang@163.com

Received February 2022; revised May 2022

Abstract. *Crowd counting in various complex scenes is a challenging problem which has attracted much attention in both academic circles and industries due to its applications in public safety. Recently, state-of-the-art methods for counting people in crowded scenes rely on deep Convolutional Neural Networks (CNNs) to estimate crowd density. In this paper, we propose a novel network for accurate and efficient crowd counting called Multi-Scale Dilated Convolutional Neural Network (MSDNet) to provide an effective deep learning method that can perform accurate estimation of crowd density. The proposed MSDNet uses multi-scale dilated kernels to aggregate features across different scales. In addition, a Global Attention Fusion Module (GAFM) is designed to merge features from different levels in a global attention mechanism. Extensive experiments on four benchmark crowd counting datasets (ShanghaiTech, UCF_CC_50, WorldExpo'10, and UCSD) demonstrate the superior performance of the proposed method over other competitive approaches.*

Keywords: Crowd counting, Deep learning, Global attention feature fusion, Multi-scale dilated network

1. **Introduction.** Crowd counting is a visual cognitive task that aims to accurately estimate the number of people in various congested scenes. With the rapid development of deep learning, current state-of-the-art approaches use deep learning methods to regress the density map and integrate over the density map to get the crowd count.

While many works have been done, crowd counting is still a challenge in both academic circles and industries on account of following difficulties. First, congestion and occlusion are common challenges for crowd counting in different crowd scenes. In addition, the diversity of crowd distribution and the crowd density across different scenes

DOI: 10.24507/ijicic.18.04.1147

FIGURE 1. The example of the variation of crowd density

makes it a troublesome problem to estimate the crowd density accurately. According to the benchmark datasets, the crowd density of the image samples varies remarkably from one to another. As shown in Figure 1, different densities often indicate different sizes of the crowd. The samples are selected from ShanghaiTech dataset. The top row lists the source crowd frames and the bottom row lists their density distribution. The crowd density varies remarkably from one to another scenes. Thus, this motivates us to design a proper framework that can capture multi-scale information to better deal with the variance of crowd density in different crowd scenes.

Our motivation comes from dilated convolutional kernel [1], which has been confirmed to be capable of delivering a larger reception field. Dilated convolution is a beneficial alternative of pooling layer that uses a filter with "holes" to alternate the pooling and convolutional layer. While dilated convolution with fixed dilation rate has been verified effective for crowd counting in the CSRNet [2], its capability of capturing multi-scale information is limited. In this paper, we designed a multi-scale dilated convolutional block by assembling dilated convolutions with different dilation rates. By the design of multi-scale dilated convolutional block, more multi-scale information can be captured for accurate crowd counting.

Feature fusion methods like element-wise addition, concatenation, and element-wise maximization, are common operations in various deep neural networks [3-5]. In order to get a more accurate estimation of the crowd count, we also adopt a feature fusion module to merge the features from different levels. Different from the common used feature fusion methods, we designed a Global Attention Fusion Module (GAFM) and experimentally verified the superior performance of the GAFM over other common used feature fusion methods. We evaluate the proposed method on four famous benchmark datasets including ShanghaiTech [6], UCF_CC_50 [7], WorldExpo'10 [8] and UCSD [9]. Extensive experiments on all of these datasets demonstrate the superiority of the proposed method over some state-of-the-art approaches. In summary, the main contributions of this paper are as follows.

i) We proposed a novel and efficient network, dubbed as Multi-Scale Dilated Network (MSDNet) to capture multi-scale information, which improves the accuracy of crowd counting in complex crowd scene.

ii) We designed a novel feature fusion module, dubbed as Global Attention Fusion Module (GAFM) to improve the representational power of the model, which further improves the performance of crowd counting.

The remainder of this paper is organized as follows. Section 2 introduces the related work on crowd counting. Section 3 presents our network architecture. Section 4 gives the implementation details of our scheme. Section 5 shows the experimental results and compares our scheme with existing methods. Section 6 concludes the whole paper.

2. **Related Work.** Crowd counting in varying density scenes has attracted much attention in both academic circles and industries due to its applications in public safety. Due to the congestion and occlusion among the pedestrains, detection-based counting methods cannot perform well in the complex crowd scenes. Therefore, current state-of-the-art methods combine the deep learning technology and density estimation to achieve an accurate count of the crowd. The model architecture of CNN-based density estimation methods can be classified into two categories: single-column CNNs and multi-column CNNs.

Single-column CNNs adopt a front-end of convolutional layers to extract features, followed by several regression layers to predict the count of crowd. Sam et al. in [10] achieved a considerable counting result by adopting a router layer to rout the feature map to different regression CNNs and using these regression CNNs to regress the density map. Due to the computational consumption of dividing the image patches and the limited regressing ability of the back-end CNN, this method can be further improved. Li et al. in [2] designed a single-column CNN dubbed as CSRNet using convolution kernels with fixed dilation rate to process the low-level features and regress the density distribution. The CSRNet [2] can be improved further for more accurate crowd scene representation because its capability of capturing multi-scale information is limited. The proposed MSDNet is a further exploration on the basis of the CSRNet.

