Lightweight Pyramid Network-Based Multi-Modal Feature Fusion For Crowd Counting

Han Shi, *Rongkai Wang, and *Deqing Li

*Corresponding author. E-mail: wangrk2024@126.com (Rongkai Wang); byoungholee@qq.com (Deqing Li)

Received May 31, 2026; accepted August 17, 2026

Accurate crowd counting under complex scenes remains challenging due to scale variation, occlusion, and background noise. This paper introduces a lightweight pyramid network that fuses multi-modal features for robust crowd density estimation. The architecture integrates VGG-16, a multi-scale feature fusion module (MFFM) and a convolutional block attention module (CBAM) to suppress noise and enhance head-region focus. A dilated convolution module (DCM) further refines features into background attention and density maps, which are element-wise multiplied to produce the final prediction. A joint loss function combining Euclidean and adaptive Bayesian losses improves density map accuracy. Evaluated on ShanghaiTech, UCF-CC-50, and NWPU-Crowd datasets, the proposed method achieves state-of-the-art MAE and RMSE across sparse and dense scenes. Ablation studies confirm the effectiveness of MFFM, CBAM, and DCM modules. The model demonstrates strong generalization and robustness, even under low-light, occlusion, and background clutter.

Keywords: Crowd counting, Lightweight pyramid network, Multi-modal feature fusion, Dilated convolution module

http://dx.doi.org/10.6180/jase.202612_35.025

1. Introduction

Crowd counting estimates the quantity, density and spatial distribution of pedestrians within images and video frames. As a fundamental task in intelligent video surveillance, it supports downstream visual analysis including behavior recognition, crowd congestion assessment, anomaly and event detection [1]. Due to its great value for public security, traffic regulation, crowd supervision and disaster emergency disposal, related technologies have drawn extensive research interest across various fields [2].

Deep learning techniques have gradually become the mainstream for crowd counting, showing outstanding overall performance. Wang et al. [3] proposed DMCount by using a deep convolutional neural network (CNN) as the feature extractor and employed the optimal transport (OT) theory to measure the distribution matching degree. It had achieved significant improvements in counting performance on multiple benchmark datasets, fully demonstrating the technical advantages of deep learning methods in this field. However, such methods highly rely on large-scale labeled samples to build high-precision and high-robust models. In dense crowd scenarios, the acquisition of high-quality labeled data faces the practical dilemma of high labor costs. It contains 1.25 million head instance annotations, and the cumulative labor time is 2000 hours in UCF-QNRF dataset. To reduce the reliance on costly and time-consuming manual annotation, semi-supervised crowd counting methods have gradually emerged [4]. These methods start model training using a small set of labeled samples. After that, unlabeled data are inferred to produce pseudo-labels, and only reliable ones are chosen to expand the training set. Through multiple rounds of iterative training, the counting performance and generalization ability of the model are improved. Thus, the quality of pseudo-label generation is the key factor determining the performance of semi-supervised crowd counting models.

A confidence threshold is applied to screen pseudo-labels during every iteration, so as to exclude low-quality

labels and suppress training noise [5]. We select reliable predictions as new training samples to dynamically upgrade the dataset. Ideally, high confidence should correspond to high reliability, meaning that the model has a high degree of trust in its prediction results, and this trust is based on reliable prediction results. However, in practical applications, high confidence does not always mean high reliability. Due to its own cognitive limitations, the model is prone to giving overly high confidence to incorrect predictions. That is, cognitive uncertainty can lead to prediction biases in scenarios not covered by the training data. At the same time, arbitrary uncertainties caused by inherent noise in the data (including annotation errors, background interference, etc.) will also introduce prediction biases [6, 7].

Aiming at the adverse impact of background noise on counting models, Zhu et al. [8] designed SFANet with dual-path structure, multi-scale fusion and attention mechanism. The spatial attention information from one path guided the density feature learning of the other, which weakened noise disturbance and yielded high-fidelity density maps. However, the quality of the generated spatial attention map was relatively low, and the counting effect was still poor in scenes with complex backgrounds. Li et al. [9] added spatial attention mechanism and channel attention mechanism to CSRNet (Network for Congested Scene Recognition), enabling it to better learn the key information in the image and reduce the influence caused by complex background environments and occlusions. However, their model structures are complex and the training was difficult.

