Computers, Materials & Continua

Tech Science Press

DOI: 10.32604/cmc.2024.048928

ARTICLE

Lightweight Res-Connection Multi-Branch Network for Highly Accurate Crowd Counting and Localization

Mingze Li, Diwen Zheng and Shuhua Lu*

College of Information and Cyber Security, People's Public Security University of China, Beijing, 102600, China

*Corresponding Author: Shuhua Lu. Email: lushuhua@ppsuc.edu.cn

Received: 22 December 2023 Accepted: 18 March 2024 Published: 15 May 2024

ABSTRACT

Crowd counting is a promising hotspot of computer vision involving crowd intelligence analysis, achieving tremendous success recently with the development of deep learning. However, there have been still many challenges including crowd multi-scale variations and high network complexity, etc. To tackle these issues, a lightweight Res-connection multi-branch network (LRMBNet) for highly accurate crowd counting and localization is proposed. Specifically, using improved ShuffleNet V2 as the backbone, a lightweight shallow extractor has been designed by employing the channel compression mechanism to reduce enormously the number of network parameters. A light multi-branch structure with different expansion rate convolutions is demonstrated to extract multi-scale features and enlarged receptive fields, where the information transmission and fusion of diverse scale features is enhanced via residual concatenation. In addition, a compound loss function is introduced for training the method to improve global context information correlation. The proposed method is evaluated on the SHHA, SHHB, UCF-QNRF and UCF_CC_50 public datasets. The accuracy is better than those of many advanced approaches, while the number of parameters is smaller. The experimental results show that the proposed method achieves a good tradeoff between the complexity and accuracy of crowd counting, indicating a lightweight and high-precision method for crowd counting.

KEYWORDS

Crowd counting; Res-connection; multi-branch; compound loss function

Symbol	Description
L_E;L_C;L	Euclidean distance loss; Global context loss; Total loss function
X_i	The i th picture
D_i^{GT}	The ground truth density map corresponding to the i th picture
Θ; α; σ	The parameters of the model; Weight coefficient; Distance threshold
$G(X_i;\theta)$	The density map obtained by the model using parameter θ
MAE	The mean absolute error
MSE	The root-mean-squared error
C_i^{ET};C_i^{GT}	The number of people in the ith image estimated by the proposed network; Truth in the i th image

TP	The distance between the predicted head position and the real marker points less than the pixel threshold
FP	The distance between them greater than the pixel threshold
FN	The extracted position matched the real marker point does not exist

1 Introduction

Crowd counting, as a significant task in computer vision, can accurately calculate the number and density distribution of the crowd in the image or video, which is widely used in extensive fields including security monitoring, urban planning, scene understanding traffic management, etc., hence attracting considerably increasing interest [1,2]. Traditional crowd counting methods often rely on handcraft features based on machine learning, suffering from low accuracy and robustness when applied to complex scenarios. In recent years, with the development of computing power and algorithms, deep learning has gained huge success in artificial intelligence [3,4]. And, crowd counting methods based on deep learning have been developed rapidly, achieving end-to-end high precision and robust counting [5–7].

In realistic applications, it is difficult to quickly and accurately calculate the number of people from images due mainly to some challenges, such as uneven illumination, complex backgrounds and large-scale variations. Many efforts have been devoted to designing various networks and tricks including multi-scale fusion structures and attention mechanisms to tackle these problems mentioned above, making remarkable progress [6–9]. As far as we know, Zhang et al. [8] have for the first time attempted to design the three-column convolution network, named multi-column convolutional neural network (MCNN), obtaining multi-scale feature fusion for crowd counting. Subsequently, two-branch [9] and multi-branch [10] networks have been proposed widely, enormously improving the accuracy of counting. Besides, the single-column network with different expansion rate convolutions has been used to effectively resolve the multi-scale variation.

In addition, to reduce the computational complexity of crowd counting methods and improve their real-time performance, many lightweight structures have been proposed such as MobileCount [11], lightweight multi-scale adaptive network (ligMSANet) [12], lightweight scale-aware network (LSANet) [13], lightweight multi-scale network (LMSNet) [14] and so on, providing a good reference for the efficient crowd counting. For instance, Jiang et al. [12] proposed LigMSANet, obtaining multi-scale fusion and real-time counting. More recently, Chavan et al. proposed CrowdDCNN [15] for real-time crowd counting on Internet of Things (IoT) edge. However, in practical applications, especially in edge devices, the high efficiency and accuracy of crowd counting methods still need to be further improved. Specifically, on the one hand, in order to improve the counting accuracy, complex network structures are usually designed, resulting in a large increase in parameters and calculation time. On the other hand, lightweight networks can improve computational efficiency, while they suffer from low counting accuracy in complicated scenarios originating from insufficient learning ability. Therefore, it is of great significance to further explore efficient and precise crowd counting methods.

In this article, a lightweight and powerful multi-branch network, named LRMBNet, for crowd counting and localization embedded in an improved ShuffleNet V2 as backbone is proposed, which can on the one hand extract multi-scale features, and on the other hand reduce considerably the number of network parameters via the design of the channel compression mechanism (CCM). Particularly, in the proposed multi-branch network, the residual connection is tailored to perform multi-scale feature fusion and enhance the diverse information transfer. In summary, the main contributions of this article are as follows:

- We propose a lightweight and powerful multi-branch network for crowd counting and localization, where an improved ShuffleNet V2 is used as a lightweight shallow extractor and then CCM has been designed to further reduce the number of network parameters.
- We stack three multi-branch modules to extract scale diversity features, where the residual connection is tailored to perform the concatenation operation between different branches to enhance the information fusion and transmission. In addition, a compound loss function is introduced to train the proposed method, gaining the global information correlation.
- Extensive experimental results demonstrate that the proposed method achieves superior performance over many advanced methods for crowd counting, indicating a good tradeoff in efficiency and accuracy. Accordingly, it is a promising method for crowd counting and localization in realistic scenarios.

2 Related Work

2.1 Crowd Counting

During the past few decades, many approaches have been proposed for crowd counting. Based on the development of image processing, the methods for crowd counting are broadly divided into two categories: Traditional methods and deep learning-based ones.

Generally, there are two main types of methods for crowd counting: Detection-based methods and regression-based methods. Early detection-based methods usually design handcrafted shallow features to detect body parts for counting, while they perform poorly in crowded scenes due mainly to large-scale variations and heavy occlusions. Regression-based methods establish a regression model for image features and number of people, to estimate the number of people in the scene by extracting features. Regression-based methods avoid these problems mentioned above but lose the ability to capture spatial information about the crowd in many cases [16].