Multi-column CNNs use several parallel CNN blocks with different kernel sizes to constitute a multi-column network architecture and concatenate the parallel blocks to merge the feature maps. In [6], Zhang et al. designed a multi-column CNN consisting of three columns of different convolutional kernels to extract the feature of the image and merge the feature maps to get the final density map. Sindagi and Patel [11] proposed a CP-CNN architecture consisting of three branches named global context estimator, density map estimator, and local context estimator. The three-branch CNNs were reported to estimate context at various levels for achieving lower count error and better-quality density maps. Shen et al. in [12] proposed a three-column Adversarial Cross-Scale Consistency Pursuit (ACSCP) network to deal with the problem of inconsistent estimation of across different scaled input patches. In [13], Deb and Ventura proposed a multi-column CNN for perspective-free counting. It used three columns of convolutional neural networks to extract the features and aggregated the features using a series of dilated CNNs. Because of the similarity of the convolution kernels among different branches, the features extracted by different branches may have overlapped parts which causes excessive computation and not all of them are beneficial to the final density estimation. IA-DCCN [28] infuses segmentation information through an inverse attention mechanism into the counting network to capture important regions in the feature maps during learning. MLCNN [30] first adaptively learns multi-level density maps and then fuses them to predict the final output.

The above methods extract multi-scale image features by combining convolution kernels of different sizes into multiple branches, trying to solve the problem of changing the size of the target scale. However, due to the multi-branch architecture, the model parameters increase exponentially, which increases the risk of model overfitting and also increases the hardware and software requirements for the machine. In this paper, we also designed a multi-column convolutional neural network with multi-scale dilated block and attention

feature fusion module to achieve an accurate estimation of the crowd count. Compared to other multi-stream CNNs, our multi-scale dilated block can capture multi-scale information with much less parameters due to the leverage of dilated convolution with different dilation rates.

3. **Network Architecture.** Following [11], we deploy a fine-tuned VGG-16 [14] as the backbone. The front ten convolutional layers of VGG-16 are utilized as the feature extractor of the proposed MSDNet. Then we assign two Multi-scale Dilated Blocks (MDB) following the fine-tuned VGG-16. In addition, we further designed a GAFM to promote the feature fusion in an attention mechanism. Finally, a 1×1 sized convolution layer is used to reshape the channels of the feature maps to 1 to produce the output density map. An overview of the proposed MSDNet is shown in Figure 2. A convolutional layer is denoted as "Conv(kernel size)_(dilation rates)". The output feature maps from the fine-tuned VGG-16 are fed into the multi-scale dilated CNN blocks. The GAFM is used to merge the feature maps from different levels. The 1×1 sized convolutional layer is used to match the channels. In the extraction of the underlying features, the VGG-16 backbone network commonly used in computer vision tasks is used, and the first 10 layers of the pre-trained model are used to extract the underlying features of the image. On this basis, two multi-scale dilated blocks in series are connected to capture multi-scale information, and a global attention fusion module is introduced to further improve the representation ability of the model. Finally, the channel information fusion and channel transformation are realized through the convolution kernel of 1×1 size, the channel number of the feature map is reduced to 1, and the feature map with the channel number of 1 is output as the predicted target density distribution map. The total number of objects in the scene image can be obtained by integrating and summing the pixel values of each position on the density distribution map.

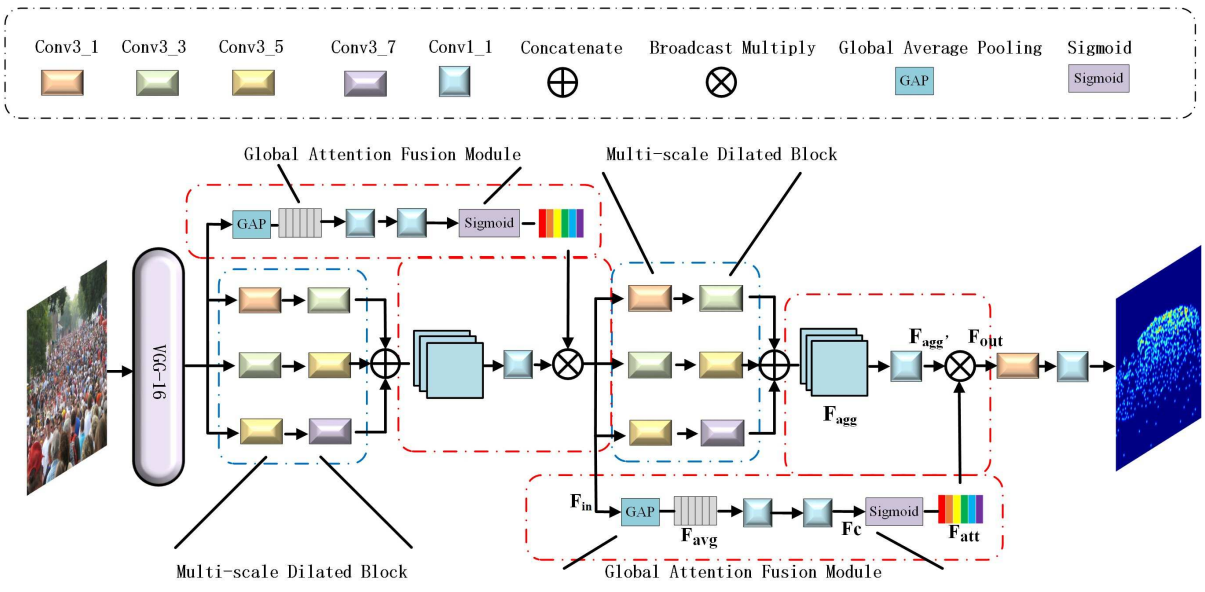

FIGURE 2. The architecture of the proposed MSDNet

3.1. **Multi-scale dilated block.** Dilated convolution, as discussed in [15], is effective for increasing of the receptive field with a linear increase in the number of parameters with respect to each hidden layer. A typical 2-D dilated convolution operation can be defined as follows:

$$f(s,t) = \sum_{m=1}^{M} \sum_{n=1}^{N} x(s + m \cdot r, t + n \cdot r) \cdot \omega(m,n) \tag{1}$$

where (s,t) and (m,n) are the coordinates of feature map and convolution kernel. $f(s,t)$ is the output feature map of the dilated convolution operation and $x(s,t)$ is the input feature map. The input $x(s,t)$ is weighed by $\omega(m,n)$ with the length M and the width N, respectively. The parameter r is the dilation rate which controls the area of receptive field. If $r = 1$, a dilated convolution turns into a normal convolution.

Dilated convolution has been verified effective in the field of semantic segmentation [1, 15] and crowd counting [2, 13]. While dilated convolution with fixed dilation rate can deliver a larger reception field, its capability of capturing multi-scale information is limited. To capture multi-scale information, we designed a multi-scale dilated convolutional block by assembling dilated convolutions with different dilation rates. However, it is verified in [16] that dilated convolution will bring "grid effect" which leads to the discontinuity of the receptive field. In order to avoid the "grid effect", it designs a Hybrid Dilated Convolution (HDC), which use different dilation rates whose common factor cannot be greater than 1. Therefore, we also follow the HDC standard to design our MDB. As shown in Figure 3, the multi-scale dilated kernels can incorporate responses of multiple local areas into the output feature map and provide a wider range of scale invariance for CNNs. The multi-scale dilated block includes L branches of dilated CNNs and the dilation rates of the two dilated convolution in the lth branch CNN are as follows:

$$r_{l1} = 2l - 1,\ r_{l2} = 2l + 1 \quad (l = 1, 2, \ldots, L) \tag{2}$$

In our architecture, we choose $L = 3$. In the following experiments, we will conduct an ablation study on the selection of L.

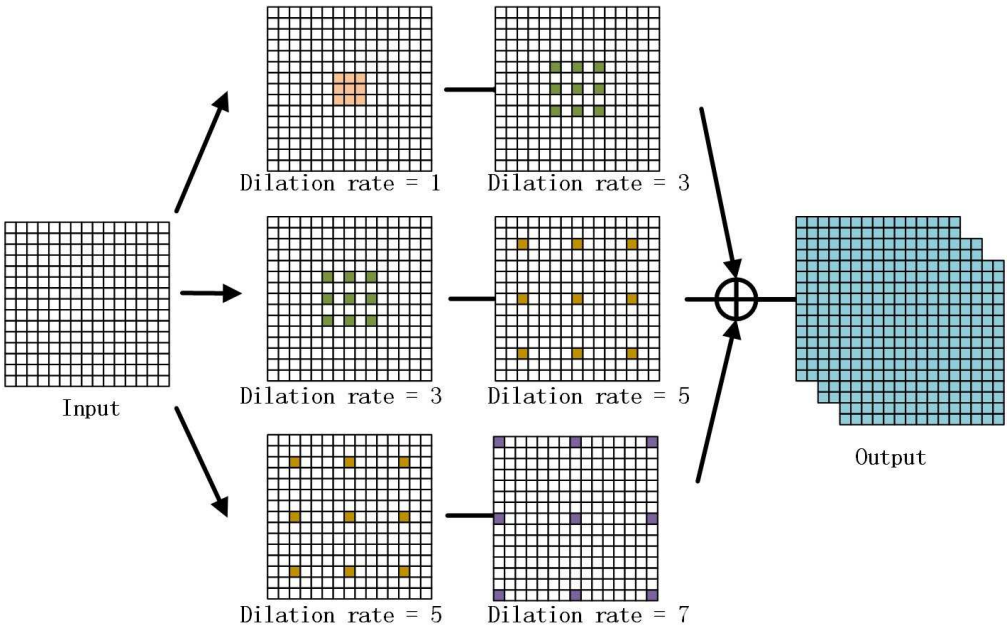

FIGURE 3. Multi-scale dilated block. Employing different values of dilation rate enlarges the receptive field of the model, enabling object encoding at multiple scales.

3.2. **Global attention fusion module.** Most commonly used methods for feature fusion are element-wise addition [3,17,18], concatenation [4,19,20], and element-wise maximization [5,21]. Element-wise addition is an operation to sum up the different feature maps element by element. Concatenation is an operation to concatenate the feature maps along the dimension of channel. Element-wise maximization is an operation to select the max value of the input feature maps element by element to form the output feature map. Element-wise addition and element-wise maximization require equal number of channels and equal size of corresponding input feature maps. Concatenation requires only equal size of corresponding input feature maps.

Given two feature maps $X = [x_1, x_2, \ldots, x_p]$ and $Y = [y_1, y_2, \ldots, y_p]$, p is the channel number of the two input feature maps. Denote the convolution parameters of convolution kernel as w_i, $i = 1, 2, \ldots, v$ ($v = p$ for element-wise addition and element-wise maximization, $v = 2p$ for channel-wise concatenetion). Considering one output channel, the output convolution feature of the element-wise addition fusion is

$$Z_{add} = \sum_{i=1}^{p} (x_i + y_i) * w_i = \sum_{i=1}^{p} x_i * w_i + \sum_{i=1}^{p} y_i * w_i \tag{3}$$

The output convolution feature of the channel-wise concatenation fusion is

$$Z_{\text{concat}} = \sum_{i=1}^{p} x_i * w_i + \sum_{i=1}^{p} y_i * w_{i+p} \tag{4}$$

The output convolution feature of the element-wise maximization fusion is

$$Z_{\max} = \sum_{i=1}^{p} [\max(x_i, y_i)] * w_i \tag{5}$$

In fact, element-wise addition is a special case of concatenation when the convolution parameters of concatenated features satisfy the condition $w^q = w^{p+q}$ ($q = 1, 2, \ldots, p$).