Above methods merely perform simple feature extraction on the input image, ignoring the large-scale pixel-level context information and failing to effectively solve the problem of the shadow occlusion affecting the counting effect. For crowd counting models, the ability to capture comprehensive multi-scale information and resist background noise directly determines their practical performance. To address the existing limitations in current methods, we design an innovative crowd counting framework that integrates multi-modal feature fusion with a lightweight pyramid structure. The entire network adopts a modular design, mainly incorporating the VGG-16 feature extractor, multi-scale feature fusion module (MFFM), convolutional block attention module (CBAM) and dilated convolution module (DCM).

The working flow of the proposed network is arranged step by step. Firstly, the VGG-16 backbone is used to extract fundamental visual features from raw input images. The acquired primary features are subsequently sent into MFFM. Through multi-scale fusion operations, this module fully excavates feature information at different scales and complements rich semantic content for subsequent processing. Afterwards, CBAM is introduced to conduct adaptive optimization on the fused feature maps. By adjusting the importance of different feature channels and spatial regions, this module effectively suppresses irrelevant background noise and highlights crowd-related feature areas.

Following the feature calibration stage, DCM takes over the subsequent processing. Relying on dilated convolution operations, it generates two key outputs: background segmentation attention maps and crowd density feature maps. We fuse these two outputs via element-wise multiplication, which can effectively eliminate the influence of cluttered backgrounds and construct high-quality density prediction maps. Benefiting from the above design, our method realizes precise estimation of crowd numbers.

We evaluate the proposed method on three widely adopted benchmark datasets: ShanghaiTech, UCF-CC-50 and NWPUCrowd. Experimental outcomes verify that our developed algorithm possesses prominent advantages and superior counting accuracy compared with conventional solutions.

2. Materials and methods

The proposed multi-modal feature fusion model based on lightweight pyramid network for crowd counting is shown in Figure 1. Given the powerful feature extraction capability and standardized network structure of VGG-16, this paper adopts VGG-16 as the backbone network with the features of conv2_2, conv3_3, conv4_3, and conv5_3 as the output features. Considering that the shallow features have more noise, only conv3_3, conv4_3, and conv5_3 are retained as the input features for the next stage. On this basis, a multi-scale feature fusion module is proposed to fuse the multi-level features through top-down and lateral connections to address the problem of scale continuity. In addition, to suppress background noise, the channel spatial attention module is introduced to calibrate the input feature maps in the channel spatial domain, highlighting the head region of the image. To further capture fine-grained details from input images, the well-calibrated feature representations are imported into dilated convolution units. Through hierarchical deep feature extraction, the network outputs two core feature maps for subsequent calculation: background segmentation attention maps and density distribution maps.

2.1. Multi-scale Feature Fusion Module (MFFM)

The multi-scale feature fusion module introduces the ASPP (Atrous Spatial Pyramid Pooling) and LPN (lightweight Pyramid Network) to build a feature pyramid to address

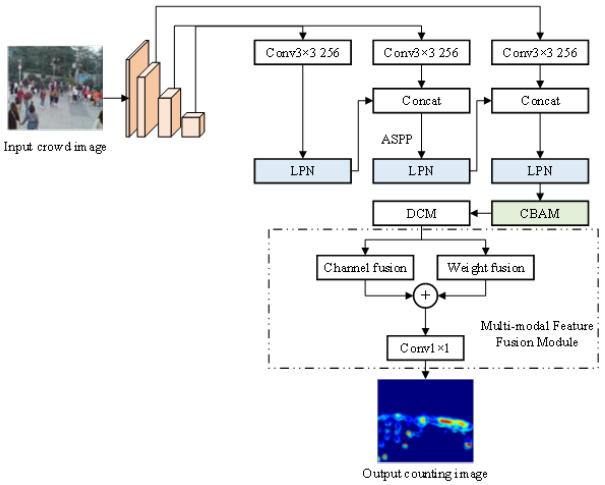

Fig. 1. Proposed crowd counting structure

the scale variation problem in dense crowd images. CNN has strong adaptability to positional changes, but its ability to cope with scale changes is somewhat lacking. Therefore, Lin et al. [10] proposed the Feature Pyramid Network (FPN), which could simultaneously utilize the strong semantic information of deep features and the high resolution of shallow features, and fuse features at different levels. By fully exploiting multi-scale spatial cues and pivotal cross-dimensional features from channel attention vectors, the LPN module strengthens the network's capacity to learn fine-grained multi-scale characteristics. Ding et al. [11] proposed an ASPP module with dilated spatial pooling, using multiple sampling rate convolution kernels to detect the input convolution features, thereby obtaining more abundant multi-scale information. Inspired by the above work, this paper designs a multi-scale feature fusion structure and introduces the ASPP and LPN modules that can extract finer-grained spatial information to fuse the branches at different scales, thereby solving the problem of continuous scale variation.

