



JMFEEL-Net: a joint multi-scale feature enhancement and lightweight transformer network for crowd counting

Mingtao Wang¹ · Xin Zhou¹ · Yuanyuan Chen¹

Received: 24 July 2023 / Revised: 29 October 2023 / Accepted: 26 December 2023 /

Published online: 30 January 2024

© The Author(s), under exclusive licence to Springer-Verlag London Ltd., part of Springer Nature 2024

Abstract

Crowd counting based on convolutional neural networks (CNNs) has made significant progress in recent years. However, the limited receptive field of CNNs makes it challenging to capture global features for comprehensive contextual modeling, resulting in insufficient accuracy in count estimation. In comparison, vision transformer (ViT)-based counting networks have demonstrated remarkable performance by exploiting their powerful global contextual modeling capabilities. However, ViT models are associated with higher computational costs and training difficulty. In this paper, we propose a novel network named JMFEEL-Net, which utilizes joint multi-scale feature enhancement and lightweight transformer to improve crowd counting accuracy. Specifically, we use a high-resolution CNN as the backbone network to generate high-resolution feature maps. In the backend network, we propose a multi-scale feature enhancement module to address the problem of low recognition accuracy caused by multi-scale variations, especially when counting small-scale objects in dense scenes. Furthermore, we introduce an improved lightweight ViT encoder to effectively model complex global contexts. We also adopt a multi-density map supervision strategy to learn crowd distribution features from feature maps of different resolutions, thereby improving the quality and training efficiency of the density maps. To validate the effectiveness of the proposed method, we conduct extensive experiments on four challenging datasets, namely ShanghaiTech Part A/B, UCF-QNRF, and JHU-Crowd++, achieving very competitive counting performance.

Keywords Crowd counting · Count estimation · Multi-scale variations · Multi-density map supervision

✉ Yuanyuan Chen
chenyuanyuan@scu.edu.cn

Mingtao Wang
2021223040008@stu.scu.edu.cn

Xin Zhou
xinzhou@scu.edu.cn

¹ College of Computer Science, Sichuan University, Chengdu 610065, Sichuan Province, China

1 Introduction

Crowd counting aims to estimate the total number of people in an image using appropriate methods, and it holds significant practical value, in fields such as video surveillance [1] and city management [2]. In addition, it serves as the foundation for advanced tasks such as multi-class object counting [3], behavior analysis, and anomaly detection [4]. Therefore, in-depth research into this technology is highly necessary.

Crowd counting in natural scenes faces many challenges, including crowd occlusion, perspective distortion, and illumination variation. Mainstream methods typically adopt CNNs to regress crowd density maps and then obtain the total number of people by integrating and summing these predicted density maps. In recent years, researchers have proposed a variety of effective strategies to mitigate issues such as high density and multi-scale variation. One feasible workaround is to utilize multi-column CNNs [5, 6] with different receptive fields to aggregate multi-scale features of crowds. However, these methods lack an effective scale fusion mechanism, leading to feature redundancy between different branches, which makes it difficult to fully utilize the network's representation ability and generalization performance. To tackle the problem, some methods [7–9] employed a lightweight architecture. Specifically, these methods utilized VGG as the primary feature extractor and expanded the model's perceptual range by incorporating dilated convolution layers in the backend network. However, the density maps generated by these networks are only 1/8 the size of the original input. Low-resolution feature maps can result in the missing of much important information about the small objects, which limits their performance in certain complex scenes. Moreover, due to the single-column structure of VGG, it is difficult to achieve multi-level feature extraction and fusion, which poses challenges in effectively modeling high-complexity scenes. In order to further mitigate the interference of complex backgrounds on counting accuracy, researchers introduced the attention mechanisms [10, 11] into counting networks. Applying attention mechanisms in the networks helps enhance the ability to understand scenes, allowing them to better focus on local details and crowded regions. These methods mentioned above have been demonstrated to be useful. However, CNN usually only consider local regions, making it difficult to capture global features for context modeling, which is extremely ineffective especially in dense crowded areas. Another feasible strategy to improve the counting performance is to utilize the ViT [12], which relies on a powerful global modeling capability and is able to capture richer semantic information of crowds. In recent studies [13–16], researchers proposed ViT-based counting models and achieved better counting performance. However, it is undeniable that the computational cost of these methods is relatively high and the models are difficult to train.

Based on the above analysis, it is necessary to construct a counting model that can balance the performance and computational cost. For this purpose, this paper proposes a novel crowd counting method that leverages the dual advantages of CNN and ViT. This method is capable of better capturing the semantic information and contextual relationships within the crowd, thereby further improving the accuracy of dense crowd prediction. To achieve this goal, we use HRNet [17] as the front-end network, which maintains the size of its feature map output at 1/4 of the original input size, thereby generating rich high-resolution representations. This helps to preserve the richness of receptive field information, leading to more accurate density map prediction. We employ multi-scale feature enhancement and visual attention mechanisms to mitigate the effects of scale variation, severe occlusion, perspective distortion, and other factors on the counting results. We additionally adopt multi-density map supervision during training to accelerate model convergence.

The contributions of this paper can be summarized as follows:

- We propose a novel network named JMFEEL-Net, which incorporates the local perception capability of CNN and the global modeling capability of ViT, and thus demonstrates excellent performance in handling complex scenes.
- We propose a structurally simple and effective multi-scale feature enhancement module (MSFEM) that better models multi-scale features and mitigates the problem of low counting accuracy caused by scale variations.
- We design a multi-attention module (MAM) to enhance the model's ability to handle challenges in natural scenes. Additionally, we adopt a multi-density map supervision training strategy for network parameter optimization, which significantly improves the convergence speed and generalization performance of the model.

2 Related works

In recent years, deep learning techniques have advanced rapidly. Extensive research has been conducted on the problem of crowd counting. To improve the expressiveness of networks, researchers have employed strategies such as multi-scale feature fusion, attention mechanisms, and dilated convolutions to improve the feature extraction process. This section provides a brief review of some mainstream works on architecture design and feature extraction that are highly relevant to our proposed method.

2.1 Multi-scale feature fusion

Due to factors such as shooting angles and different scenes, the scale of objects in images exhibits non-uniform variations. Multi-scale feature fusion aims to address the scale variation problem by extracting the features of different scales using different receptive fields. In previous work, based on MCNN [5] and Switching-CNN [6], IG-CNN [18] automatically divided the density into different levels during training, and different patches selected the corresponding network branches. Similarly, SANet [19] used a feature map encoder and a density map estimator to extract multi-scale features and generate high-resolution density maps. In [8, 20], researchers used multi-scale feature fusion networks to extract contextual and semantic information from crowd scenes to reduce crowd feature loss. These architectures are still instructive and informative, and many subsequent methods have followed or extended this design idea.

2.2 Dilated convolution and deformable convolution

Dilated convolutions and deformable convolutions are two classical techniques used in image convolution operations. Dilated convolutions allow the capturing of higher-level features with larger receptive fields without increasing the number of parameters, computational complexity, or network complexity. CSRNet [7] used dilated convolutions to understand highly congested scenes and to perform accurate count estimation. DSSINet [21] adopted a multi-scale structural similarity loss function with dilation convolutions to guide the network in learning the local correlations of people in regions of different sizes, thereby generating high-quality density maps. Deformable convolutions can adaptively adjust the shape and size of receptive fields based on input features. DADNet [22] used deformable convolu-

tions to achieve precise spatial transformations of crowd positions in the generated density maps. ADCrowdNet [23] used deformable convolutions to improve the model's perception of detailed information, allowing it to capture crowd features more effectively.