Recently, deep learning-based methods have dominated the development of related technologies, most of which usually generate density maps and sum all pixels in those maps to get the total number of people [7,17]. Depending on the network structure, crowd counting algorithms can be divided into multi-column and single-column networks. Multi-column-based methods usually divide the network into multiple columns to extract multi-scale features and combine them at the output layer. For instance, to solve the challenging problem of scale variations in images, Zhang et al. [8] propose a multi-column counting network MCNN to extract features at different scales. Since then, multi-scale fusion structures with multi-column/branches have developed rapidly [10,18]. Zeng et al. [18] propose a multi-branch crowd counting network, which consists of a front end network and a back-end network. The front end is a conventional convolutional neural network. The back end uses a maximum scale combination strategy to learn different levels of scale information. On the other hand, single-column structures with different convolution kernels have also been proposed to achieve multi-scale fusion. Among them, congested scene recognition network (CSRNet) [7] is the representative of the single column methods, where the front-end uses the first 10 layers of VGG16 to extract features and the back-end uses dilated convolution to expand receptive fields. This network is simple in structure, but good at processing multi-scale variation information. Furthermore, in order to improve feature scale continuity and information transfer capability, Dai et al. [19] propose a single-column deep counting network, which consists of three densely expanded convolutional blocks. The blocks of convolutions are connected by residual connection. In addition, the application of the Transformer model for crowd counting is developing rapidly. TransCrowd [20] uses an approach based on the attentional mechanism to focus on the most informative regions of the crowd, leading to more accurate crowd counting.

2.2 Lightweight Crowd Counting

Although the methods based on density estimation have achieved excellent counting results, there have also been some new problems to resolve, such as redundant network structures, large numbers of model parameters, and some difficulties in training, resulting in poor performance in real-time counting. Therefore, to tackle these issues, various lightweight networks for crowd counting have been designed [11–14], which are roughly divided into two categories: Lightweight structure and model compression methods.

To achieve highly efficient counting, many lightweight structures have been designed based on CNN. MCNN [8] is an early multi-column network with a lightweight structure that extracts head features at different scales according to the different sizes of the convolutional kernel. Cascaded multi-task learning (CMTL) [21] is a multi-task framework that uses prior knowledge of classification as an auxiliary branch of the model to improve counting performance. Perspective crowd counting network (PCCNet) [22] is an improvement network based on CMTL, which uses a priori knowledge of background segmentation to improve counting accuracy. To improve the accuracy and efficiency of crowd counting, many lightweight methods with high accuracy have been proposed [11,23,24]. MobileCount [11] is an example of a lightweight framework used directly for crowd counting. It is a combination of the lightweight networks MobileNetV2 [25] and RefineNet [26]. The lightness of MobileNetCount is mainly due to the use of the first 4 bottleneck blocks of MobileNetV2 as a front-end, resulting in a significant reduction in the model parameter.

Alternatively, the model compression methods have been proposed via several operations such as pruning and knowledge distillation on the original complex CNN crowd counting framework to reduce the number of parameters and improve the counting speed without affecting the accuracy. Shi et al. [27] propose Compact-CNN (C-CNN). They directly compressed the multicolumn framework of MCNN, considering that a layer of convolutional kernels of different sizes can extract different spatial features, thus reducing the multicolumn redundancy of the MCNN. The method of knowledge distillation requires an effective but large parametric model of the teacher to induce a small parametric model of the student for training. Liu et al. [28] propose a new multi-layer knowledge distillation method. This method uses the original CSRNet network as the teacher model and 1/4 channel CSRNet as the student model, and trains using knowledge distillation. This structure allows the small model to achieve similar performance as the original model, but a significantly lower number of parameters and an improved efficiency.

3 The Proposed Network

3.1 Overview

To extract multi-scale features, and yet reduce the computational complexity, we propose a lightweight and powerful multi-branch network for crowd counting and localization. The overview framework is shown in Fig. 1, which is mainly composed of lightweight shallow extractor and three multi-scale fusion modules. The former is used to extract shallow features, reducing the number of channels in the network. The latter is proposed to extract multi-scale features, improving the multi-scale fusion capability. Specifically, the lightweight shallow extractor, composed of an improved ShuffleNet V2 and CCM, is designed as an extremely simple backbone to extract shallow features efficiently, where CCM is devised to replace 1024 convolution of ShuffleNet to reduce parameters. The mid-end of the proposed network consists of three residual connection multi-branch modules (RCMBs), where each RCMB is designed with five branches, adopting various expansion rate

convolutions in different branches to obtain multi-scale features. That is, different from other multi-branch structures, we innovatively propose a Res-connection to link different branches, enhancing the fusion and transmission of multiple feature information. At the end of the proposed network, 1×1 convolution is used to generate density maps. And a compound loss function is introduced to train the proposed method, enhancing the global information correlation.

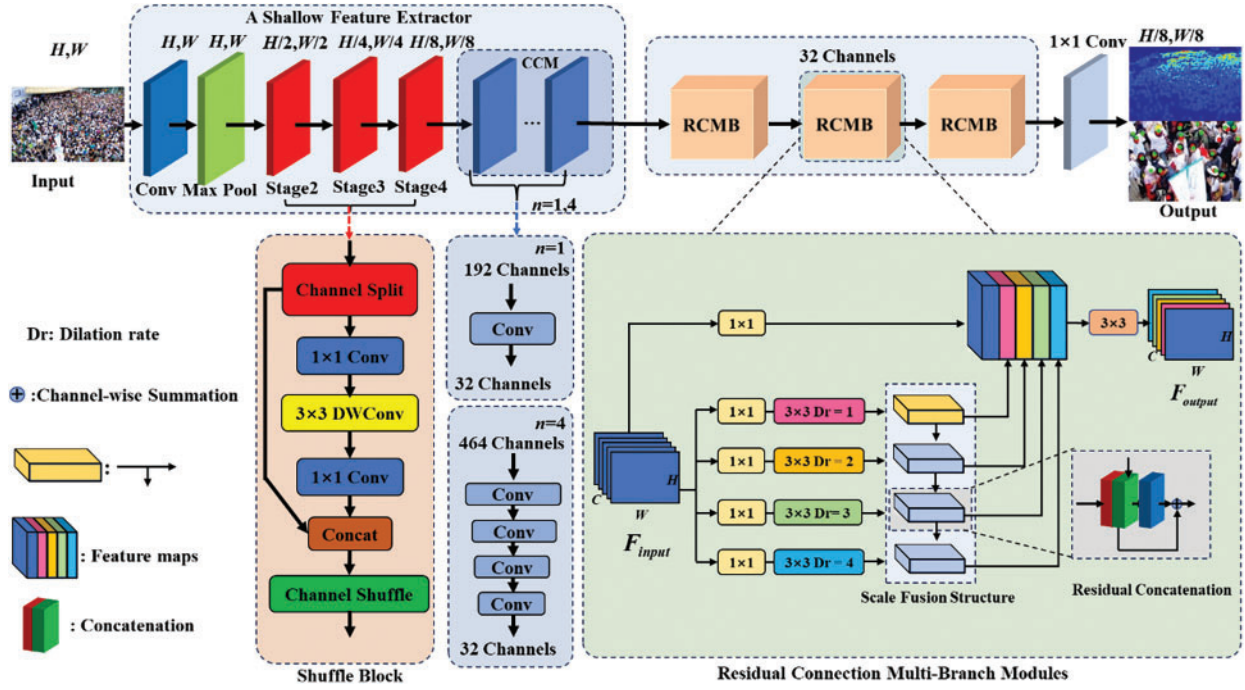

Figure 1: The overview framework of the proposed crowd counting network

3.2 The Lightweight Backbone

As shown in Fig. 1, using the improved ShuffleNet V2 and CCM, the lightweight shallow extractor is designed as backbone to reduce the number of model parameters and simplify the calculation, where CCM replaces the final convolution and pooling layers of the ShuffleNet V2 block. It is noted that the number of block channels on ShuffleNet V2 is scaled by 0.5, 1.0, 1.5, and 2.0 times to generate ShuffleNet V2 networks with different complexity, respectively. They are marked correspondingly with ShuffleNet V2 0.5, ShuffleNet V2 1.0, ShuffleNet V2 1.5, and ShuffleNet V2 2.0. Here, the pre-trained ShuffleNet V2 1.0 and ShuffleNet V2 0.5 are utilized to extract features, respectively, and the corresponding methods are called as V1.0 and V0.5.