Different from the common used feature fusion methods, we designed a GAFM to selectively emphasize informative channels of features and suppress less useful ones. As shown in Figure 2, the feature maps of different dilated convolution branches have encoded information of different scales. However, different branches are independent, which will limit the representational power of the model. In addition, the feature map before encoded by the MDB retains the global information, which can be used to guide the attention of different scales. Therefore, like the related work [23], we design a global attention fusion module to give different importance to different channels, which greatly improves the representational power of the model.

Taking one of the GAFMs shown in Figure 2 as an example, we denote the input feature map as $F_{in} \in \mathbb{R}^{b \times c_1 \times h \times w}$ and the aggregated feature map as $F_{agg} \in \mathbb{R}^{b \times c_2 \times h \times w}$. c_1 and c_2 are the channel numbers of the corresponding feature maps. w and h are the width and the height of the feature maps, respectively. b is the batch size.

First, the input feature map F_{in} is pooled by a global average pooling layer [22] to get $F_{avg} \in \mathbb{R}^{b \times c_1 \times 1 \times 1}$, which encodes the global information of the different channels. In order to decrease the parameters, we use a 1×1 sized convolution on the F_{avg} to decrease the channel number from c_1 to $\frac{c_1}{\alpha}$. α is the factor for decreasing the channel number, which is set to 16. Then, another 1×1 sized convolution is used to get the attention information $F_c \in \mathbb{R}^{b \times c_1 \times 1 \times 1}$. Next, we get the attention feature map F_{att} by conducting a Sigmoid function on F_c along the channel dimension:

$$F_{att} = \text{sigmoid}\,(F_c) \tag{6}$$

In order to keep the same channel number, we use a 1×1 sized convolution on the aggregated feature map F_{agg} to transoform its channel number to c_1, and the output is denoted as $F_{agg'}$. Finally, the output feature map is got by conducting a broadcast multiply on F_{att} and $F_{agg'}$:

$$F_{out} = F_{att} \otimes F_{agg'} \tag{7}$$

The operations introduced above build an attention mechanism between the input feature and the aggragated feature after the MDB. So we name it Global Attention Fusion Module (GAFM), and we will verify its performance in the ablation study.

4. Implementation.

The fundamental idea of the proposed crowd counting scheme is to deploy a deep neural network to estimate the crowd density map and integrate over the estimated density map to get the crowd count. In this section, we will introduce the specific implementation details including generation of ground-truth density map, training and testing details.

4.1. Generation of ground-truth density map.

The ground-truth density map is generated for training the proposed MSDNet. Referring to the method of generating density maps in [6], we adopt the geometryadaptive kernels to deal with the highly congested crowd scene. We generate the ground-truth density map by blurring each dot of the head annotation using a Gaussian kernel which is normalized to 1. Different from most methods of generating ground-truth maps [2, 6], we make a minor change to adapt to the variance of crowd density by using a density-aware variance parameter σ. The adaptive 2-D Gaussian kernel is given by

$$G(x, y, g(\cdot)) = \sum_{i=1}^{N} \delta(x - x_i)\delta(y - y_i) \cdot \exp\left(-\frac{(x - x_i)^2 + (y - y_i)^2}{2g(\sigma_i)^2}\right) \tag{8}$$

where (x, y) is the pixel coordinate in the generated density map, (x_i, y_i) is the coordinate annotation of the head given by the dataset, and N is the total crowd count of the given image.

The density-aware variance parameter $g(\sigma_i)$ is defined as

$$g(\sigma_i) = \sum_{j=1}^{k} \gamma_j d_{ij} \tag{9}$$

We use the k nearest neighbors' distance d_{ij} weighted by a coefficient γ_j to determine $g(\sigma_i)$. the variance of Gaussian kernel. Different values of γ_j and k are set for different datasets according to the average crowd density.

4.2. Training details.

We train the proposed model on a single machine with a GTX-1080Ti GPU using Pytorch for implementation. We adopt a sample-based strategy for data augmentation, which is to randomly crop the subimage from the original image and resize it by different ratios ranging from 0.9 to 1.1. Here, our intention is to increase the number of images in the dataset and perform simple cropping and scaling operations on the images in the original image set. The ratio of the scaling operation should not deviate too much from the cropped image size, so the range of 0.9-1.1 is used. Similar operations are conducted on the ground-truth density maps. This strategy for data augmentation is effective for training a better model of the proposed MSDNet since it provides different scale information of the training images, which can further reinforce the model's adaptation to different scales.

Similar to other works [10,27], we choose the Euclidean distance between the ground-truth and the estimated density map regressed by our MSDNet to define the loss function. Specifically, the loss function is defined as

$$L(\Theta) = \frac{1}{2B}\sum_{b=1}^{B}\left\|Y\left(X_b;\Theta\right) - Y_b^{Groundtruth}\right\|_2^2 \tag{10}$$

In the above loss function, B is the batch size of training and X_b is the input image. $Y\left(X_b;\Theta\right)$ is the output density map generated by the proposed MSDNet with parameters represented as Θ while $Y_b^{Groundtruth}$ is the ground-truth density map of the input image X_b.