2.3 Attention mechanism

Attention mechanisms have been widely used in various computer vision tasks to help models cope with complex scene problems. To deal with factors such as occlusion, illumination variations, and perspective distortion, some researchers have used attention mechanisms to enable models to focus on the crowd regions, further improving the accuracy and adaptability of the network. SCAR [10] introduced spatial and channel attention mechanisms into the crowd counting task. Among them, the spatial attention is used to encode pixel-level contextual information of the whole image, which improves the accuracy of the model in predicting pixel-level density maps. The channel attention is used to extract different feature information, making the model more robust to complex backgrounds. In other works [11, 24–26], researchers used attention mechanisms to filter out noisy information, reducing errors caused by background interference and further improving the network's generalization performance.

2.4 Vision transformer

Recently, the ViT has demonstrated remarkable performance across diverse visual domains, primarily attributed to its powerful capacity for capturing global context. Consequently, some researchers have begun to use ViT to improve crowd counting models. Liang et al. [13] proposed TransCrowd, a crowd counting network based on ViT, which achieved promising counting results in a weakly supervised manner. In other similar research works, MAN [14] used an improved transformer as an auxiliary feature extractor and incorporated global attention, learnable region attention, and instance attention loss into the network to improve the overall performance of the model. CCTrans [15] employed an efficient ViT as a backbone network and integrated a pyramid feature aggregation module to better cope with scale variation problem. The model achieved significant counting performance improvement in both fully and weakly supervised methods. In this paper, an improved lightweight ViT encoder is used in the backend network to model the global features of the crowd scene, which further reduces computational overhead and counting errors.

3 Proposed method

3.1 Overview

In this paper, the proposed JMFEEL-Net utilizes HRNet [17] as the backbone network. We construct the multi-attention module (MAM), a CNN branch, a transformer branch, and a regression decoder in the backend network. Among them, the CNN branch consists of an atrous convolution module (ACM) and a deformable convolution module (DCM), which mainly addresses the issue of multi-scale variations. The transformer branch contains an improved MobileViTBlock [27], which is utilized to capture semantic relationships within the entire region. An overview of the network is presented in Fig. 1. For each input image $I \in \mathbb{R}^{3 \times H \times W}$, we initially extract primary features using the backbone HRNet [17], resulting

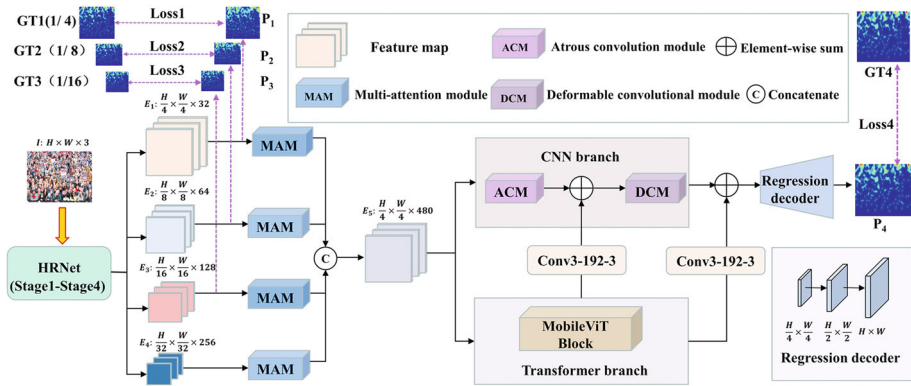


Fig. 1 The overview of our proposed JMFEEL-Net

in four primary high-resolution feature maps with different resolutions and channel numbers, namely $E_1 \in \mathbb{R}^{32 \times \frac{H}{4} \times \frac{W}{4}}$, $E_2 \in \mathbb{R}^{64 \times \frac{H}{8} \times \frac{W}{8}}$, $E_3 \in \mathbb{R}^{128 \times \frac{H}{16} \times \frac{W}{16}}$, and $E_4 \in \mathbb{R}^{256 \times \frac{H}{32} \times \frac{W}{32}}$. In order to enable the network to learn features at different stages, we use the three upper branches of HRNet to predict the three primary density maps separately; we define these as P_1 , P_2 , and P_3 , with their height and width being, respectively, 1/4, 1/8, and 1/16 of the original input size. Subsequently, these four primary feature maps are fed into the MAM and fused into a new attention feature map via channel concatenation, resulting in the feature $E_5 \in \mathbb{R}^{480 \times H \times W}$. The parallel CNN branch and transformer branch are then used for multi-scale feature enhancement and global contextual modeling. Next, we merge the outputs of these two branches using channel-wise concatenation. Finally, the fused feature map is fed into the decoder module for decoding, and predicting the final density map, defined as P_4 , which is the same as the original input size. The decoder module consists of two 4×4 transposed convolutional layers and one 1×1 convolutional layer.

3.2 Backbone

For crowded areas or small target crowds, high-resolution representations are crucial. Unlike many existing solutions, this paper uses HRNet [17] as a backbone to generate high-quality feature maps. Compared to VGG, HRNet performs better in feature extraction and maintaining high-resolution representations. It employs a strategy of repeatedly connecting and fusing multiple high- to low-resolution sub-networks in parallel, thus maintaining high-resolution features while being able to fully fuse multi-scale features. Considering the computational cost, we only use the lightweight HRNet V32 as the backbone network. The resolution of the output feature map is 1/4 of the input size, which makes the predicted feature map more spatially accurate.

3.3 Multi-attention module

In crowd counting tasks, applying attention mechanisms can help the network to distinguish between different crowd distributions and complex backgrounds. We utilize HRNet [17] to generate high-resolution representations and address the limitations of low-resolution feature

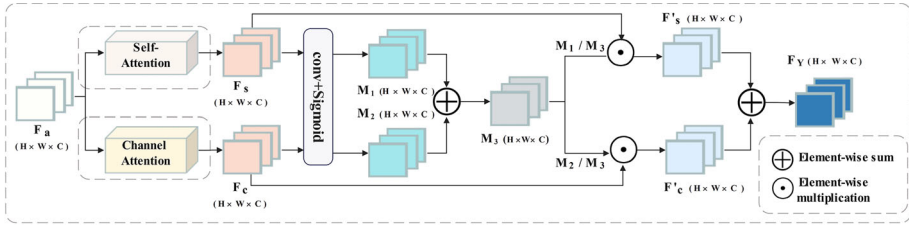


Fig. 2 The architecture of the proposed multi-attention module (MAM)

maps. However, during the repeated fusion process, the network introduces redundant noise information. Therefore, we perform the attention operation after HRNet.