CCM is designed with different 3×3 convolutions, where the convolutional number and channel number are adjusted according to the different network structures, resulting in controlling 32 output channels. That is, the number of output channels is compressed from 464 to 32 in ShuffleNet V2 1.0 via four 3×3 convolution operation. And the number of output channels is compressed from 192 to 32 in ShuffleNet V2 1.0 via one 3×3 convolution operation. As a result, the redundant information can be vastly removed by CCM, reducing the calculation amount remarkably.

3.3 Residual Connection Multi-Branch Module

To enhance the fusion and transmission of multi-scale features, we propose a residual-connection improved multi-branch module, as shown in Fig. 1, which consists of one 1×1 convolution branch and another 4 branches. Specifically, among 4 branches, each branch is composed of one 1×1 convolution and various 3×3 convolutions with dilation rate of 1, 2, 3 and 4, respectively. The pipeline of the multi-scale feature fusion and transfer is described as follows. Firstly, each branch adjusts the channels of the input feature maps by 1×1 convolution, and expands the receptive fields with the expansion rates of 1, 2, 3 and 4, to enhance the relevance of context information and extract the features of the corresponding scales, coping with large-scale crowd variations. Secondly, the different scale features from 4 branches are fused via residual connection, as shown in Fig. 1. The number of map channels is adjusted by 3×3 convolution, and then the feature maps of each branch are concatenated to improve the information transfer and fusion ability of multi-scale features, and hence leading to an improvement in counting accuracy. Finally, the feature maps originated from 4 branches by residual connection are concatenated directly with the feature maps from one 1×1 convolution branch. And then, the output maps are obtained by controlling the number of channels with a 3×3 convolution.

3.4 Loss Function

Euclidean distance loss L_E is widely used as a supervisor to train the methods for crowd counting by calculating the difference between the ground truth map and the predicted density one [8,7,29]. Considering the characteristics of the proposed multi-branch structure, a global context loss L_C is introduced as the partner of L_E to focus as much as possible on the correlation of global context information. Hence, a compound loss function with different weights of L_E and L_C is adopted to improve the counting accuracy by jointly training the proposed methods, which is defined as follows:

$$L_E = \frac{1}{N}\sum_{i=1}^{N}\left|G\left(X_i;\theta\right) - D_i^{GT}\right|_2^2 \tag{1}$$

$$L_C = \frac{1}{N}\sum_{i=1}^{N}\left|P_{avg}\left(G\left(X_i;\theta\right)\right) - P_{avg}\left(D_i^{GT}\right)\right| \tag{2}$$

$$L = L_E + \alpha L_C \tag{3}$$

where N is the number of training set images. X_i and D_i^{GT}, respectively, represent the i th picture and the ground truth density map corresponding to the i th picture. θ represents the parameters of the model and $G\left(X_i;\theta\right)$ represents the density map obtained by the model using parameter θ. Based on extensive experiments on the SHHA datasets, the weight coefficients α is set to 1000.

4 Experiments and Result Analysis

4.1 Evaluation Metrics

4.1.1 Counting Metrics

To evaluate quantitatively the prediction accuracy and concentricity of the proposed network, the mean absolute error (MAE) and the root-mean-squared error (RMSE), widely employed in crowd counting, are introduced as the evaluation metrics, defined as follows:

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}\left|C_i^{ET} - C_i^{GT}\right| \tag{4}$$

$$\mathrm{MSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left|C_i^{ET} - C_i^{GT}\right|^2} \tag{5}$$

where N stands for the total number of the test pictures. C_i^{ET} represents the number of people in the ith image estimated by the proposed network. C_i^{GT} represents the number of people labelled in the ith image.

4.1.2 Localization Metrics

To perform the quantitative evaluation of the proposed network for crowd localization, we adopt Precision, Recall and F-measure as localization Metrics. When the distance between the predicted point Pp and the labelled point Pg is less than the distance threshold σ, it indicates that Pp and Pg are successfully matched. In the SHHA dataset, two fixed thresholds ($\sigma = 4, 8$) are selected for evaluation. The specific formulas are defined as follows:

$$Precision = \frac{TP}{TP + FP} \tag{6}$$

$$Rcall = \frac{TP}{TP + FN} \tag{7}$$

$$F - measure = \frac{2 * Precision * Rcall}{Precision + Rcall} \tag{8}$$

where TP represents the distance between the predicted head position point and the real marker point less than the pixel threshold. FP represents the distance between them greater than the pixel threshold. FN represents that the extracted position matched the real marker point does not exist.

4.2 Datasets

We have conducted experiments on several current popular datasets, including sparse scenarios, dense scenarios, and weather variations.

4.2.1 ShangHaiTech

The ShangHaiTech dataset, proposed by Zhang et al. [8] in 2016, is divided into two parts depending on the crowd density: Part_A and Part_B, referred to as SHHA and SHHB, respectively. It has a total of 1198 images and 330,165 annotation headers. SHHA has 482 images, mainly derived from crowd images on the Internet. 300 images are used for training and 182 images are used for testing. SHHB has 716 images, mainly sourced from images of Shanghai city area. 400 images are used for training and 316 images are used for testing.

4.2.2 UCF_QNRF

The UCF-QNRF dataset, proposed by Ideers et al. [30] in 2018, contains 1535 crowd images and 1,251,642 annotated headers in total, among which 1201 images for training and 334 images for testing. It is the dataset with the largest number of individual images at the time. The images in the UCF-QNRF dataset are mainly derived from web searches and sights like the Hajj pilgrimage to Mecca.

4.2.3 UCF_CC_50

The UCF_CC_50 dataset [31] contains 50 greyscale images and has a total of 63,874 annotation headers containing a variety of complex scenes. The number of people in each image is between 94

and 4543. In the experiments, we adopt respectively 40 and 10 of these images to train and validate the proposed method with a 5-fold cross-validation paradigm.

4.2.4 JHU-CROWD++

The JHU-CROWD++ dataset [32] is a large-scale crowd counting dataset including 4372 images and 1,515,005 annotations. It contains a variety of challenging scenarios such as density variations, light variations, and weather variations.