The main steps of LPN are as follows. First, the input vector X is divided into S groups from the channel, and each group undergoes convolution with different-sized convolution kernels to extract information of different scales. Results are fused in a stacked manner to obtain the tensor X'. Secondly, the channel attention of different scale feature maps is extracted using the attention weight module, and the channel attention vector is obtained. Then, the Softmax function is used to re-calibrate the multi-scale channel attention vectors to obtain the attention weight vector Z after multi-scale channel interaction. Finally, the element-wise dot product of Z and X' is performed to obtain a more refined feature map with richer multi-scale feature information [12]. ASPP captures multi-scale information through several different sampling rates of dilated convolutions. The difference between dilated convolution layers and general convolutions lies in the dilation rate, which controls the padding and dilation during convolution. Different padding and dilation can obtain different scales of receptive fields, thereby achieving free multi-scale feature extraction.

The output features conv5_3, conv4_3 and conv3_3 from the backbone network are respectively represented as C1, C2 and C3, which are used as the input of MFFM. Then, 3×3 convolution operations are performed on C1, C2 and C3 to unify the channel numbers and obtain $C1'$, $C2'$ and $C3'$. The $C1'$ is input into the LPN module to obtain a feature map P1 with richer multi-scale information expression ability. P1 is up-sampled to double its size and concatenated with $C2'$. Then it is input into the ASPP module and LPN module to obtain different scales of receptive fields and extract more detailed spatial information, resulting in the feature map P2. P2 is up-sampled and concatenated with $C2'$, and then input into the LPN module to generate feature map P3. Feature map P3 is the final output feature map of the multi-scale feature fusion module and is used as the input of the CBAM module. The rich semantic information and multi-scale pedestrian feature representation can effectively solve the problem of scale feature variation and improve the accuracy of pedestrian counting in small and medium scales.

2.2. CBAM module

Collaborating the VGG-16 backbone with multi-scale fusion modules allows the system to acquire optimized feature representations that contain diverse scale characteristics. This approach addresses the core challenge of scale fluctuation among crowded pedestrians. Furthermore, the attention unit is leveraged to recalibrate spatial feature weights. Rather than capturing redundant environmental noise including buildings and trees, the model prioritizes effective head-related features to boost counting accuracy [13]. In this paper, a CBAM module is inserted after the multi-scale feature fusion module. The CBAM module not only focuses on the channel domain information, but also on the spatial domain information, learning from two dimensions. Given an intermediate feature map, by learning the channel relationship and spatial relationship between features, a channel attention map and a spatial attention map are generated. These features are adaptively adjusted, so that the weight of the crowd area in the feature map can be improved. Specifically, this is represented by Eqs. (1) and (2).

$$F' = M_c(F) \otimes F \tag{1}$$

$$F'' = M_s(F') \otimes F' \tag{2}$$

Where, $\otimes$ represents element-wise multiplication. F represents the input feature map. M_c is the channel attention operation. F' is the output feature map of the channel attention module. M_s is the spatial attention operation. F'' is the output feature map of the spatial attention module.

The output of the multi-scale feature fusion stage is a multi-scale feature map that integrates features from different depths. Different depth feature maps have different perception capabilities when they are applied to heads of different sizes. CBAM adjusts the channel dimension to increase the correlation on the dimension, enabling the model to have strong adaptability when facing scenarios with varying crowd scales and uneven distribution. In addition, CBAM recalibrates the feature maps in the spatial dimension, allowing the network to learn the key information in the crowd images, enhancing the model's background suppression ability and effectively solving common problems such as occlusion and complex environments in the images.

2.3. Multi-modal Feature Fusion Module

After completing the multi-modal feature interaction enhancement, this paper designs a multi-modal feature fusion module, which consists of two parts: channel fusion and global-local attention interaction fusion. The crowd features F1 and F2 obtained through feature extraction are concatenated with the coarse channels of multi-modal information as shown in Eq. (3).

$$f = C(F1, F2) \tag{3}$$

Where, C represents the concatenation operation. f is the feature of the coarse channel. Then, the feature map f is aggregated using the double pooling method to enhance the discriminative ability of multi-modal crowd feature channels, and the aggregated feature dimension M_c is obtained, as shown in Eq. (4).

$$M_c = \sigma\left(FC\left(GAvgPool(f)\right) + FC\left(GMaxPool(f)\right)\right) \tag{4}$$

Here, $GAvgPool$ represents average pooling. $GMaxPool$ represents maximum pooling. FC stands for fully connected. σ represents the Gaussian activation function.