We design a module called MAM that seamlessly integrates self-attention and channel-attention. This module is primarily used to balance the global information and local details of primary feature maps, thereby enabling the model to better understand complex crowd distributions while reducing feature redundancy. The structure of MAM is shown in Fig. 2. The input feature $F_a \in \mathbb{R}^{C \times H \times W}$ is first fed into the self-attention and channel attention submodules, generating the self-attention output $F_s \in \mathbb{R}^{C \times H \times W}$ and the channel attention output $F_c \in \mathbb{R}^{C \times H \times W}$, respectively. In order to allow the model to flexibly balance self-attention and channel attention during the learning process, we use a dynamic weight generation mechanism (a network consisting of convolutional layers and a sigmoid activation function) to compute weights for these two attentions, namely generating the weights $M_1 \in \mathbb{R}^{C \times H \times W}$ and $M_2 \in \mathbb{R}^{C \times H \times W}$. Subsequently, we add these two weights to obtain the total weight $M_3 \in \mathbb{R}^{C \times H \times W}$, which is used to normalize the two weights. We then multiply the pre-generated F_s and F_c with their normalized weights. Finally, we add them together to generate the fused attention feature map $F_Y \in \mathbb{R}^{C \times H \times W}$, defined as follows:

$$\begin{cases} M_i = \mathcal{F}_{sa}(F_i; \theta), \\ F_Y = \sum_{i=1}^N \left(F_i \odot \frac{M_i}{\sum_{i=1}^N M_i} \right), \end{cases} \quad (1)$$

where M_i is the attention weight, namely M_1 and M_2 . \mathcal{F}_{sa} is a network consisting of convolutional layers and a sigmoid function, F_i is the attention output feature map, namely F_s and F_c . The symbol \odot denotes the element-wise multiplication operation.

Self-attention The self-attention is an effective method for feature representation in neural networks. It primarily captures global contextual information by computing internal relationships within the input features. The process of the self-attention mechanism is shown in Fig. 3.

We replace fully connected layers with convolutional layers to achieve linear mapping. Specifically, during initialization, linear mappings for key, query, and value are performed by different convolutional layers with a kernel size of 1. These convolutional layers map the input features to a higher-dimensional subspace. Similarly, the final projection layer uses a convolutional layer with a kernel size of 1×1 . Compared to traditional self-attention operations, using convolutional layers with a kernel size of 1×1 helps to reduce the number of model parameters, lower computational complexity, and improve the efficiency of model optimization. Additionally, mapping the input features to a higher-dimensional subspace allows the model to learn different feature dependencies in multiple subspaces, which further enhances the feature representation capacity.

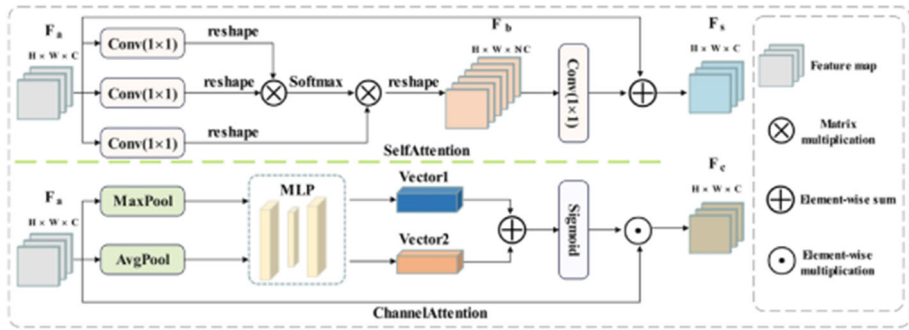


Fig. 3 Diagram of the self-attention and channel attention process

Channel attention We first apply global max pooling and global average pooling separately to the input feature maps. Next, we compress the feature maps based on two dimensions to obtain two different-dimensional feature descriptors. These two feature maps share a multilayer perceptron (MLP) network. In the MLP, the channel dimension is reduced by a fully connected layer and then restored by another fully connected layer. The two feature maps are then stacked along the channel dimension and the weights of each channel are normalized using the sigmoid activation function. Finally, the normalized weights are multiplied by the input feature maps to obtain the final weighted feature map.

3.4 Multi-density map supervision

We use multi-density map supervision (MMS) to guide the network in learning better models. By aggregating feature information from different layers and resolutions of the network, the model can adapt to different density levels in real scenarios and accelerate convergence. As shown in the dashed part of Fig. 1, we calculate the sum of the weighted losses between P_1 , P_2 , P_3 , and P_4 with their ground truth (GT) density maps to perform the MMS. The purpose of primary density map supervision is to enhance the robustness of intermediate feature maps and promote the accuracy of the final density regression. In view of the lower resolution (1/32) of the predicted primary feature maps in the fourth branch of HRNet, using lower-resolution density maps for supervision training is likely to increase prediction errors, especially in scenarios with smaller target crowds. Therefore, we do not use the primary density maps predicted by this branch for supervision training. Subsequent ablation experiments show that the MMS training-based strategy helps to fully exploit the correlations between density maps of different resolutions, thereby facilitating learning of the crowd distribution in the scene, producing finer density maps, and further improving training efficiency.

3.5 CNN branch

The output feature maps of HRNet [17] are only 1/4 the size of the original input image, making it challenging to predict dense crowds or smaller targets. Therefore, it is necessary to further increase the resolution of the feature map. Inspired by previous works such as CSRNet [7], ADCrowdNet [23], DCN [28], and DADNet [22], we construct a multi-scale feature enhancement module (MSFEM) in the CNN branch. This module consists of two

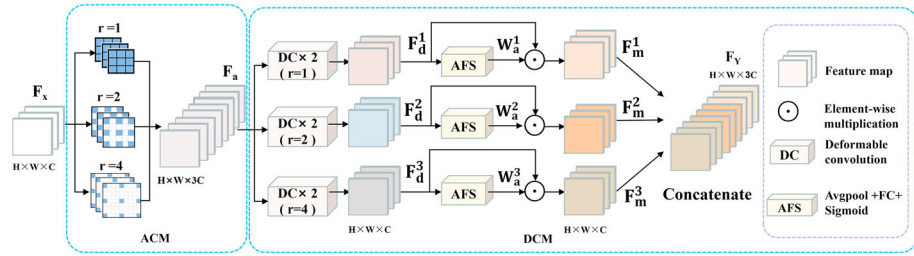


Fig. 4 The architecture of the proposed multi-scale feature enhancement module (MSFEM)

submodules: the atrous convolution module (ACM) and the deformable convolution module (DCM).

Atrous Convolution Module The architecture settings of the ACM submodule are shown in Fig. 4. This submodule comprises three parallel branches, each containing convolutional layers with dilation rates of 1, 2, and 4, respectively. The purpose is to further expand the receptive field and integrate features of different scales. This design strategy effectively compensates for the potential loss of feature detail during pooling and sampling operations, enabling the network to better recognize small target clusters, edges, and other local features. It also facilitates the modeling of multi-scale features, resulting in more accurate counting results.

Deformable Convolution Module Compared to traditional convolutions, deformable convolutions exhibit enhanced representational capabilities. They can learn richer spatial distribution features of crowds, which improves the model’s performance in counting tasks. In this submodule, the parameter settings are similar to those in the ACM. We use a three-branch deformable convolutional group design. These three sub-branches process the input feature maps in parallel, allowing the model to capture more diverse geometric shapes and structural information at different feature levels. We set the convolution kernel size for all three sets of deformable convolutional layers to 3×3 . Compared to larger kernel sizes, the 3×3 convolution kernel has a clear advantage in terms of parameter efficiency while still effectively capturing local features. In addition, we apply attention after the last deformable convolutional layer in each branch to improve counting accuracy. Specifically, the DCM takes the feature $F_a \in \mathbb{R}^{3C \times H \times W}$ as input and extracts the features $F_d^i \in \mathbb{R}^{C \times H \times W}$ ($1 \leq i \leq 3$) layer by layer through three sets of deformable convolutional layers at different scales. These features are then fed into the AFS network to generate the attention weights W_a^i . The AFS network consists of an average pooling layer, a fully connected layer, and a sigmoid activation function. Next, the feature $F_m^i \in \mathbb{R}^{C \times H \times W}$ is obtained by performing a multiplication operation between the pre-generated features F_d^i and the attention weights W_a^i , which is defined as follows:

$$F_m^i = F_d^i \odot W_a^i, \tag{2}$$

where F_d^i are the pre-generated features for each branch, W_a^i are the corresponding attention weights. The symbol \odot denotes the element-wise multiplication operation.