4.3 Implementation Details

The training and testing processes have been conducted under Ubuntu 22.04 system. The deep learning framework is PyTorch 1.12 and the programming language is Python 3.8. The GPU used for training is NVIDIA RTX 3090 with 24 GB of video memory and the CPU computer memory is 64 GB. Iteration epochs are set to 1000 and batchsize is set to 16 during training. At the same time, Adam optimizer is used to adjust the learning rate, where the initial value of the learning rate is set to 0.0001 and the decay rate is set to 0.5 every 100 iterations. The images are limited to a minimum width and height of 512 and a maximum width and height of 1920, while maintaining the original image scale and being able to be divisible by 16. Each image is randomly scaled at [0.8, 1.2] and fixed size patches are cropped at random locations. And then random mirroring with 50% probability and [0.5, 1.5] gamma contrast transformation with 30% probability are used. Finally, the color image is changed to grey image with 10% probability.

4.4 Results and Analysis

4.4.1 Crowd Counting

The accuracy of the crowd counting. To demonstrate quantitatively the accuracy of the proposed network, extensive experiments have been conducted on the public challenging datasets, e.g., SHHA, SHHB, UCF-QNRF, UCF_CC_50 and JHU-CROWD++, whose results are compared with other state of the arts (SOTA) methods in terms of MAE, MSE and network parameters, summarized in Table 1. As can be seen from the top half of Table 1, the proposed method (V1.0) with MAE of 53.82 and MSE of 87.35 on SHHA show superiority over most advanced methods with heavy weight networks including CNN and transformer-based approaches, while it exhibits a much lighter structure. Among them, compared to our previous multi-scale feature fusion and attention (MSFFA) method [10], the accuracy of the proposed method (V0.5) is roughly equal, while its number of parameters is much smaller than that of MSFFA. This may be attributed to the lightweight feature extractor, especially to the channel compression mechanism. Objectively, the accuracy of the proposed model is slightly lower than that of Point to point network (P2PNet), but its number of network parameters is much smaller. From the bottom half of Table 1, When compared with lightweight networks for crowd counting, the proposed method outperforms the comparative approaches by a large margin across all the four datasets, indicating excellent robustness both in sparse and highly crowded scenarios. Amongst, compared to other similar multi-branch methods like LigMSANet [12], LMSNet [14] and Lightweight multiscale feature fusion network (LMSFFNet) [33], our multi-branch network performs well in terms of accuracy and parameters. We ascribe it largely to the introduction of residual connection in the multi-scale fusion structure, improving the diversity information transmission and fusion. In the dataset JHU-CROWD++, which contains different locations with weather variations, the proposed method (V1.0) achieves SOTA results, indicating its effectiveness in dealing with different

scenarios. In summary, the proposed method has strong competitiveness in the accuracy and efficiency of crowd counting.

Table 1: Comparison results on the different datasets

Methods	Params. (M)	SHHA		SHHB		UCF-QNRF		UCF_CC_50		JHU-CROWD++	
		MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE
Switching CNN (2017) [6]	15.1	90.4	135.0	21.6	33.4	228.0	445.0	318.1	439.2	–	–
CSRNet (2018) [7]	16.26	68.2	115.0	10.6	16.0	120.3	208.5	266.1	397.5	85.9	309.2
LSC-CNN (2020) [34]	20.6	66.4	117.0	8.1	12.7	–	–	225.6	302.7	–	–
TransCrowd (2022) [20]	90.4	66.1	105.1	9.3	16.1	97.2	168.5	–	–	–	–
BL (2019) [35]	21.5	62.8	101.8	7.7	12.7	88.7	154.8	229.3	308.2	105.4	454.2
CAN (2019) [36]	18.10	62.3	100.0	7.8	12.2	107.0	183.0	212.2	**243.7**	100.1	314.0
UEPNet (2021) [37]	26.21	54.64	91.15	6.38	10.88	**81.13**	**131.68**	**165.24**	275.9	–	–
CTASNet (2022) [38]	30.3	54.3	87.8	6.5	10.7	85.6	148.3	211.0	291.5	–	–
P2PNet (2021) [39]	18.34	**52.74**	**85.06**	**6.25**	**9.9**	85.32	154.5	172.72	256.2	–	–
SFPANet (2023) [40]	–	65.7	106.4	7.4	11.7	–		–	–	–	–
CHS-Net (2023) [41]	–	59.2	97.8	7.1	12.1	83.4	144.9	–	–	–	–
PESSNet (2023) [42]	21.75	57.3	95.9	6.4	9.9	85.3	154.5	–	–	–	–
DMCNet (2023) [43]	–	58.46	84.55	8.64	13.67	96.52	163.99	–	–	–	–
MSFFA (2023) [10]		59.5	96.3	6.7	11.6	94.6	170.6	190.2	263.2	–	–
LRMBNet V1.0 (Ours)	**2.29**	53.82	87.35	6.64	11.14	84.43	147.78	184.0	263.4	**52.5**	**201.8**
MCNN (2016) [8]	0.13	110.2	173.2	26.4	41.3	243.5	364.7	467.0	498.5	188.9	483.4
LCN (2020) [24]	**0.032**	93.3	157.0	15.1	23.3	262.0	358.6			–	–
MobileCount (2020) [11]	3.4	89.4	146.0	9.0	15.4	131.1	222.6	284.8	392.8	–	–
C-CNN (2020) [27]	0.073	88.1	141.7	14.9	22.1	224.2	331.0	–	–	–	–
LCNet (2019) [24]	0.062	87.0	143.3	13.9	22.4	–	–	–	–	–	–
LigMSANet (2022) [12]	0.63	76.6	121.4	10.9	17.5	–	–	231.5	339.7	–	–
PCCNet (2019) [22]	0.55	73.5	124.0	11.0	19.0	148.7	247.3	240.0	315.5	–	–
SANet (2018) [44]	1.39	67.0	104.5	8.4	13.6	152.6	247.0	–	–	91.1	320.4
LSANet (2020) [13]	0.20	66.1	110.2	8.6	13.9	112.3	186.9	–	–	–	–
LMSFFNet (2023) [33]	4.58	85.85	139.9	9.2	15.1	112.8	201.6	105.7	120.3	–	–
LMSNet (2023) [14]	0.73	62.9	108.4	8.2	13.5	110.7	178.7	223.5	281.0	–	–
LRMBNet V0.5 (Ours)	0.25	**59.94**	**97.38**	**7.62**	**12.82**	**93.90**	**155.08**	**198.1**	**251.9**	**56.0**	**223.0**

To present the distribution of predicted crowd counting points, the scatter and fitting plots are shown in Fig. 2, where the x-axis and y-axis represent the ground truth and the predicted results, respectively. As can be seen, on SHHA, the fitting line is slightly deviated from the theoretical line ($y = x$), which maybe ascribe to the complexity scene in crowd counting but the lightweight network of the proposed method. Fig. 3 shows the comparative results between the predicted density maps generated by our proposed model and the ground truth on the four datasets, where the image samples with sparse and highly congested scenes are randomly selected from the SHHA, SHHB, UCF-QNRF and UCF_CC_50 datasets. It can be seen that the predicted results are approximately consistent with the ground truth in various scenes, which indicates excellent generalization and robustness.