Finally, two stacked convolutions are utilized to obtain the channel-fused crowd feature f_{RT}^c, as shown in Eq. (5).

$$f_{RT}^c = Conv_{3\times3}^2\left(M_c\right) \tag{5}$$

Here, $Conv_{3\times3}^2$ represents two stacked 3×3 convolutions.

To enhance the complementarity among multi-modal features and improve the performance of multi-modal feature fusion, a local and global attention interaction fusion module is further designed. Through the collaboration of local and global attention, the global context information of crowd features and the local detailed feature information are obtained, achieving deep fusion and effectively enhancing the feature representation ability to improve the performance of crowd counting.

In the interaction fusion module, the global attention mechanism is used to focus on the overall information of the crowd images, and the global feature F_G can be calculated using formula (6).

$$F_G = b(Comv_{1\times1}(Relu(b(Conv_{1\times1}(GAP(\widehat{F}_{RGB} + \widehat{F}_t)))))) \tag{6}$$

Where, $\widehat{F}_{RGB}$ and $\widehat{F}_t$ represent the input features in the visible light and thermal image branches. GAP represents global average pooling. b represents BN batch normalization.

The local attention mechanism focuses on the local areas in the crowd image, directing its attention to the fine-grained features of the crowd region. The local features are calculated using Eq. (7).

$$F_L = b(Conv_{1\times1}(Relu(b(Conv_{1\times1}(GAP(\widehat{F}_{RGB} + t)))))) \tag{7}$$

The integration of global and local features facilitates the acquisition of fusion weight W. This parameter is utilized to score and optimize the input features of the visible light branch, and the processed weighted feature F_{WRGB} is presented in Eq. (8).

$$F_{WRGB} = (W) \times \widehat{F}_t = \left(1 - \delta(F_G \oplus F_L) \times \widehat{F}_t\right) \tag{8}$$

Here, δ represents the sigmoid activation function.

For the thermal image branch, the fusion weight $(1 - W)$ is obtained to get the weighted feature Fr as shown in Eq. (9).

$$F_{WT} = (1 - W) \times \widehat{F}_t = \left(1 - \delta\left(F_G \oplus \widehat{F}_L\right) \times \widehat{F}_t\right) \tag{9}$$

Table 1. Experimental dataset information

Parameter	ShanghaiTech		UCF-CC-50	NWPU-Crowd
	Part A	Part B		
Image number	482	761	50	5 109
Average size	589 × 868	768 × 1024	2101 × 2888	2191 × 3209
Average number of crowd	501	123	1 279	418
Maximum number of crowd	3139	578	4633	20033
Minimum number of crowd	33	9	94	0
Total crowd	241677	88488	63974	2133375

Table 2. Comparison results with different algorithms on the ShanghaiTech, UCF-CC-50 and NWPU-Crowd

Method	ShanghaiTech Part A		ShanghaiTech Part B		UCF-CC-50		NWPU-Crowd	
	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
MCNN [14]	110.2	173.2	26.4	41.3	377.6	509.1	245.8	700.3
SwitchCNN [15]	90.4	135.0	21.6	33.4	318.1	439.2	218.5	700.6
CSRNet [16]	68.2	115.0	10.6	16.0	266.1	397.5	171.2	471.5
TEDNet [17]	64.2	109.1	8.2	15.7	249.4	354.5	141.4	630.7
DUBNet [18]	64.6	106.8	7.7	12.5	243.8	329.3	138.6	625.1
DPDNet [19]	66.6	120.3	7.9	12.4	251.2	337.1	124.8	610.2
URC [20]	72.8	111.6	12.0	18.7	293.9	443.0	113.6	569.1
MPS [21]	71.1	110.7	9.6	15.0	267.1	368.5	108.8	553.7
AutoScale [22]	65.8	112.1	8.6	13.9	258.3	323.4	105.8	504.4
FusionCount [23]	62.2	101.2	6.9	11.8	205.7	289.5	104.9	433.5
Proposed	63.6	100.2	6.9	11.3	196.2	274.6	102.3	416.2

2.4. Loss Function

To optimize the network, we integrate Euclidean distance loss with Bayesian loss to build a hybrid loss function. Among them, Euclidean loss $L(\theta)$ measures the deviation between predicted density maps and corresponding annotated data, as defined in Eq. (10).

$$L(\theta) = \frac{1}{2N}\sum_{i=1}^{N} \|A_i(Y_i;\theta) - B_i^{GT}\|_2^2 \tag{10}$$

Where θ represents the learning parameters. N represents the number of training samples. $A_i(Y_i;\theta)$ and B_i^{GT} represent the predicted density map and the true density map of the $i-th$ training image respectively.