Finally, the multi-scale features $F_Y \in \mathbb{R}^{C \times H \times W}$ are aggregated:

$$F_Y = F_m^1 \oplus F_m^2 \oplus F_m^3, \tag{3}$$

where F_m^i are the features obtained from each branch. The symbol \oplus denotes the channel concatenation operation.

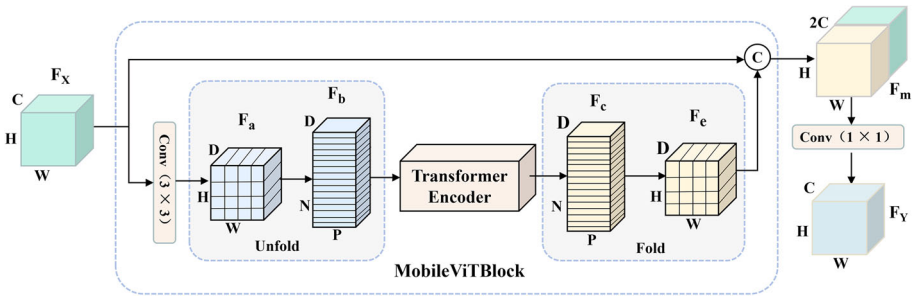


Fig. 5 The process diagram of the MobileViTBlock

3.6 Transformer branch

In the transformer branch, we use a lightweight ViT encoder called MobileViTBlock [27] to capture global context. The MobileViTBlock is suitable for deployment and execution on resource-constrained devices. It is smaller than conventional ViT models while still providing remarkable global modeling capabilities.

As shown in Fig. 5, the workflow of this module is as follows: firstly, for the features $F_X \in \mathbb{R}^{C \times H \times W}$ extracted from MAM, MobileViT employs a shallow convolutional network for local modeling and dimension transformation, generating features $F_a \in \mathbb{R}^{D \times H \times W}$. In order to learn a global representation with spatial inductive bias, the feature F_a is divided into patches and unfolded into N non-overlapping features $F_b \in \mathbb{R}^{D \times N \times P}$. Subsequently, the encoder from the ViT model is used for feature interaction and encoding. The encoder consists of a multi-head attention (MA) module, layer normalization (LN), and a feed-forward network (FFN), defined as follows:

$$\begin{cases} F'_x = LN(FFN(LN(F_x)) + LN(F_x)), \\ F_x = MA(I_x) + I_x, \\ I_x = UF(x), \\ MA(I_x) = softmax\left(\frac{QK^T}{\sqrt{c}}\right)V, \end{cases} \quad (4)$$

where F'_x is the final output of the global contextual features from the encoder, $UF(x)$ represents the operation of unfolding the feature F_a . K , Q , and V are, respectively, the key, query, and value in the multi-head attention operation, and they are three learnable weight matrices. $\frac{1}{\sqrt{c}}$ is a scaling factor.

After the aforementioned process, resulting in intermediate feature $F_c \in \mathbb{R}^{D \times N \times P}$, which is then folded into features $F_e \in \mathbb{R}^{D \times H \times W}$. Finally, the residual F_X is concatenated, and a 1×1 convolutional layer is used to project the features into a lower-dimensional space, obtaining the fused global contextual feature map $F_Y \in \mathbb{R}^{C \times H \times W}$.

The multi-head attention operation incurs high computational cost. The challenge lies in performing matrix multiplication operations with high time and space complexity based on the context. We align the output channels of the 3×3 convolutional layer in the local modeling stage with the embedding dimension (embed-dim) of the transformer encoder and remove the two 1×1 convolutional layers between them, further enhancing the module's lightweightness

and efficiency. To further reduce computational cost, only one MobileViTBlock is used in this branch, and the number of attention heads is set to 4.

3.7 Loss function

We utilize the Mean Squared Error (MSE) as the optimization objective function, which measures the difference between the predicted density maps and the ground truth density maps. Its definition is as follows:

$$\mathcal{L}_m = \frac{1}{2N} \sum_{i=1}^N \|Den_i^{EST} - Den_i^{GT}\|_2^2, \quad (5)$$

where N is the number of training images, Den_i^{GT} and Den_i^{EST} denote the GT and predicted density map, respectively.

As discussed above, during training, we use multi-density maps to supervise the model training. The density maps P1, P2, P3, and P4 have different resolutions and semantic information, which leads to the fact that they need to be optimized to different degrees. Therefore, we balance the weight of each loss term by a hyperparameter λ_m . The final training loss function is defined as follows:

$$\mathcal{L}_{total} = \sum_{m=1}^4 \lambda_m \mathcal{L}_m, \quad (6)$$

where \mathcal{L}_m ($m=1,2,3,4$) denotes the MSE loss between the predicted density maps (P1, P2, P3, and P4) and GT density maps, λ_m is the weight assigned to each loss term. We set the weights $\lambda_1, \dots, \lambda_4$ to 0.1, 0.15, 0.2, and 0.3, respectively.

4 Experiments

4.1 Datasets

We evaluate the performance of the proposed method on four public datasets, the characteristics of which are outlined below.

ShanghaiTech [5] consists of two parts, Part A and Part B, with a total of 1,198 images. Part A contains 300 training images and 182 testing images, while Part B contains 400 training images and 316 testing images. Part A has a higher crowd density and is collected from web images, while Part B has uniformly sized images with relatively lower crowd density and is collected from street-view images.

UCF-QNRF [29] is a challenging dataset. This dataset contains a total of 1,535 images with varying sizes, and the number of people ranges from 49 to 12,865. Both crowd density and image resolution show significant variation in this dataset.

JHU-Crowd++ [30] contains 4,822 images, divided into 2,722 training images, 500 validation images, and 1,600 testing images. The number of people in each image varies greatly, ranging from 0 to 25,791 individuals. Additionally, this dataset includes images captured under extreme weather conditions such as snow, rain, and haze.

4.2 Evaluation metrics

We use two commonly used evaluation metrics, Mean Absolute Error (MAE) and Mean Squared Error (MSE), to validate the accuracy and robustness of the proposed JMFEEL-Net. These two metrics can be defined as follows:

$$MAE = \frac{1}{m} \sum_{i=1}^m |y_i - \bar{y}_i|, \quad (7)$$

$$MSE = \sqrt{\frac{1}{m} \sum_{i=1}^m (y_i - \bar{y}_i)^2}, \quad (8)$$

where m is the total number of images, y_i and \bar{y}_i are the GT and predicted count of the i -th image, respectively.

4.3 Implementation details

In this work, we generate GT density maps using a fixed Gaussian kernel of size 15. During the training process, we use the Adam optimizer. Initially, we train the model for 300 epochs with an initial learning rate of $1e-4$ and a weight decay rate of $5 \times 1e-4$. We then continue training for another 200 epochs with a learning rate of $1e-5$. We also use random cropping and horizontal flipping to augment the training data. The cropping size for the Part A dataset is 256×256 , while that for the other datasets is 512×512 .