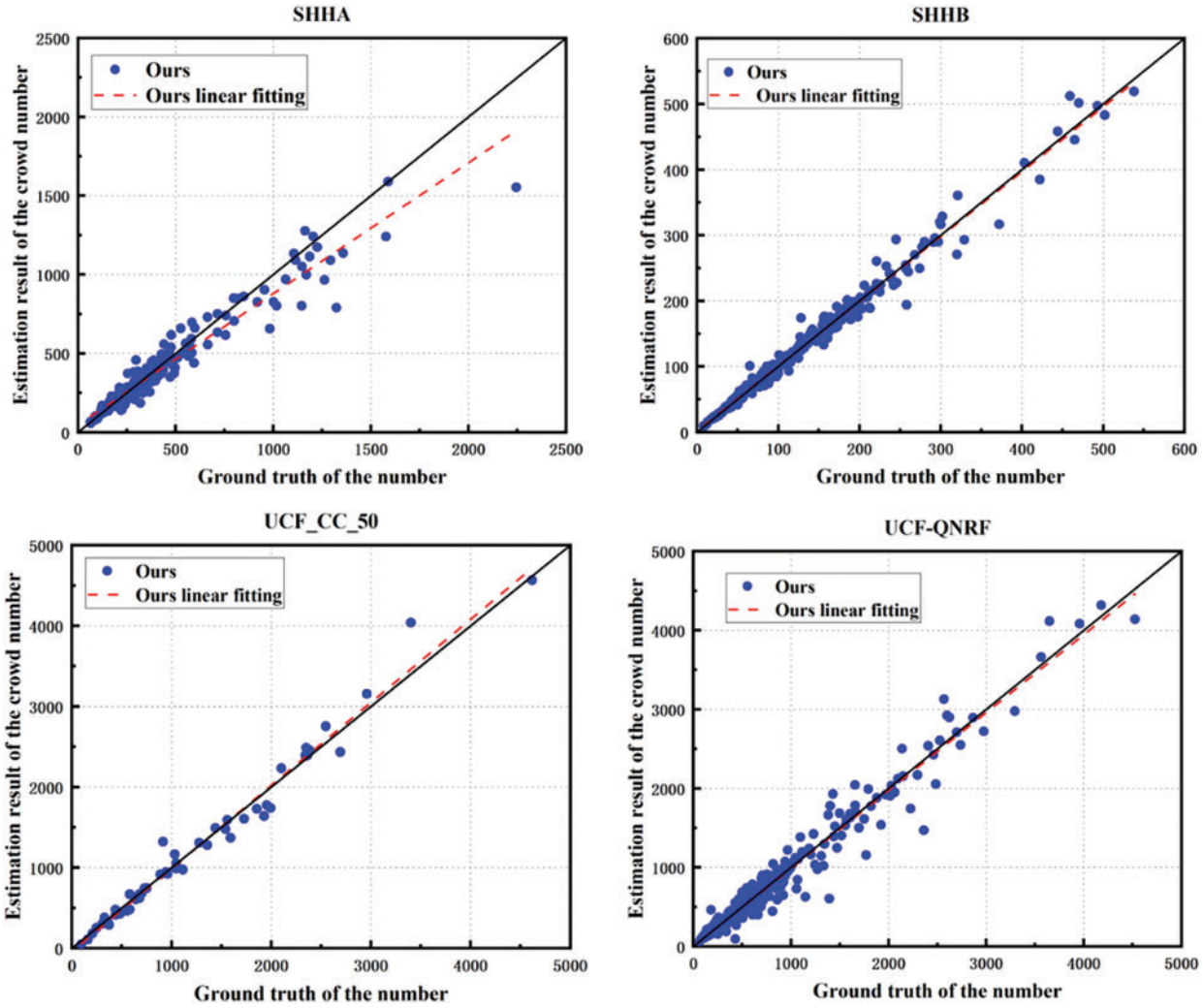

Figure 2: Comparison of the theoretical and estimated values on the SHHA, SHHB, UCF-QNRF and UCF_CC_50 datasets. The blue dots for the results predicted by our LRMBNet V1.0. and the red dotted line for its fitting

The inference speeds. To characterize the efficiency of the proposed method, the experimental results of inference speeds, FLOPs and time on SHHA have been tested on two types of GPU (RTX3090 and GTX1080ti) chips. Table 2 shows the comparative results with other SOTA methos. It is worth pointing out that the experimental results conducted on different GPU are relative values, while they are still of some comparative significance. As can been seen, the inference speeds and time of our method with high accuracy are better than those of most existing SOTA methods, showing a good tradeoff between accuracy, speed, and computational resources. Fig. 4 depicts the visualization comparation results of our method in terms of inference speeds, parameters and MAE with other classic methods including CSRNet [7], MobileCount [11], LigMSANet [12] and MCNN [8]. As shown in Fig. 4, the size of the circle represents the speed, where the larger the circle, the higher the speed. And the proposed methods are in leading positions, indicating their significant advantages in both counting accuracy and network efficiency.

Figure 3: Examples of generation density maps

Table 2: Comparison of the inference speed and FLOPs of different models

Method	Backbone	FLOPs(G)	Input	GPU	FPS	Time (ms)	MAE	MSE
CSRNet [20]	VGG16	108.34	384 × 384	RTX TitanXP	21.67	46.1	68.2	115.0
TransCrowd [20]	Transformer		384 × 384	RTX TitanXP	46.73	21.4	66.1	105.1
MCNN [13]		24.81	1280 × 720	RTX 2080	32.70	30.6	110.2	173.2
CAN [36]	VGG16	114.83	1024 × 768	RTX 3090	21.28	47.0	62.3	100.0
SCAR [38]	VGG16	108.44	1024 × 768	RTX 3090	21.28	47.0	66.3	114.1
CTASNet [38]	VGG16	102.22	1024 × 768	RTX 3090	25.00	40.0	54.3	87.8
MobileCount [11]	MobileNet V2	7.26	1920 × 1080	GTX1080Ti	22.40	44.6	89.4	146.0
LigMSANet [12]	MobileNet V2	**1.3**	400 × 400	RTX2080Ti	23.80	42.0	76.6	121.4
LSANet [13]	VGG16	6.34	1280 × 720	RTX 2080	24.90	40.2	66.1	110.2
Ours (V1.0)	ShuffleNet V2	16.17	Original size	RTX3090	37.45	26.7	**53.82**	**87.35**
				GTX1080Ti	20.44	48.9		
Ours (V0.5)	ShuffleNet V2	2.62	Original size	RTX3090	**49.19**	20.3	59.94	97.38
				GTX1080Ti	**40.44**	**24.7**		