4.3. **Testing and evaluation.** For testing the well-trained model, raw images can be fed into the network. The final output feature map of the network is the estimated density map. Some visual results of the output density maps and the counting results are shown in Figure 4. From top to bottom, the scale of pedestrians is decreasing and the crowd density is increasing. It shows the superior adaptation ability to different scales and densities of crowded scenes. The estimated crowd count C can be calculated by integrating over the

FIGURE 4. Some examples of the estimated crowd counts with different crowd densities and scales using the proposed method

estimated density map:

$$C = \sum_{w=1}^{W} \sum_{h}^{H} P_{w,h} \tag{11}$$

where W and H represent the width and height of the density map, respectively. $p_{w,h}$ is the pixel value at (w, h) of the generated density map.

For evaluation and comparison of the proposed approach and other methods, the widely used Mean Absolute Error (MAE) and the Root Mean Squared Error (RMSE) are adopted. Lower MAE and RMSE indicate a better model. The MAE and RMSE are defined as follows:

$$\text{MAE} = \frac{1}{M} \sum_{i=1}^{M} \left| C_i - C_i^{GT} \right| \tag{12}$$

$$\text{RMSE} = \sqrt{\frac{1}{M} \sum_{i=1}^{M} \left| C_i - C_i^{GT} \right|^2} \tag{13}$$

Here, M is the total number of images in the test datasets. i is the image indicator of the test dataset. C_i^{GT} represents the ground-truth crowd count and C_i represents the estimated crowd count of the ith image.

5. Experiments.

In this section, we will make a general introduction to the four different public datasets and make comparisons on the performance between the proposed approach and the previous state-of-the-art methods on all of the four benchmark datasets. To further verify the effectiveness of different components of the proposed crowd counting approach, we make an extensive ablation study on different variants of the proposed multi-scale dilated convolutional neural network.

5.1. Experimental results on benchmark datasets.

We evaluate our approach on four different public datasets: ShanghaiTech [6], UCF_CC_50 [7], WorldExpo'10 [8], and UCSD [9]. ShanghaiTech dataset [6] is a large-scale dataset for crowd counting, which includes 1,198 annotated images with a total amount of 330,165 pedestrians. The proposed approach is evaluated and compared to other twelve recent works including one traditional method LBP + RR (using manual feature Local Binary Pattern) and eleven CNN-based deep learning methods [2, 6, 8, 10-12, 24, 27-30]. The results are presented in Table 1. As shown in Table 1, all the deep learning methods outperform handcrafted features and our approach achieves the lowest MAE in both of Part A and Part B subsets. Compared to CSRNet which uses a single-column dilated architecture with fixed dilation rate, our multi-scale dilated block architecture shows a superior performance on crowd counting. The UCF_CC_50 dataset [13] is collected from the Internet and contains 50 images with different resolutions and densities. We evaluate and compare our approach with ten recent works and the result is shown in Table 2. As illustrated in the table, our approach achieves the lowest RMSE and competitive MAE among the recent works. Compared to the single-column dilated CSRNet with fixed dilation rate, the proposed multi-scale dilated block architecture achieves an obviously higher accuracy by 36.4 in terms of the MAE metric and 86.2 in terms of the RMSE metric.

The WorldExpo'10 dataset [8] is a highly congested crowd counting dataset including 3,980 annotated frames with a total of 199,923 labeled pedestrians from 1,132 video sequences captured by 108 different surveillance cameras. Experimental results are shown in Table 3. It illustrates the proposed MSDNet achieves the highest accuracy on average. The average MAE decreased 1.2 compared to CSRNet [2], which uses fixed dilation rates.

TABLE 1. Results on ShanghaiTech dataset

Methods	Part A		Part B	
	MAE	RMSE	MAE	RMSE
LBP+RR	303.2	371	59.1	81.7
Zhang et al. [8]	181.8	277.7	32.0	49.8
MCNN [6]	110.2	173.2	26.4	41.3
NetVLAD [27]	107.6	169.3	21.4	33.9
StackPooling [24]	93.98	150.59	18.02	35.64
Switch-CNN [10]	90.4	135.0	21.6	33.4
CP-CNN [11]	73.6	106.4	20.1	30.1
ACSCP [12]	75.7	102.7	17.2	27.4
CSRNet [2]	68.2	115.0	10.6	16.0
MC-CNN [29]	96.1	150.5	17.8	27.7
IA-DCCN [28]	66.9	108.4	10.2	16.0
MLCNN [30]	71.2	112.5	12.1	19.3
MSDNet (ours)	**62.9**	**96.1**	**7.5**	**12.1**

TABLE 2. Results on UCF_CC_50 dataset

Methods	MAE	RMSE
Zhang et al. [8]	467.0	498.5
MCNN [6]	377.6	509.1
Swith-CNN [10]	318.1	439.2
NetVLAD [27]	311.3	401.8
CP-CNN [11]	295.8	320.9
AMDCN [13]	290.8	/
ACSCP [12]	291.0	404.6
CSRNet [2]	266.1	397.5
MC-CNN [29]	295.0	443.7
IA-DCCN [28]	264.2	394.4
MLCNN [30]	242.4	317.8
MSDNet (ours)	**229.7**	**311.3**

The UCSD dataset [9] includes 2,000 frames with a relatively sparse density of crowd scenes captured by a surveillance camera on the UCSD campus. The annotation in each frame ranges from 11 to 46, which is obviously sparser than the aforementioned three datasets. The experimental results are listed in Table 4. As presented in the table, our approach achieves the best accuracy on UCSD dataset with lowest MAE and RMSE.