4.4 Results and analysis

4.4.1 Counting results

We compare our method with 17 other classical methods, including CSRNet [7], FIDTM [31], CLTR [32] and CHS-Net [33]. The comparative results are shown in Table 1. Overall, our method achieves excellent counting performance on the ShanghaiTech, UCF-QNRF, and JHU-Crowd++ datasets, significantly outperforming the majority of methods. Particularly, the proposed method obtains the best counting results on the ShanghaiTech Part A dataset. Compared to methods based on other backbones, there is a significant reduction in count error. For the Shanghai Tech PartB dataset, the counting performance of our proposed JMFEEL-Net is comparable to CLTR [32] and SGANet [34], obtaining the third best counting performance. Compared with DFRNet [35], MAE and MSE are improved by 2.9 % and 9.9%, respectively. For the UCF-QNRF dataset, our method achieves substantial reductions in both MAE and MSE. In terms of MAE, JMFEEL-Net is tied for second place with CHS-Net [33]. Additionally, for the large-scale JHU dataset, our method is highly competitive, showing significant improvements in both MAE and MSE, ranking third and second, respectively.

We further adopt visual density maps to illustrate the counting performance of JMFEEL-Net under different density scenarios and compare it with density maps predicted by other methods. Figure 6 presents examples of density maps predicted by our method on datasets with varying scales. It can be seen that our method shows excellent adaptability in coping with changes in crowd density. This strongly demonstrates our method can efficiently extract rich contextual feature at multiple scales. Figure 7 displays the results of comparing the density maps predicted by different methods. As can be seen, when dealing with crowd

Table 1 Comparison of different methods on the ShanghaiTech, UCF-QNRF, and JHU-Crowd++ datasets

Method	Venue	SHA		SHB		UCF-QNRF		JHU-Crowd++	
		MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE
CSRNet [7]	2018 CVPR	68.2	115	10.6	16	–	–	85.9	309.2
CAN [8]	2019 CVPR	62.3	100.0	7.8	12.2	107	183	100.1	314.0
PaDNet [36]	2019 TIP	59.2	98.1	8.1	12.2	96.5	170.2	–	–
AMRNet [37]	2020 ECCV	61.6	98.4	7.0	11.0	86.6	152.2	–	–
MSPNet [38]	2020 ICASSP	59.8	98.2	7.5	14.1	–	–	–	–
NoisyCC [39]	2020 NIPS	61.9	99.6	7.4	11.3	85.8	150.6	67.7	258.5
DS-CNN [40]	2021 AJSE	–	–	–	–	115.2	175.7	–	–
AutoScale [41]	2021 IJCV	64.2	98.4	7.2	11.1	104.7	173.6	–	–
Khan et al. [42]	2021 TVC	–	–	–	–	112	173	–	–
GL [43]	2021 CVPR	61.3	95.4	7.3	11.7	84.3	147.5	59.9	259.5
Khan et al. [44]	2021 IJGIS	77.58	129.7	14.1	21.10	218.2	357.4	–	–
SGANet [34]	2022 TITS	57.6	101.1	6.6	10.2	87.6	152.5	–	–
FIDTM [31]	2022 TMM	57.0	103.4	6.9	11.8	89.0	153.5	66.6	253.6
CLTR [32]	2022 ECCV	56.9	95.2	6.5	10.6	87.3	142.4	59.5	240.6
Meng et al. [45]	2022 arXiv	57.0	98.6	7.1	12.3	85.3	129.4	66.6	254.9
DFRNet [35]	2023 ACM	59.6	100.9	6.9	12.1	80.2	145.5	–	–
CHS-Net [33]	2023 ICASSP	59.2	97.8	7.1	12.1	83.4	144.9	–	–
JMFEEL-Net(ours)	–	56.1	94.5	6.7	10.9	83.4	147.8	60.4	241.6

Best results are highlighted in bold

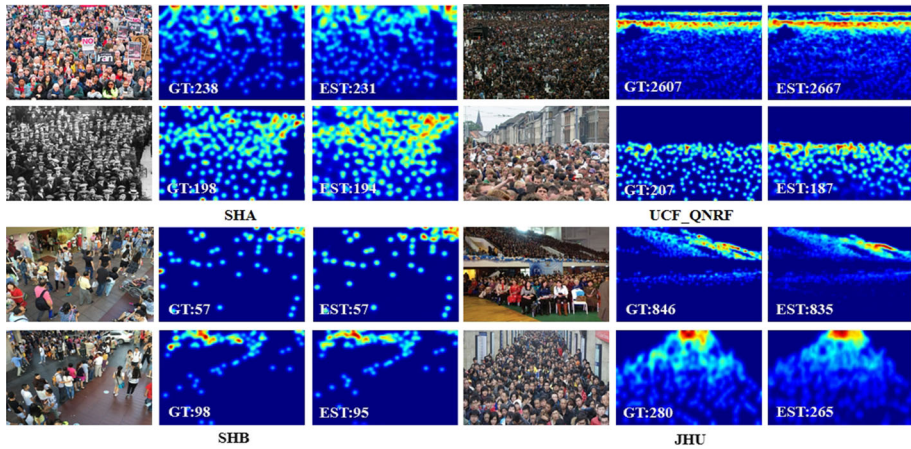


Fig. 6 The visualization results of density maps on the ShanghaiTech, UCF-QNRF, and JHU-Crowd++ datasets

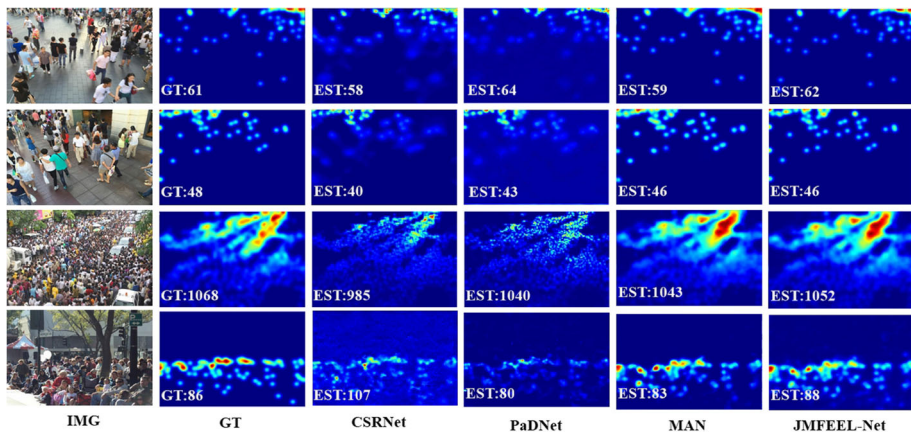


Fig. 7 The visualization results of density maps generated by different methods

scenes with complex backgrounds, CSRNet [7] and PaDnet [36] cannot recognize crowds and backgrounds effectively, especially in crowded areas. In comparison, the quality of the density map predicted by our method is excellent and closer to GT. These results demonstrate that our method exhibits better accuracy and greater robustness in complex crowd scenarios. This can be largely attributed to our adoption of the MSFEM and MAM, which help to address the multi-scale variation problem and capture global features, thus facilitating the prediction of more detailed and accurate density maps. However, as shown in Fig. 8, in specific scenarios, the density maps generated by JMFEEL-Net have lower quality, resulting in a considerable difference in the counting results as compared to the GT. The reason may be that the relatively low resolution and excessively high crowd density in some images, which result in the loss of some detailed information during the sampling process, making it challenging to obtain sufficient features for accurate estimation.