Based on the Eq. (10), by combining Bayesian loss, the global density distribution of the crowd and the local counting accuracy are further enhanced. Bayesian loss builds a density probability model based on the marked points, which can associate the counts of the marked points with the pixels in the entire crowd image. However, in dense crowd areas, the noise of the marked points may cause gradient fluctuations during the training process, affecting the stability of optimization. Therefore, in this paper, an adaptive Gaussian kernel width the neighborhood distance of the marked points is introduced before the Bayesian loss. By dynamically adjusting the standard deviation of the Gaussian kernel for each marked point, the influence of noise can be effectively weakened. The calculation is as shown in Eqs. (11) and (12).

$$d_i = \frac{1}{k}\sum_{j=1}^{k} \|X_i - X_j^{(i)}\|_2 \tag{11}$$

$$\sigma_i = \beta \times d_i \tag{12}$$

Where, X_i represents the position of the $i-th$ labeled point. $X_j^{(i)}$ is the coordinate of the $j-th$ $(j \in (1,.,k))$ nearest neighbor labeled point of the $i-th$ labeled point. d_i is the average distance between $X_j^{(i)}$ and the k nearest neighboring labeled points. β is a learnable parameter.

The process of constructing the real density map $C^{gt}(X_m)$ through adaptive σ_i is as shown in Eq. (13).

$$C^{gt}(X_m) = \sum_{i=1}^{K} \frac{1}{2\pi\sigma_i^2} \exp\left(-\frac{\|X_m - X_{m_i}\|^2}{2\sigma_i^2}\right) \tag{13}$$

Where, K represents the total number of labeled points in the current image. X_m denotes the position of a pixel. X_{m_i} is the position of the $i-th$ labeled point. The Bayesian loss L_b is calculated using the constructed density map, and the specific process is as shown in Eq. (14).

$$L_b = \sum_{m=1}^{M} F\left(C^{gt}(X_m) - C^{pre}(X_m)\right) \tag{14}$$

Here F represents the distance function, while $C^{gt}(X_m)$ and $C^{pre}(X_m)$ denote the true density map and the predicted density map respectively. M represents the total number of pixels in the density map.

The final loss L_{loss} is composed of the Euclidean distance loss $L(\theta)$ and the Bayesian loss L_b, as shown in Eq. (15).

$$L_{loss} = L(\theta) + \alpha L_b \tag{15}$$

Here, α is a hyperparameter. In this paper, we set α= 0.1.

3. Results and discussion

Three widely recognized crowd counting datasets, including ShanghaiTech, UCF-CC-50 as well as NWPU-Crowd, are employed for experimental validation. The fundamental attributes of the above benchmarks are listed in

To verify the effectiveness of the proposed method, this paper conducts training on three challenging datasets and compares the counting accuracy with some excellent algorithms. Results are shown in

From Table 2, it can be seen that Proposed method achieves MAE and RMSE of 63.6 and 100.2 respectively on the Part A. Compared with the sub-optimal FusionCount, although MAE slightly decreases, RMSE decreases by 0.98%, indicating that Proposed method has better stability. On the Part B, MAE and RMSE reach the current optimal values of 6.9 and 11.3, indicating that Proposed method also has good performance in scenarios with relatively sparse crowds. On the UCF_CC_50 with small samples and a large density range, compared with URC, MAE decreases by 33.2% and RMSE decreases by 38.0%, indicating that the algorithm in this paper also has good counting performance even with a small number of samples. On the latest large-scale dataset NWPU-Crowd, although the negative samples increases the training difficulty to some extent, it also improves the generalization ability of the model. Compared with other algorithms, MAE and RMSE of proposed method reach the current optimal values of 102.3 and 416.2 respectively, indicating that Proposed method still has good counting accuracy in cases with a large number of samples and significant differences in image resolution.