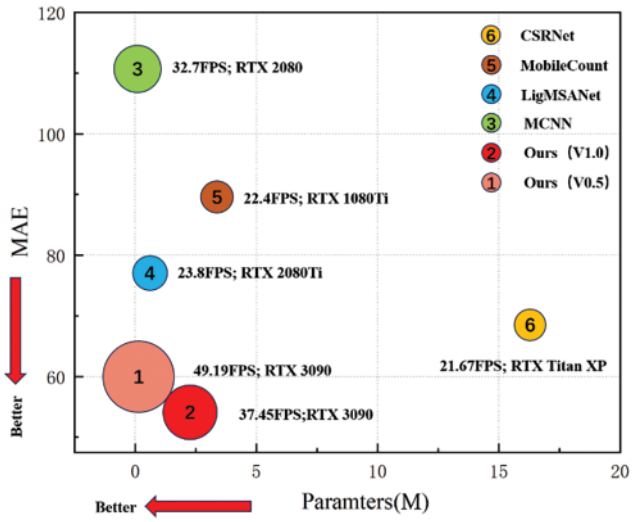

Figure 4: Comparisons with SOTA crowd counting methods in terms of inference speed, MAE, and model parameters on the SHHA dataset

4.4.2 Crowd Localization

Crowd localization is a challenging task associated with crowd counting, which helps to predict the position of each person. To reflect the precise position of individuals, the extensive experiments of crowd localization have been carried out based on the FIDTM framework [45], where the proposed method is adjusted to match FIDTM. Specifically, in the channel compression mechanism, the transposed convolution is used to replace the convolution so as to get 8 times up-sampling, resulting in that the output density map is the same size as the labeled map. The results compared to other methods on SHHA are summarized in Table 3. As can be seen, the crowd localization performance of the proposed method is slightly lower than that of FIDTM, which can be attributed to the lightweight structure. However, it is still higher than many advanced methods. It is noted that the number of FIDTM parameters is about 66 M reproduced in this article according to the open codes. Fig. 5 shows the visualization results of crowd localization estimated on SHHA, where head labeled points are shown in red dots and prediction points are shown in green dots, respectively. It can be seen that the prediction results of the proposed method are roughly in good agreement with the ground truth, demonstrating its excellent performance in accurate crowd localization.

Table 3: Crowd location results on the SHHA dataset

Methods	$\sigma = 4$			$\sigma = 8$		
	P	R	F	P	R	F
FIDTM(LCFCN) [45]	43.3%	26.0%	32.5%	75.1%	45.1%	56.3%
FIDTM (Method in) [45]	34.9%	20.7%	25.9%	67.7%	44.8%	53.9%
FIDTM (TopoCount) [45]	41.7%	40.6%	41.1%	74.6%	72.7%	73.6%
FIDTM (LSC-CNN) [45]	33.4%	31.9%	32.6%	63.9%	61.0%	62.4%
FIDTM (HRNET) [45]	**59.1%**	**58.2%**	**58.6%**	**78.2%**	**77.0%**	**77.6%**
CSRNet [7]	37.7%	35.6%	36.7%	60.0%	56.5%	58.1%

(Continued)

Table 3 (continued)

Methods	$\sigma = 4$			$\sigma = 8$		
	P	*R*	*F*	*P*	*R*	*F*
MSFFA [10]	37.0%	34.2%	35.6%	60.7%	56.1%	58.3%
LRMBNet V1.0 (Ours)	56.4%	55.2%	55.8%	77.7%	76.0%	76.9%

Figure 5: Crowd location visualization results. *P* for *Precision*, *R* for *Recall* and *F* for *F*-measure

To further evaluated the performance under varying conditions such as lighting, weather, or occlusions, the experiments have been conducted using selected images and the results are shown in Fig. 6. As can be seen, our method achieves good results under some various conditions. However, in some extremely hard conditions, the performance degrades due to its relatively simple structure, limiting the learning ability.

Figure 6: (Continued)

Figure 6: Crowd location visualization results in complex scenarios

4.5 Ablation Studies and Network Design

4.5.1 The Effectiveness of Each Component

To evaluate the effectiveness of different components, ablation studies have been performed on the SHHA dataset, in which the network including the lightweight shallow extractor, Euclidean distance loss and density map generator is used the baseline. Based on the baseline, the multi-branch structure without/with residual connection and global loss function are introduced successively. The ablation study results are shown in Table 4, where with the introduction of each component, the performance of the proposed network grows gradually. As can be seen, the MAE and MSE decrease from 63.56 and 105.78 to 53.82 and 87.35, respectively. Compared with the baseline, the introduction of loss function improves the accuracy significantly. And, using multi-branch structure and residual connection, the counting accuracy is gradually improved. This indicates that the global context loss and RCMB can effectively enhance diversity feature fusion and information transfer. To demonstrate the effectiveness of the components more intuitively, the ablation experiment results are shown in Fig. 7, where the accuracy of the crowd counting is enhanced with the introduction of various tricks.

Table 4: Ablation studies on the SHHA dataset

Components	MSE	MAE	FLOPs (G)	Params.	Time (s)
Baseline	105.78	63.56	15.93	2.23 M	0.0231
Baseline + L_C	91.40	57.20	15.93	2.23 M	0.0236
Baseline + L_C + RCMB*	89.51	55.49	16.12	2.28 M	0.0253
Baseline + L_C + RCMB	**87.35**	**53.82**	16.17	**2.29 M**	**0.0267**

Note: "*" represents for the multi-branch module without residual connection.

4.5.2 Network Design

To design the quantities of RCMB modules, a series of experiments have been conducted on SHHA. The results are shown in Table 5, where the performance gradually improves with the number of modules increasing. In contrast, the performance decreases when the number of modules exceeds 3. As a consequence, considering the performance and the number of parameters, we design a network structure with three RCMB modules. Table 6 shows the experimental results of the effect on the lightweight network of the channel compression mechanism. As can been from Table 6, the parameter amounts of the compressed network (V1.0 and V0.5) decrease from 13.96 and 2.99 M to 2.29 and 0.25 M, respectively. Furthermore, the accuracy of the compressed network (V1.0 and V0.5) improves slightly. We ascribe to the lightweight design possibly reduces redundant information and effectively

improve the generalization of the model, enhancing the counting accuracy. It is concluded that the proposed method has a good tradeoff between the complexity and accuracy.