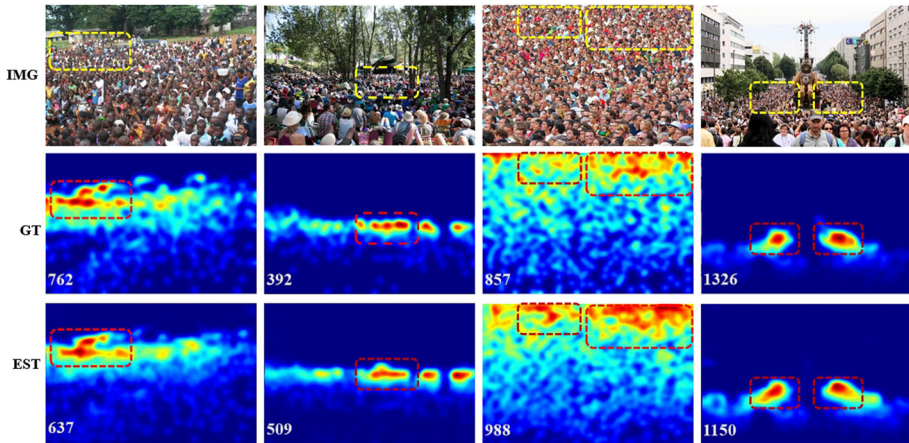


Fig. 8 The visualization results of density maps with large prediction errors

Table 2 The comparisons of complexity

Method	Backbone	Parameters(M)	FLOPs(G)
CSRNet [7]	VGG	16.3	54.2
AutoScale [41]	VGG	24.9	53.7
TransCrowd [13]	Transformer	86.4	49.3
GL [43]	VGG	21.5	52.3
FIDTM [31]	HRNet V48	66.6	80.1
MAN [14]	VGG	30.9	58.2
CLTR [32]	ResNet50	43.4	78.6
JMFEEL-Net(ours)	HRNet V32	33.8	62.4

The experiments are conducted on a 3090 GPU, and the size of the input image is 512×512

4.4.2 Complexity analysis

Table 2 presents the results of a computational cost comparison with other methods. Our proposed JMFEEL-Net does not have an advantage in model parameters and FLOPs, primarily because we utilize HRNet as the backbone, which increases the network's complexity to some extent compared to VGG. However, our method still has a lower computational cost than methods such as FIDTM [31], TransCrowd [13] and CLTR [32], while delivering superior counting results. Our next objective is to streamline JMFEEL-Net using appropriate pruning algorithms, further enhancing the model's efficiency.

4.5 Ablation study

To validate the effectiveness of the MAM, MSFEM, MobileViTBlock, and MMS, we conduct multiple ablation experiments on the ShanghaiTech Part A dataset. Firstly, we evaluate the effect of MAM on the counting results. Then, we predict the results using CNN branch and Transformer branch separately, and further evaluate the prediction results combining the two branches. Next, we use two different strategies for supervised training of the model and further

Table 3 Ablation experiments of the key components on the ShanghaiTech Part A dataset

HRNet	MAM	MSFEM	MobileViTBlock	MMS	MAE	MSE
✓					69.1	113.2
✓	✓				66.5	110.8
✓		✓			65.6	106.7
✓			✓		65.1	102.3
✓		✓	✓		62.4	101.6
✓	✓	✓	✓		58.4	98.3
✓	✓	✓	✓	✓	56.1	94.5

Best results are highlighted in bold

explore the effect of λ_m on the counting results. The quantitative results of counting accuracy are listed in Table 3. Finally, to provide a more intuitive demonstration of the effectiveness of different components of the model, we display density maps for various combinations of outputs, as shown in Fig. 9

Effectiveness of MAM To explore the effectiveness of MAM, we introduce MAM after four branches of the backbone. The experimental results are shown in Table 3. We can observe a significant improvement in counting error, specifically, the MAE and MSE are reduced by 3.8% and 2.1%, respectively. By comparing the predicted density maps, it can be seen that the density maps predicted by backbone only are rather blurred. In particular, the dense regions are less distinguishable between the background and crowd regions, while the quality of the density maps generated by adding MAM is much improved and the local and global textures of the images are clearer. This shows that in our work, MAM can better help the network to better recognize the distribution of crowd and background regions, and can effectively solve the semantic imbalance problem.

Effectiveness of MSFEM We incorporate a MSFEM after the backbone to compensate for the details lost during the sampling operations and verify its positive impact on the counting results. As shown in Table 3, both MAE and MSE are significantly reduced and the quality of the density map is greatly improved. This proves that the module is able to enhance the representation ability of the network by expanding the receptive field and effectively mitigate the multi-scale variation problem. In addition, Table 4 shows that the network achieves optimal results when the dilation rates are set to 1, 2, and 4, respectively. Considering that the computational complexity becomes excessive when the number of parallel modules or the number of layers in the dilated convolution is set to 4 or higher, we do not add any additional modules beyond this point.

Effectiveness of MobileViTBlock Similarly, we introduce the lightweight MobileViTBlock after the backbone and evaluate its effect on the counting results. The results are shown in Table 3, the MAE and MSE are improved by 5.8% and 9.6%, respectively. As can be seen from Fig. 9, the density map generated by MobileViT is closer to GT, especially the prediction results in crowded areas are better than those predicted by MSFEM, which indicates that MobileViTBlock is more capable of modeling dense areas and can learn richer semantic information. In order to prove that the joint learning of MSFEM and MobileViT is beneficial to improve the global and local recognition performance of the network, we use both CNN branch and transformer branch to predict the density map and compare the quality of their density maps. As shown in Fig. 9, the density maps output from the joint MSFEM and MobileViT have higher density quality, and show great clarity and accuracy in both congested

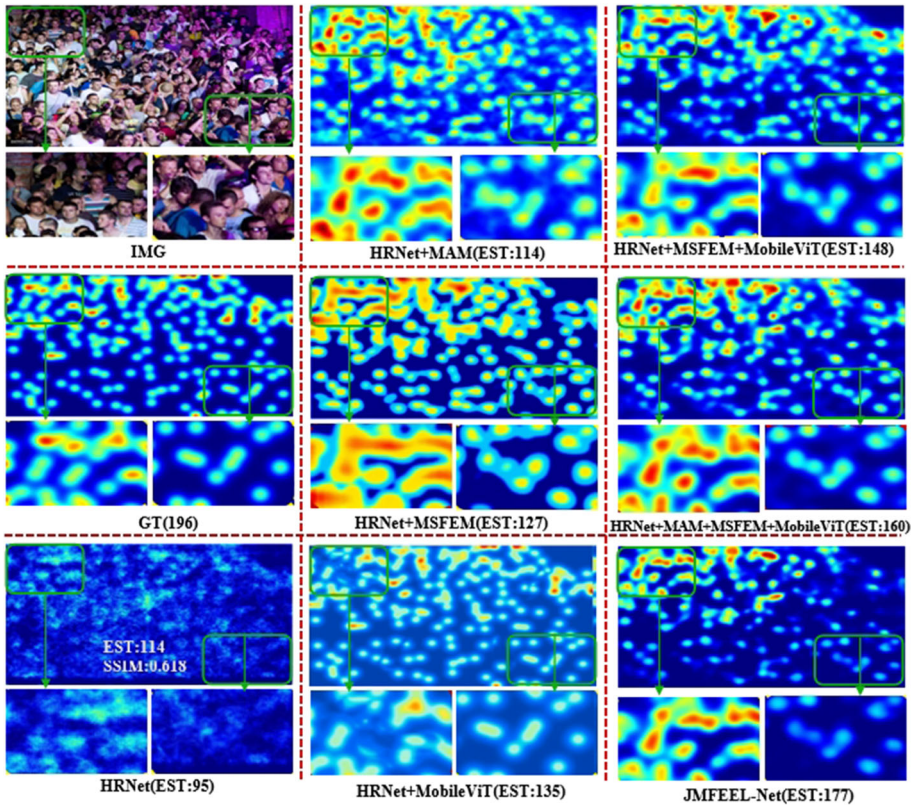


Fig. 9 The visualization results of predicted density maps for different components

and sparse regions, which proves that our method can cope with multi-scale features as well as efficiently model global features.