The experimental data in Table 2 demonstrate that regardless of whether the crowd is dense or sparse, the new algorithm in this paper has high counting accuracy and exhibits excellent generalization ability. Even in complex scenarios such as significant changes in crowd density, blurred faces, and background occlusion, it still maintains good counting accuracy, indicating that the new algorithm has strong robustness. Several qualitative visualization cases on the ShanghaiTech dataset are displayed in Figure 2, aiming to intuitively reflect the overall performance of the proposed approach. Specifically, GT corresponds to the ground-truth crowd number, and Pre stands for the counting result estimated by our algorithm.

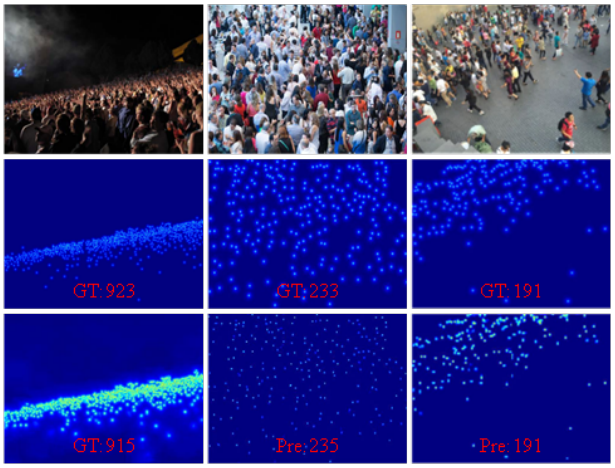

Fig. 2. Visualized results with proposed method

From the first column in Figure 2, it can be seen that for images with dim lighting and a large number of people, the proposed method can also maintain a relatively small relative error. From the second column in Figure 2, it can be seen that for images with dense population distribution, the proposed method predicts the crowd number very close to the actual number. From the third column in Figure 2, it can be seen that for images with a relaxed crowd and a simple scene, the proposed method accurately predicts the correct crowd number. In conclusion, the presented algorithm can effectively adapt to scenarios with scale variations and severe background interference in dense crowds. The generated predicted density map is closer to the actual density map, demonstrating the superior performance of proposed method in the field of crowd counting.

The proposed method is mainly composed of MFFM, CBAM and DCM modules. To further verify the rationality and effectiveness of each part in the new structure, ablation experiments are conducted on the ShanghaiTech dataset Part A. Specifically, experiments are carried out on MFFM, CBAM, DCM, MFFM+CBAM, MFFM+DCM, CBAM+DCM, and MFFM+CBAM+DCM. The result is shown in

From Table 3, it can be seen that each of the MFFM, CBAM, and DCM can obtain population information to a certain extent, but the counting accuracy is relatively low. After integrating MFFM and DCM, the counting accuracy

Table 3. Ablation experiment results

Module			MAE	RMSE
MFFM	CBAM	DCM		
√	×	×	70.2	106.5
×	√	×	70.6	114.5
×	×	√	68.2	115.0
√	√	×	67.8	107.1
√	×	√	66.1	108.3
×	√	√	65.6	117.6
√	√	√	63.6	100.2

Table 4. Ablation experiment results of α value

α	Part A		Part B	
	MAE	MSE	MAE	MSE
0	65.8	101.5	7.3	11.8
0.1	63.6	100.2	6.9	11.3
0.2	65.2	99.6	7.6	11.5
0.5	64.7	102.3	7.2	12.0

has increased, indicating that MFFM can enhance the expression ability of multi-scale features and the counting performance. When CBAM is combined with MFFM and DCM respectively, compared with using a single MFFM and DCM, MAE has decreased by 3.4% and 3.8% respectively, indicating that CBAM has a good inhibitory effect on background noise. The above experimental results show that the multi-scale feature fusion module and the channel spatial attention module have a strong improvement effect on the performance.

In the joint loss function used in this paper, the hyperparameter α is employed to balance the weights between $L(\theta)$ and L_b. To verify the impact of different hyperparameters on the experimental results, this paper conducts ablation experiments on the ShanghaiTech dataset for different α. Experimental results are shown in

From Table 4, it can be seen that when α=0.1, the MAE values on both Part A and Part B reach the lowest values with 63.6 and 6.9 respectively. At the same time, the RMSE on Part B achieves the best result. When α=0.2, the RMSE result on Part A is the best, but its MAE is slightly inferior to that of α= 0.1. Meanwhile, the MAE and RMSE on Part B do not reach the best results either. Therefore, this paper selects 0.1 as the value of the hyperparameter α.