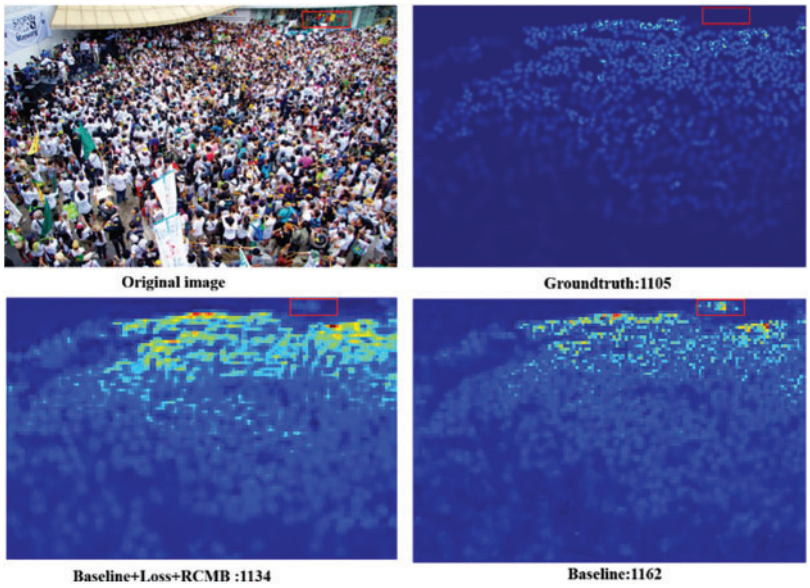

Figure 7: Visualization of the ablation experiment

Table 5: Results of RCMB with different quantities on the SHHA dataset

Metrics	1 RCMB	2 RCMB	3 RCMB	4 RCMB	5 RCMB
MAE	57.08	56.87	**53.94**	55.59	56.21
MSE	92.78	93.71	**89.37**	91.61	90.64

Table 6: Comparison of results with or without channel compression module in SHHA dataset

Compression	MAE	FLOPs (G)	Params.	FPS
V0.5 (w/o)	60.17	13.81	2.99 M	45.5
V1.0 (w/o)	55.26	63.97	13.96 M	25.3
V0.5 (w)	59.94	2.62	0.25 M	49.2
V1.0 (w)	53.82	16.17	2.29 M	37.5

5 Conclusions and Outlook

In this article, we propose a lightweight and powerful multi-branch network, named LRMBNet, improved by a residual connection to enhance the accuracy and efficiency of crowd counting and localization. Principally, we demonstrate a powerful multi-branch structure improved by residual connection to extract multi-scale features, enhancing the information transfer and fusion of diverse scale features. In addition, a lightweight shallow extractor is designed using the improved ShuffleNet

V2 and channel compression mechanism, reducing enormously the number of network parameters. Besides, to improve global context information correlation, a compound loss function is introduced. Extensive experimental results show that the proposed method outperforms many SOTA methods in terms of counting precision and speed, achieving a good tradeoff in efficiency and accuracy of crowd counting. In our future work, application research on edge devices will be conducted to enhance the performance of crowd counting and localization in practical scenarios.

In the outlook, although we have reduced the number of parameters of our method enormously and improved the inference speed, while maintaining accuracy as high as possible, the future deployment on edge devices needed to be conducted as well as for real-time implementation. In addition, the performance under complex conditions in realistic scenarios needs to be further improved by designing more excellent network structures. Besides, the generalization between different domains needs to be further studied in the future when integrating with existing surveillance or traffic systems.

Acknowledgement: None.

Funding Statement: This work is supported by Double First-Class Innovation Research Project for People's Public Security University of China (2023SYL08).

Author Contributions: Li Mingze: Writing, Conceptualization, Methodology, Implementation; Zheng Diwen: Writing–review & editing, Formal analysis, Validation; Lu Shuhua: Funding acquisition, Project administration, Writing–review & editing, Supervision. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: The data that support the findings of this study are available from the corresponding author, Lu S.H., upon reasonable request.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] F. Zhu, H. Yan, X. Chen, T. Li, and Z. Zhang, "A multi-scale and multi-level feature aggregation network for crowd counting," *Neurocomput.*, vol. 423, pp. 46–56, 2021. doi: 10.1016/j.neucom.2020.09.059.

[2] M. Zhu, X. Wang, J. Tang, N. Wang, and L. Qu, "Attentive multi-stage convolutional neural network for crowd counting," *Pattern Recognit. Lett.*, vol. 135, pp. 279–285, 2020. doi: 10.1016/j.patrec.2020.05.009.

[3] R. Sundararaman, C. de Almeida Braga, E. Marchand, and J. Pettre, "Tracking pedestrian heads in dense crowd," in *Proc. Comput. Vis. Pattern Recognit.*, 2021, pp. 3865–3875.

[4] M. Duan, K. Li, K. Li, and Q. Tian, "A novel multi-task tensor correlation neural network for facial attribute prediction," *ACM Trans. Intell. Syst. Technol.*, vol. 12, no. 1, pp. 1–22, 2020.

[5] V. A. Sindagi and V. M. Patel, "A survey of recent advances in cnn-based single image crowd counting and density estimation," *Pattern Recognit. Lett.*, vol. 107, pp. 3–16, 2018. doi: 10.1016/j.patrec.2017.07.007.

[6] D. Babu Sam, S. Surya, and R. Venkatesh Babu, "Switching convolutional neural network for crowd counting," in *2017 IEEE Conf. Comput. Vis. Pattern Recognit.*, Honolulu, Hawaii, USA, 2017, pp. 4031–4039.

[7] Y. Li, X. Zhang, and D. Chen, "CSRNet: Dilated convolutional neural networks for understanding the highly congested scenes," in *Proc. Comput. Vis. Pattern Recognit.*, Salt Lake City, Utah, USA, 2018, pp. 1091–1100.