5.2. **Ablation experiments.** In addition to comparing with other methods, we also made an extensive ablation study on different variants of the proposed multi-scale dilated convolutional neural network to further verify the effectiveness of different components of the proposed crowd counting approach.

5.2.1. *Multi-scale dilated block.* To further understand the benefits of the proposed multi-scale dilated block, we design an extensive comparison experiment. Specifically, we study the number of branches of the proposed multi-scale dilated block. We set the number

TABLE 3. Results of MAE on the WorldExpo'10 dataset

Methods	Scene 1	Scene 2	Scene 3	Scene 4	Scene 5	Average
Zhang et al. [8]	9.8	14.1	14.3	22.2	3.7	12.9
MCNN [6]	3.4	20.6	12.9	13	8.1	11.6
NetVLAD [27]	3.7	15.9	10.2	15.2	6.7	10.3
Switch-CNN [10]	4.4	15.7	10	11	5.9	9.4
D-ConvNet [28]	**1.9**	12.1	20.7	8.3	**2.6**	9.1
CP-CNN [11]	2.9	14.7	10.5	10.4	5.8	8.86
ic-CNN [26]	17	12.3	9.2	**8.1**	4.7	10.3
CSRNet [2]	2.9	11.5	8.6	16.6	3.4	8.6
MLCNN [30]	2.7	13	13.4	17.1	3.5	9.94
MSDNet (ours)	2.7	**10.6**	**8.1**	12.9	2.8	**7.4**

TABLE 4. Results on the UCSD dataset

Methods	MAE	RMSE
Zhang et al. [8]	1.6	3.31
MCNN [6]	1.07	1.35
Switch-CNN [10]	1.62	2.1
ACSCP [12]	1.04	1.35
CSRNet [2]	1.16	1.47
MSDNet (ours)	**1.01**	**1.21**

TABLE 5. Results on different architectures for MDB on UCF_CC_50

Setting of dilation rates	MAE	RMSE
[1, 3]	269.1	396.3
[1, 3; 3, 5]	251.6	355.9
[1, 3; 3, 5; 5, 7]	243.1	332.7
[1, 3; 3, 5; 5, 7; 7, 9]	264.7	393.2
[1, 3; 3, 5; 5, 7; 7, 9; 9, 11]	289.5	401.3

of branches to be 1, 2, 3, 4, and 5. We do not use any feature fusion methods to only verify the performance of the MDB. Then we observe the counting results respectively. We conduct this experiment on the most challenging UCF_CC_50 dataset. The experimental results are shown in Table 5.

As illustrated in the table, the estimation accuracy becomes higher with the increase of branches at first, and then becomes lower later. The multi-scale dilated block achieves the best accuracy when the number of dilated branches is 3, that is $L = 3$. Excessive branches of dilated convolution also bring harm since it makes the model difficult to converge. This accounts for that we design the multi-scale dilated block with 3 columns. Compared to the architecture with single branch and fixed dilation rate, like CSRNet [2], the proposed multi-scale dilated block architecture achieves a superior performance on both MAE and RMSE.

5.2.2. *Methods for feature fusion.* In Section 3.2, we analyzed the commonly used methods for feature fusion and designed a global attention feature fusion module to merge features

from different levels. To further illustrate the effectiveness of the GAFM on this crowd counting task, we make a comparison experiment using the commonly used feature fusion methods and the GAFM.

We keep the other architecture the same and explore the influence of different feature fusion methods. For element-wise addition, element-wise maximization, and the GAFM, the channel numbers of the output fused feature are the same as the input feature which is not encoded by the MDB. While for concatenation, the channel numbers of the output fused feature are doubled. So the corresponding channel number of convolution kernels for concatenation is also doubled. The experimental results are shown in Table 6. As shown in the table, the GAFM achieves better performance than other commonly used fusion methods. It demonstrates the superiority of the GAFM over other feature fusion modules on this crowd counting task.

TABLE 6. Results of different methods for feature fusion on UCF_CC_50

Fusion methods	MAE	RMSE
Without fusion	243.1	332.7
Element-wise addition	240.9	323.2
Concatenation	236.6	315.6
Element-wise maximization	249.9	338.3
GAFM (ours)	229.7	311.3

It can be also concluded from the table that concatenation performs better than element-wise addition and element-wise maximization. This can be explained to some extent from the analysis in Section 3.2. In Section 3.2, we have analyzed the relationship between element-wise addition and concatenation and pointed out that elementwise addition is a special case of concatenation when the convolution parameters of concatenated features satisfy certain conditions. Concatenation can provide more presentation diversity than element-wise addition and it can cover the case of element-wise addition. Therefore, concatenation can at least achieve similar performance to element-wise addition as long as the convolution parameters are updated to satisfy this condition during the training process. However, it does not mean that concatenation is always better than element-wise addition. The fact cannot be ignored that concatenation often requires more parameters than element-wise addition, which will increase the computation. Therefore, sometimes element-wise addition may be a good complement to concatenation for reducing parameters. As for element-wise maximization, it selects the max value across different channels to form the new feature. However, it can hardly model the correlation between different dimensions of heterogeneous features since different channels often represent different semantic information. In terms of this crowd counting task, it seems to make little sense to simply select the max value across different semantic information. The experimental results also show the performance of element-wise maximization is inferior to other fusion methods, even inferior to without any feature fusion strategies.