A series of contrast experiments are performed to evaluate the lightweight property of the proposed network. We benchmark parameter size, FLOPs and running FPS between our model and other prevalent lightweight approaches under the same RTX 3090 hardware environment, as recorded in

The proposed method achieves the smallest parameter size (12.1M) and FLOPs (6.5G), while reaching the highest inference speed (17.3FPS). Compared with FusionCount, our method reduces parameters by 10.4% and increases FPS by 9.5%. The results confirm that the lightweight pyramid network and efficient multi-modal fusion module strike a good balance between accuracy and efficiency, meeting the real-time deployment requirements of edge devices in video surveillance.

To simulate complex real-world scenarios, we add Gaussian noise ($\sigma = 0.01, 0.05$) and salt-and-pepper noise (density $= 0.05, 0.1$) to test images on ShanghaiTech Part A. The robustness results are shown in Table 6. Under low-level noise interference, the proposed method only has a slight increase in MAE ($\leq$ 5.2) and RMSE ($\leq$ 5.5). Even under high-level noise, MAE remains below 76, which outperforms most existing methods under the same noise setting. This further verifies that the CBAM and DCM modules effectively suppress noise and enhance foreground feature representation, enabling the model to maintain stable counting performance under complex noise interference.

4. Conclusions

This paper presents a lightweight pyramid network integrating multi-modal feature fusion, CBAM-based attention, and dilated convolutions for accurate crowd counting under scale variation and background noise. Experimental results across multiple mainstream crowd counting benchmarks reveal that the developed approach can maintain outstanding counting performance under various complex scenarios. Additional ablation analyses validate the functionality of individual modules. Furthermore, this study summarizes three avenues for subsequent research: (1) extending the framework to video-based counting by incorporating temporal consistency and motion cues; (2) exploring self-supervised and few-shot learning to reduce reliance on dense annotations; and (3) optimizing the model for real-time edge deployment through neural architecture search and quantization techniques.

Table 5. Comparison of model complexity and inference efficiency

Method	Params (M)	FLOPs (G)	FPS
CSRNet	16.2	8.7	12.5
DUBNet	14.8	7.9	14.2
FusionCount	13.5	7.2	15.8
Proposed	12.1	6.5	17.3

Table 6. Robustness to noise interference (ShanghaiTech Part A)

Noise Type	Noise Level	MAE	RMSE
Gaussian	σ=0.01	66.8	105.7
Gaussian	σ=0.05	72.3	114.9
Salt-and-Pepper	0.05	68.5	108.2
Salt-and-Pepper	0.1	75.1	118.6

Acknowledgement

This work was supported by the 2023 Annual Basic Research and Academic Work Expenses Project of Provincial Higher Education Institutions of Heilongjiang Province (2023-KYYWF-0582); Jiamusi City's Municipal Science and Technology Plan Innovation Incentive Project (GY2025JL0007); 2023 Youth Innovation Talent Cultivation Support Project of Jiamusi University (JMSUQP23007).

References

[1] G. Gao, J. Gao, Q. Liu, Q. Wang, and Y. Wang, (2025) *"A survey of deep learning methods for density estimation and crowd counting"* **Vicinagearth 2**(1): 2. DOI: 10.1007/s44336-024-00011-8.

[2] Y. Yu, F. Zhu, J. Qian, H. Fujita, J. Yu, K. Zeng, and E. Chen, (2025) *"CrowdFPN: crowd counting via scale-enhanced and location-aware feature pyramid network: Y. Yu et al."* **Applied Intelligence *55***(5): 359. DOI: 10.1007/s10489-025-06263-1.

[3] H. Lin, X. Hong, Z. Ma, Y. Wang, and D. Meng, (2024) *"Multidimensional measure matching for crowd counting"* **IEEE Transactions on Neural Networks and Learning Systems *36***(5): 9112–9126. DOI: 10.1109/TNNLS.2024.3435854.

[4] M. Ling, J. Chen, Y. Liu, W. Fang, and X. Geng, (2025) *"Dual-branch adjacent connection and channel mixing network for video crowd counting"* **Pattern Recognition *167***: 111709. DOI: 10.1016/j.patcog.2025.111709.

[5] S. Yin, L. Wang, T. Chen, H. Huang, J. Gao, J. Zhang, M. Liu, P. Li, and C. Xu, (2026) *"LKAFormer: A lightweight kolmogorov-arnold transformer model for image semantic segmentation"* **ACM Transactions on Intelligent Systems and Technology *17***(3): 1–24. DOI: 10.1145/3759254.