[8] Y. Zhang, D. Zhou, S. Chen, S. Gao, and Y. Ma, "Single-image crowd counting via multi-column convolutional neural network," in *Proc. Comput. Vis. Pattern Recognit.*, Las Vegas, Nevada, USA, 2016, pp. 589–597.
[9] Y. Wang, W. Zhang, Y. Liu, and J. Zhu, "Two-branch fusion network with attention map for crowd counting," *Neurocomput.*, vol. 411, pp. 1–8, 2020. doi: 10.1016/j.neucom.2020.06.034.
[10] Z. Li, S. Lu, Y. Dong, and J. Guo, "MSFFA: A multi-scale feature fusion and attention mechanism network for crowd counting," *Vis. Comput.*, vol. 39, no. 3, pp. 1045–1056, 2023. doi: 10.1007/s00371-021-02383-0.
[11] P. Wang, C. Gao, Y. Wang, H. Li, and Y. Gao, "MobileCount: An efficient encoder-decoder framework for real-time crowd counting," *Neurocomput.*, vol. 407, pp. 292–299, 2020. doi: 10.1016/j.neucom.2020.05.056.
[12] G. Jiang, R. Wu, Z. Huo, C. Zhao, and J. Luo, "LigMSANet: Lightweight multi-scale adaptive convolutional neural network for dense crowd counting," *Expert. Syst. Appl.*, vol. 197, pp. 116662, 2022. doi: 10.1016/j.eswa.2022.116662.
[13] F. Zhu, H. Yan, X. Chen, and T. Li, "Real-time crowd counting via lightweight scale-aware network," *Neurocomput.*, vol. 472, pp. 54–67, 2022. doi: 10.1016/j.neucom.2021.11.099.
[14] M. Xi and H. Yan, "Lightweight multi-scale network with attention for accurate and efficient crowd counting," *Vis. Comput.*, vol. 30, pp. 1–14, 2023. doi: 10.1007/s00371-023-03099-z.
[15] R. Chavan, G. Rani, P. Thakkar, and V. S. Dhaka, "CrowdDCNN: Deep convolution neural network for real-time crowd counting on IoT edge," *Eng. Appl. Artif. Intel.*, vol. 126, pp. 107089, 2023. doi: 10.1016/j.engappai.2023.107089.
[16] A. B. Chan and N. Vasconcelos, "Bayesian poisson regression for crowd counting," in *2009 IEEE 12th Int. Conf. Comput. Vis.*, 2009, pp. 545–551.
[17] X. Jiang *et al.*, "Attention scaling for crowd counting," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 4706–4715.
[18] X. Zeng, Y. Wu, S. Hu, R. Wang, and Y. Ye, "DSPNet: Deep scale purifier network for dense crowd counting," *Expert. Syst. Appl.*, vol. 141, pp. 112977, 2020. doi: 10.1016/j.eswa.2019.112977.
[19] F. Dai, H. Liu, Y. Ma, X. Zhang, and Q. Zhao, "Dense scale network for crowd counting," in *2021 Int. Conf. Multimed. Retr.*, New York, NY, USA, 2021, pp. 64–72.
[20] D. Liang, X. Chen, W. Xu, Y. Zhou, and X. Bai, "TransCrowd Weakly-supervised crowd counting with transformers," *Sci. China Inf. Sci.*, vol. 65, no. 6, pp. 160104, 2022. doi: 10.1007/s11432-021-3445-y.
[21] V. A. Sindagi and V. M. Patel, "CNN-based cascaded multi-task learning of high-level prior and density estimation for crowd counting," in *Proc. IEEE Int. Conf. Advanced Video Signal Based Surveillance*, Lecce, Italy, 2017, pp. 1–6.
[22] J. Gao, Q. Wang, and X. Li, "PCC Net Perspective crowd counting via spatial convolutional network," *IEEE Trans. Circ. Syst. Video Technol.*, vol. 30, pp. 3486–3498, 2019. doi: 10.1109/TCSVT.2019.2919139.
[23] X. Wu, B. Xu, Y. Zheng, H. Ye, J. Yang and L. He, "Fast video crowd counting with a temporal aware network," *Neurocomput.*, vol. 403, pp. 13–20, 2020. doi: 10.1016/j.neucom.2020.04.071.
[24] X. Ma, S. Du, and Y. Liu, "A lightweight neural network for crowd analysis of images with congested scenes," in *2019 IEEE Int. Conf. Image Process. (ICIP)*, Taipei, Taiwan, 2019, pp. 979–983.
[25] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L. C. Chen, "MobileNetV2 Inverted residuals and linear bottlenecks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, 2018, pp. 4510–4520.
[26] G. Lin, A. Milan, C. Shen, and I. Reid, "RefineNet Multi-path refinement networks for high-resolution semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Honolulu, HI, USA, 2017, pp. 1925–1934.
[27] X. Shi, X. Li, C. Wu, S. Kong, J. Yang and L. He, "A real-time deep network for crowd counting," in *IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, USA, 2020, pp. 2328–2332.
[28] L. Liu, J. Chen, H. Wu, T. Chen, G. Li and L. Lin, "Efficient crowd counting via structured knowledge transfer," in *Proc. 28th ACM Int. Conf. Multimed.*, New York, NY, USA, 2020, pp. 2645–2654.
[29] S. Wang, Y. Lu, T. Zhou, H. Di, L. Lu and L. Zhang, "SCLNet: Spatial context learning network for congested crowd counting," *Neurocomput.*, vol. 404, pp. 227–239, 2020. doi: 10.1016/j.neucom.2020.04.139.

[30] H. Idrees *et al.*, "Composition loss for counting, density map estimation and localization in dense crowds," in *Eur. Conf. Comput. Vis.*, Glasgow, UK, 2018.
[31] H. Idrees, I. Saleemi, C. Seibert, and M. Shah, "Multi-source multi-scale counting in extremely dense crowd images," in *2013 IEEE Conf. Comput. Vis. Pattern Recognit.*, Portland, Oregon, USA, 2013, pp. 2547–2554.
[32] V. A. Sindagi, R. Yasarla, and V. M. Patel, "JHU-CROWD++ Large-scale crowd counting dataset and a benchmark method," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 5, pp. 2594–2609, 2020. doi: 10.1109/TPAMI.2020.3035969.
[33] J. Yi, Z. Shen, F. Chen, Y. Zhao, S. Xiao and W. Zhou, "A lightweight multiscale feature fusion network for remote sensing object counting," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–13, 2023. doi: 10.1109/TGRS.2023.3238185.
[34] D. B. Sam, S. V. Peri, M. N. Sundararaman, A. Kamath, and R. V. Babu, "Locate, size, and count: Accurately resolving people in dense crowds via detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 8, pp. 2739–2751, 2020.
[35] Z. Ma, X. Wei, X. Hong, and Y. Gong, "Bayesian loss for crowd count estimation with point supervision," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 6142–6151.
[36] W. Liu, M. Salzmann, and P. Fua, "Context-aware crowd counting," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 5099–5108.
[37] C. Wang *et al.*, "Uniformity in heterogeneity: Diving deep into count interval partition for crowd counting," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 3234–3242.
[38] Y. Chen, J. Yang, B. Chen, and S. Du, "Counting varying density crowds through density guided adaptive selection CNN and transformer estimation," *IEEE Trans. Circ. Syst. Video Technol.*, vol. 33, no. 3, pp. 1055–1068, 2022. doi: 10.1109/TCSVT.2022.3208714.
[39] Q. Song *et al.*, "Rethinking counting and localization in crowds: A purely point-based framework," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Montreal, QC, Canada, 2021, pp. 3365–3374.
[40] L. Y. Xiong, H. Deng, H. Yi, P. Huang, and Q. Zhou, "SFPANet: Separation and fusion pyramid attention network for crowd counting," *Multimed. Tools Appl.*, vol. 24, pp. 1–17, 2023. doi: 10.1007/s11042-023-17219-3.
[41] M. Dai, Z. Huang, J. Gao, H. Shan, and J. Zhang, "Cross-head supervision for crowd counting with noisy annotations," in *IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Rhodes Island, Greece, 2023, pp. 1–5.
[42] J. Yi, Y. Pang, W. Zhou, M. Zhao, and F. Zheng, "A perspective-embedded scale-selection network for crowd counting in public transportation," *IEEE Trans. Intell. Transp. Syst.*, 2023. doi: 10.1109/TITS.2023.3328000.
[43] M. Wang, H. Cai, Y. Dai, and M. Gong, "Dynamic mixture of counter network for location-agnostic crowd counting," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, Waikoloa, HI, USA, 2023, pp. 167–177.
[44] X. Cao, Z. Wang, Y. Zhao, and F. Su, "Scale aggregation network for accurate and efficient crowd counting," in *Proc. Eur. Conf. Comput. Vis.*, Munich, Germany, 2018, pp. 757–773.
[45] D. Liang, W. Xu, Y. Zhu, and Y. Zhou, "Focal inverse distance transform maps for crowd localization," *IEEE Trans. Multimed.*, vol. 25, pp. 6040–6052, 2022.