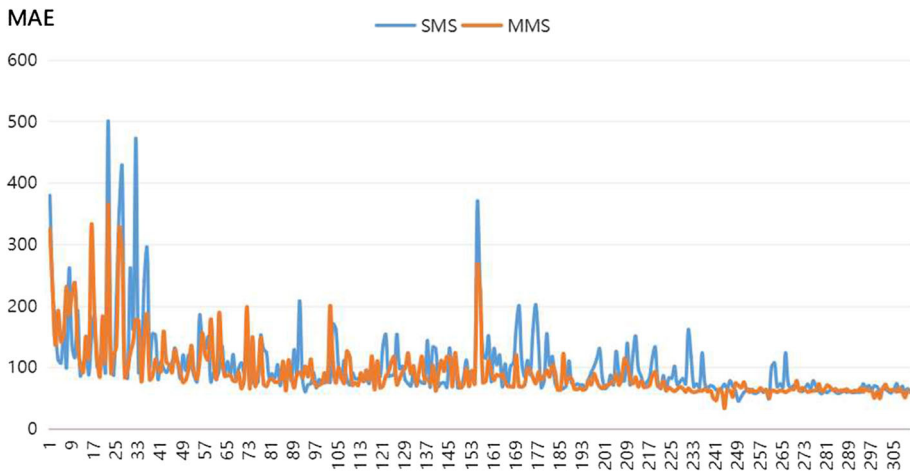
Effectiveness of MMS In order to verify the advantages of the multi-density map supervised training strategy, we compare the single-density map supervised training and multi-density map supervised training. The experimental results are shown in Table 3, Figs. 9 and 10. By jointly supervising P1, P2, P3 and P4, the MAE and MSE on SHA are improved by 3.9% and 3.8%, respectively. As depicted in Fig. 10, the MMS strategy for training exhibits greater stability and converges more quickly. This indicates that the MMS significantly outperforms the results of single-density map supervision, which can enhance the semantic perception of the network, fully utilize the correlation between feature maps of different resolutions to further reduce the loss, and promote the convergence speed of the model.

Effect of different λ_m In training, we balance the global correlation and consistency of different density maps by λ_m . We further evaluate the effect of different weights λ_m on the counting results. The detailed operation is as follows: we first keep the weights of the four losses the same and set them as the baseline, which also means optimizing the predicted four density maps equally. However, the final counting results are not satisfactory. In addition, as shown in Table 5, the counting accuracy of JMFEEL-Net declines obviously with the gradual increase of λ_m . We then fine-tune the value of λ_m and gradually reduce the interval between the four weights. It can be observed that the network obtains the lowest MAE and MSE when

Table 4 The impact of different dilation rates of the MSFEM on the counting results

Module	Dilation rate	MAE	MSE
MSFEM	r = 1, 2, 3	59.6	102.4
	r = 2, 3, 4	63.2	104.5
	r = 1, 2, 4	56.1	94.5
	r = 1, 3, 4	58.8	100.2

Best results are highlighted in bold

**Fig. 10** Comparison results of different supervised training methods on SHA. SMS and MMS denote single-density map supervised training and multi-density map supervised training**Table 5** The effect of changing λ_m on the counting results

Module	λ_1 - λ_4	MAE	MSE
MMS	0.1, 0.1, 0.1, 0.3	57.8	100.2
	0.1, 0.1, 0.1, 0.5	59.3	102.3
	0.1, 0.2, 0.3, 0.5	57.4	98.5
	0.1, 0.15, 0.2, 0.3	56.1	94.5
	0.1, 0.15, 0.2, 0.35	56.8	96.8
	1, 1, 1, 1	58.1	105.4

Best results are highlighted in bold

the weights of the four loss terms are set to 0.1, 0.15, 0.20, and 0.3. Therefore, we choose these four weight values to construct our proposed JMFEEL-Net.

5 Conclusion

In this paper, we propose a novel method, which effectively enhances the accuracy of dense crowd prediction by combining the strengths of CNN and ViT. We construct a multi-scale feature enhancement module in the CNN branch to supplement the lost detailed information during pooling and convolution operations, effectively addressing the multi-scale

problems. We integrate multiple attention mechanisms into the model to adaptively select global semantic information and local detailed information. We also introduce multi-density map supervision, effectively combining density maps from different stages and resolutions, learning the correlation between high-resolution and low-resolution features, reducing counting errors, and accelerating the model convergence. Experimental results demonstrate that our method achieves promising counting accuracy on four classic datasets. Our ablation experiments demonstrate the effectiveness of the proposed MSFEM, MobileViT, MAM, and MMS. In future work, we will explore ways to determine the weights of individual loss terms through learning automatic weighting, thereby further improving the counting performance of the model.

Author Contributions MW: Conceptualization, Methodology, Writing- Original draft preparation. XZ: Supervision, Validation, YC: Conceptualization, Supervision, Validation, Writing- Reviewing and Editing.

Declarations

Conflict of interest The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

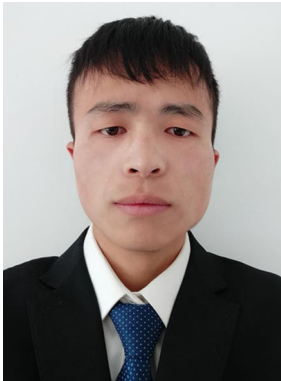
1. Chan AB, Liang Z-SJ, Vasconcelos N (2008) Privacy preserving crowd monitoring: counting people without people models or tracking. In: 2008 IEEE conference on computer vision and pattern recognition. IEEE, pp 1–7
2. Sindagi VA, Patel VM (2018) A survey of recent advances in cnn-based single image crowd counting and density estimation. *Pattern Recogn Lett* 107:3–16
3. Liu Z, Wang Q, Meng F (2022) A benchmark for multi-class object counting and size estimation using deep convolutional neural networks. *Eng Appl Artif Intell* 116:105449
4. Ko T (2008) A survey on behavior analysis in video surveillance for homeland security applications. In: 2008 37th IEEE applied imagery pattern recognition workshop. IEEE, pp 1–8
5. Zhang Y, Zhou D, Chen S, Gao S, Ma Y (2016) Single-image crowd counting via multi-column convolutional neural network. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 589–597
6. Babu Sam D, Surya S, Venkatesh Babu R (2017) Switching convolutional neural network for crowd counting. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 5744–5752
7. Li Y, Zhang X, Chen D (2018) CSRNet: dilated convolutional neural networks for understanding the highly congested scenes. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1091–1100
8. Liu W, Salzmann M, Fua P (2019) Context-aware crowd counting. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 5099–5108
9. Basalamah S, Khan SD, Ullah H (2019) Scale driven convolutional neural network model for people counting and localization in crowd scenes. *IEEE Access* 7:71576–71584
10. Gao J, Wang Q, Yuan Y (2019) Scar: spatial-/channel-wise attention regression networks for crowd counting. *Neurocomputing* 363:1–8
11. Jiang X, Zhang L, Xu M, Zhang T, Lv P, Zhou B, Yang X, Pang Y (2020) Attention scaling for crowd counting. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 4706–4715
12. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, et al (2020) An image is worth 16x16 words: Transformers for image recognition at scale. [arXiv:2010.11929](https://arxiv.org/abs/2010.11929)
13. Liang D, Chen X, Xu W, Zhou Y, Bai X (2022) Transcrowd: weakly-supervised crowd counting with transformers. *Sci China Inf Sci* 65(6):160104
14. Lin H, Ma Z, Ji R, Wang Y, Hong X (2022) Boosting crowd counting via multifaceted attention. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 19628–19637