5.3. **The advantages of our work.** For the problem of target scale change, MCNN [6] uses multi-branch convolutional neural networks to extract multi-scale features to improve the robustness of the model to target scale changes. However, the introduction of traditional multi-branch convolutional neural networks will greatly increase the amount of model parameters and then will greatly reduce the inference efficiency of the model while increasing the computational complexity of the model and the risk of overfitting. CSRNet [2] uses a dilated convolution kernel with a fixed dilation rate to process image features and regress the density distribution. Due to the limited ability of the dilated

convolution kernel with a fixed expansion rate to aggregate multi-scale information, its counting accuracy also needs to be improved. Therefore, based on the characteristics that dilated convolution can reduce the network parameters while expanding the network receptive field, this paper proposes to use dilated convolutions with different dilation rates to build a multi-scale dilated convolution module to aggregate multi-scale information and improve the model's accuracy. Adaptability to target scale changes. At the same time, this paper also introduces a new feature fusion method, global attention fusion, to further improve the representation ability of the model and greatly improve the accuracy of dense target counting in complex scenes. The algorithm can improve the dense target counting performance in the scene of target scale change and target density uneven distribution, which is of great significance for practical applications.

It can be seen from the visualization result that the density distribution map predicted by the scheme described in this paper is very close to the real density distribution map of the scene image, and the predicted number of targets is also less different from the actual number of targets, indicating that the scheme can more accurately predict the number of objects in the scene image. At the same time, it can be seen from the comparison of the decreasing target scale and the increasing density distribution from top to bottom that the scheme described in this paper can better adapt to target scenes of different scales and densities, which reflects the high robustness of the scheme described in this paper to scale and density changes.

6. **Conclusion.** In this paper, we proposed an efficient CNN-based approach for accurate crowd counting in complex crowd scene. We designed a deep convolutional neural network with multi-scale dilated block and channel attention feature fusion module to achieve an accurate estimation of the crowd count. A group of corresponding experiments were conducted on the benchmark datasets to evaluate the superior performance of the proposed approach for crowd counting. We designed ablation experiments to further study the architecture of the multi-scale dilated block and the global attention fusion module. We also specifically analyzed other popular methods for feature fusion and experimentally verified the superiority of the global attention fusion module. Remarkable improved results on four different challenging datasets with different crowd densities and much less parameters established the superiority of the proposed model over other state-of-the-art approaches. Future work will focus on using some other ideas [31] to further improve our method.

Acknowledgment. This work is supported in part by the National Key Research and Development Program of China under Grant No. 2020AAA014004. This research work is also partially supported by Ningbo Science and Technology innovation 2025 major project under Grants No. 2021Z010 and No. 2021Z063, and the Public Good Research Project of Science and Technology Program of Zhejiang Province under Grant No. LGG21F020005. The authors also gratefully acknowledge the helpful comments and suggestions of the reviewers, which have improved the presentation.

REFERENCES

[1] L. C. Chen, G. Papandreou, I. Kokkinos, K. Murphy and A. L. Yuille, DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs, *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol.40, no.4, pp.838-848, 2018.

[2] Y. Li, X. Zhang and D. Chen, CSRNet: Dilated convolutional neural networks for understanding the highly congested scenes, *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.1091-1100, 2018.