15. Tian Y, Chu X, Wang H (2021) CCTrans: simplifying and improving crowd counting with transformer. [arXiv:2109.14483](https://arxiv.org/abs/2109.14483)
16. Qian Y, Zhang L, Hong X, Donovan C, Arandjelovic O, Fife U, Harbin P (2022) Segmentation assisted u-shaped multi-scale transformer for crowd counting. In: 2022 British machine vision conference. The British Machine Vision Association (BMVA)
17. Wang J, Sun K, Cheng T, Jiang B, Deng C, Zhao Y, Liu D, Mu Y, Tan M, Wang X et al (2020) Deep high-resolution representation learning for visual recognition. *IEEE Trans Pattern Anal Mach Intell* 43(10):3349–3364
18. Sam DB, Sajjan NN, Babu RV, Srinivasan M (2018) Divide and grow: capturing huge diversity in crowd images with incrementally growing CNN. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3618–3626
19. Cao X, Wang Z, Zhao Y, Su F (2018) Scale aggregation network for accurate and efficient crowd counting. In: Proceedings of the European conference on computer vision (ECCV), pp 734–750
20. Sindagi VA, Patel VM (2017) Generating high-quality crowd density maps using contextual pyramid CNNs. In: Proceedings of the IEEE international conference on computer vision, pp 1861–1870
21. Liu L, Qiu Z, Li G, Liu S, Ouyang W, Lin L (2019) Crowd counting with deep structured scale integration network. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 1774–1783
22. Guo D, Li K, Zha Z-J, Wang M (2019) DADNet: dilated-attention-deformable convnet for crowd counting. In: Proceedings of the 27th ACM international conference on multimedia, pp 1823–1832
23. Liu N, Long Y, Zou C, Niu Q, Pan L, Wu H (2019) ADCrowdNet: an attention-injective deformable convolutional network for crowd understanding. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 3225–3234
24. Zou Z, Cheng Y, Qu X, Ji S, Guo X, Zhou P (2019) Attend to count: crowd counting with adaptive capacity multi-scale CNNs. *Neurocomputing* 367:75–83
25. Zhang A, Shen J, Xiao Z, Zhu F, Zhen X, Cao X, Shao L (2019) Relational attention network for crowd counting. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 6788–6797
26. Xie J, Pang C, Zheng Y, Li L, Lyu C, Lyu L, Liu H (2022) Multi-scale attention recalibration network for crowd counting. *Appl Soft Comput* 117:108457
27. Mehta S, Rastegari M (2021) MobileViT: light-weight, general-purpose, and mobile-friendly vision transformer. [arXiv:2110.02178](https://arxiv.org/abs/2110.02178)
28. Dai J, Qi H, Xiong Y, Li Y, Zhang G, Hu H, Wei Y (2017) Deformable convolutional networks. In: Proceedings of the IEEE international conference on computer vision, pp 764–773
29. Idrees H, Tayyab M, Athrey K, Zhang D, Al-Maadeed S, Rajpoot N, Shah M (2018) Composition loss for counting, density map estimation and localization in dense crowds. In: Proceedings of the European conference on computer vision (ECCV), pp 532–546
30. Sindagi VA, Yasarla R, Patel VM (2020) JHU-Crowd++: large-scale crowd counting dataset and a benchmark method. *IEEE Trans Pattern Anal Mach Intell* 44(5):2594–2609
31. Liang D, Xu W, Zhu Y, Zhou Y (2022) Focal inverse distance transform maps for crowd localization. *IEEE Transactions on Multimedia*
32. Liang D, Xu W, Bai X (2022) An end-to-end transformer model for crowd localization. In: European conference on computer vision. Springer, pp 38–54
33. Dai M, Huang Z, Gao J, Shan H, Zhang J (2023) Cross-head supervision for crowd counting with noisy annotations. In: ICASSP 2023–2023 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, pp 1–5
34. Wang Q, Breckon TP (2022) Crowd counting via segmentation guided attention networks and curriculum loss. *IEEE Trans Intell Transp Syst* 23(9):15233–15243
35. Gao X, Xie J, Chen Z, Liu A-A, Sun Z, Lyu L (2023) Dilated convolution-based feature refinement network for crowd localization. *ACM Trans Multimed Comput Commun Appl* 19(6):1–16
36. Tian Y, Lei Y, Zhang J, Wang JZ (2019) Padnet: pan-density crowd counting. *IEEE Trans Image Process* 29:2714–2727
37. Liu X, Yang J, Ding W, Wang T, Wang Z, Xiong J (2020) Adaptive mixture regression network with local counting map for crowd counting. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIV 16. Springer, pp 241–257
38. Wei B, Yuan Y, Wang Q (2020) MSPNet: multi-supervised parallel network for crowd counting. In: ICASSP 2020–2020 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, pp 2418–2422
39. Wan J, Chan A (2020) Modeling noisy annotations for crowd counting. *Adv Neural Inf Process Syst* 33:3386–3396
40. Khan SD, Basalamah S (2021) Sparse to dense scale prediction for crowd counting in high density crowds. *Arab J Sci Eng* 46(4):3051–3065

41. Xu C, Liang D, Xu Y, Bai S, Zhan W, Bai X, Tomizuka M (2022) AutoScale: learning to scale for crowd counting. *Int J Comput Vision* 130(2):405–434
42. Khan SD, Basalamah S (2021) Scale and density invariant head detection deep model for crowd counting in pedestrian crowds. *Vis Comput* 37(8):2127–2137
43. Wan J, Liu Z, Chan AB (2021) A generalized loss function for crowd counting and localization. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp 1974–1983
44. Khan SD, Salih Y, Zafar B, Noorwali A (2021) A deep-fusion network for crowd counting in high-density crowded scenes. *Int J Comput Intell Syst* 14(1):168
45. Meng Y, Bridge J, Wei M, Zhao Y, Qiao Y, Yang X, Huang X, Zheng Y (2022) Counting with adaptive auxiliary learning. [arXiv:2203.04061](https://arxiv.org/abs/2203.04061)

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

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.



Mingtao Wang is currently pursuing a master's degree at the College of Computer Science, Sichuan University. His research interests include computer vision and crowd density estimation.



Xin Zhou received his Ph.D. from the College of Automation, Chongqing University, China, in 2003. He is currently an associate professor in the College of Computer Science, Sichuan University. He specializes in image processing and pattern recognition, computer vision technology, and intelligent transportation systems.



Yuanyuan Chen received her BA in pattern recognition and intelligent systems from Shanghai Jiaotong University in 2007 and then became a lecturer for teaching and research in the Department of Computer Science at Sichuan University. She received her Ph.D. in machine intelligence from Sichuan University in 2015. Currently, she is an associate professor and master supervisor of Sichuan University. She has always been devoted to researching the theory and applications of neural networks and medical image analysis.