[3] K. He, X. Zhang, S. Ren and J. Sun, Deep residual learning for image recognition, *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.770-778, 2016.
[4] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed and D. Anguelov, Going deeper with convolutions, *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.1-9, 2015.
[5] W. Yin, S. Ebert and H. Schütze, Attention-based convolutional neural network for machine comprehension, *arXiv.org*, arXiv: 1602.04341, 2016.
[6] Y. Y. Zhang, D. Zhou, S. Chen, S. Gao and Y. Ma, Single-image crowd counting via multi-column convolutional neural network, *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.589-597, 2016.
[7] H. Idrees, I. Saleemi, C. Seibert and M. Shah, Multi-source multiscale counting in extremely dense crowd images, *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.2547-2554, 2013.
[8] C. Zhang, H. Li, X. Wang and X. Yang, Cross-scene crowd counting via deep convolutional neural networks, *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.833-841, 2015.
[9] A. B. Chan, Z.-S. J. Liang and N. Vasconcelos, Privacy preserving crowd monitoring: Counting people without people models or tracking, *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.1-7, 2008.
[10] D. B. Sam, S. Surya and R. V. Babu, Switching convolutional neural network for crowd counting, *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.4031-4039, 2017.
[11] V. A. Sindagi and V. M. Patel, Generating high-quality crowd density maps using contextual pyramid CNNs, *IEEE International Conference on Computer Vision (ICCV)*, pp.1879-1888, 2017.
[12] Z. Shen, Y. Xu, B. Ni, M. Wang, J. Hu and X. Yang, Crowd counting via adversarial cross-scale consistency pursuit, *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.5245-5254, 2018.
[13] D. Deb and J. Ventura, An aggregated multicolumn dilated convolution network for perspective-free counting, *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp.308-309, 2018.
[14] K. Simonyan and A. Zisserman, Very deep convolutional networks for large-scale image recognition, *arXiv.org*, arXiv: 1409.1556v6, 2014.
[15] L.-C. Chen, G. Papandreou, F. Schroff and H. Adam, Rethinking atrous convolution for semantic image segmentation, *arXiv.org*, arXiv: 1706.05587v3, 2017.
[16] P. Wang, P. Chen and Y. Yuan, Understanding convolution for semantic segmentation, *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2018.
[17] T. Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan and S. Belongie, Feature pyramid networks for object detection, *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.2117-2125, 2017.
[18] M. Liang, L. Jiao, S. Yang, F. Liu, B. Hou and H. Chen, Deep multiscale spectral-spatial feature fusion for hyperspectral images classification, *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol.11, no.8, pp.2911-2924, 2018.
[19] G. Huang, Z. Liu, L. V. D. Maaten and K. Q. Weinberger, Densely connected convolutional networks, *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.4700-4708, 2017.
[20] J. P. Robinson, Y. Li, N. Zhang and Y. Fu, Laplace landmark localization, *arXiv.org*, arXiv: 1903.11633, 2019.
[21] L. Chen, C. Wu, W. Fan, J. Sun and S. Naoi, Adaptive local receptive field convolutional neural networks for handwritten Chinese character recognition, *Chinese Conference on Pattern Recognition (CVPR)*, pp.455-463, 2014.
[22] M. Lin, Q. Chen and S. Yan, Network in network, *arXiv.org*, arXiv: 1312.4400, 2013.
[23] J. Hu, L. Shen, S. Albanie, G. Sun and E. Wu, Squeeze-and-excitation networks, *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol.42, no.8, pp.2011-2023, 2020.
[24] S. Huang, X. Li, Z.-Q. Cheng, Z. Zhang and A. Hauptmann, Stacked pooling: Improving crowd counting by boosting scale invariance, *arXiv.org*, arXiv: 1808.07456, 2018.
[25] L. Wang, B. Yin, A. Guo, H. Ma and J. Cao, Skip-connection convolutional neural network for still image crowd counting, *Applied Intelligence*, vol.48, no.10, pp.3360-3371, 2018.
[26] V. Ranjan, H. Le and M. Hoai, Iterative crowd counting, *arXiv.org*, arXiv: 1807.09959, 2018.
[27] Z. Shi, L. Zhang, Y. Sun and Y. Ye, Multiscale multitask deep NetVLAD for crowd counting, *IEEE Trans. Industrial Informatics*, vol.14, no.11, pp.4953-4962, 2018.

[28] V. A. Sindagi and V. M. Patel, Inverse attention guided deep crowd counting network, *The 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pp.1-8, 2019.
[29] D. Ttito, R. Quispe, A. R. Rivera and H. Pedrini, Where are the people? A multi-stream convolutional neural network for crowd counting via density map from complex images, *2019 International Conference on Systems, Signals and Image Processing (IWSSIP)*, pp.241-246, 2019.
[30] X. Jiang, L. Zhang, P. Lv, Y. Guo, R. Zhu, Y. Li and M. Xu, Learning multi-level density maps for crowd counting, *IEEE Trans. Neural Networks and Learning Systems*, vol.31, no.8, pp.2705-2715, 2020.
[31] T. T. H. Vu and T.-W. Chang, Artificial neural network-based milk pasteurization quality prediction system, *ICIC Express Letters, Part B: Applications*, vol.13, no.2, pp.171-178, 2022.

Author Biography

Meilei Lv received the B.Sc. degree in industrial electrical automation from Zhejiang College of Technology, China, 1991; the M.Sc. degree in mechanical and electronic engineering from Zhejiang University of Technology, China, 2005.

Prof. Lv is currently a full-time professor in the College of Electrical and Information Engineering, Quzhou University. Her main research interests include online monitoring, fault diagnosis, signal processing, information security, etc. She has published over 30 papers in journals and conferences.

Kuncai Zhang received the B.Sc. degree in communication engineering from Harbin Engineering University, Harbin, China, 2018; the M.Sc. degree in areospace information technology from Zhejiang University, China, 2021.

Mr. Zhang is currently working in Alibaba, China. His main research interests include deep learning and image analysis.

Xiaoyun Zheng received the B.Sc. degree in industrial and civil architecture from Tongji University, China, 1997.

Mr. Zheng is currently a full-time senior engineer at State Grid Wenzhou Power Supply Company, China. He formerly worked for State Grid Quzhou Power Supply Company, China. His main research interests include power network planning, new energy construction, power transmission engineering, etc.

Wei Yang received the B.Sc. degree in computer science and technology from Northeast Electric Power College (now Northeast Electric Power University), China, 2003; the M.Sc. degree in electrical engineering from Zhejiang University, China, 2011.

Mr. Yang is currently a full-time senior engineer at State Grid Quzhou Power Supply Company, China. His main research interests include application of technology informatization in power systems, smart grid, etc. Also he is an Enterprise Level 1 Human Resources Teacher.

Zhe-Ming Lu received the B.Sc. degree in electrical engineering from Harbin Institute of Technology, China, 1995; the M.Sc. degree in electrical engineering from Harbin Institute of Technology, China, 1997; the Ph.D. degree in instrument science and technology from Harbin Institute of Technology, China, 2001.

Prof. Lu is currently a full-time professor at the School of Aeronautics and Astronautics, Zhejiang University, China. His research interests include multimedia single analysis and processing, information hiding and astronautics signal processing, etc. He has published over 300 papers in journals and conferences.