[6] Y. Hu, Y. Liu, G. Cao, and J. Wang, (2025) *"CrowdCL: unsupervised crowd counting network via contrastive learning"* **IEEE Internet of Things Journal *12***(12): 21704–21719. DOI: 10.1109/JIOT.2025.3547898.

[7] Y. Qian, L. Zhang, Z. Guo, X. Hong, O. Arandjelović, and C. R. Donovan, (2025) *"Perspective-assisted prototype-based learning for semi-supervised crowd counting"* **Pattern Recognition *158***: 111073. DOI: 10.1016/j.patcog.2024.111073.

[8] Z. Zou, Y. Cheng, X. Qu, S. Ji, X. Guo, and P. Zhou, (2019) *"Attend to count: Crowd counting with adaptive capacity multi-scale CNNs"* **Neurocomputing *367***: 75–83. DOI: 10.1016/j.neucom.2019.08.009.

[9] Y. Li, X. Zhang, and D. Chen. "Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes". In: *2018 IEEE/CVF conference on computer vision and pattern recognition*. IEEE. 2018, 1091–1100. DOI: 10.1109/CVPR.2018.00120.

[10] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie. "Feature pyramid networks for object detection". In: *2017 IEEE conference on computer vision and pattern recognition (CVPR)*. Ieee. 2017, 936–944. DOI: 10.1109/CVPR.2017.106.

[11] P. Ding, H. Qian, Y. Zhou, S. Yan, S. Feng, and S. Yu, (2023) *"Real-time efficient semantic segmentation network based on improved ASPP and parallel fusion module in complex scenes"* **Journal of Real-Time Image Processing *20***(3): 41. DOI: 10.1007/s11554-023-01298-4.

[12] S. Yin, H. Li, A. A. Laghari, T. R. Gadekallu, G. A. Sampedro, and A. Almadhor, (2024) *"An anomaly detection model based on deep auto-encoder and capsule graph convolution via sparrow search algorithm in 6G Internet of Everything"* **IEEE Internet of Things Journal *11***(18): 29402–29411. DOI: 10.1109/JIOT.2024.3353337.

[13] L. Chen, H. Yao, J. Fu, and C. T. Ng, (2023) *"The classification and localization of crack using lightweight convolutional neural network with CBAM"* **Engineering Structures *275***: 115291. DOI: 10.1016/j.engstruct.2022.115291.

[14] Y. Zhang, D. Zhou, S. Chen, S. Gao, and Y. Ma. "Single-image crowd counting via multi-column convolutional neural network". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, 589–597. DOI: 10.1109/CVPR.2016.70.

[15] D. Babu Sam, S. Surya, and R. Venkatesh Babu. "Switching convolutional neural network for crowd counting". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, 5744–5752. DOI: 10.1109/CVPR.2017.429.

[16] X. Tian and H. Hiraishi, (2025) *"Design of crowd counting system based on improved CSRNet"* **Artificial Life and Robotics *30***(1): 3–11. DOI: 10.1007/s10015-024-00993-0.

[17] J. Yi, F. Chen, Z. Shen, Y. Xiang, S. Xiao, and W. Zhou, (2023) *"An effective lightweight crowd counting method based on an encoder–decoder network for internet of video things"* **IEEE Internet of Things Journal *11***(2): 3082–3094. DOI: 10.1109/JIOT.2023.3294727.

[18] J. Xu, Z. Zhang, X. Li, W. Li, and K. Yu. "Attention Mixture Network for Crowd Counting via Binarization Transfer". In: *Proceedings of the 2nd International Workshop on Multimedia Content Generation and Evaluation: New Methods and Practice*. 2024, 45–53. DOI: 10.1145/3688867.3690172.

[19] L. Zhou, P. Wang, W. Li, J. Leng, and B. Lei, (2022) *"Semantic-refined spatial pyramid network for crowd counting"* **Pattern Recognition Letters *159***: 9–15. DOI: 10.1016/j.patrec.2022.04.029.

[20] J. Yu and H. Hu, (2025) *"Multiscale regional calibration network for crowd counting"* **Scientific Reports *15***(1): 2866. DOI: 10.1038/s41598-025-86247-w.

[21] K. Liu, Z. Dou, F. Wang, X. Xia, and J. Sang, (2025) *"Cross-level attention multi-scale context-enhanced crowd counting network for transportation cyber-physical systems"* **IEEE Transactions on Intelligent Transportation Systems *26***(9): 14250–14263. DOI: 10.1109/TITS.2025.3566718.

[22] B. Yan, Y. Li, L. Dong, Z. Ren, H. Liu, X. Gao, and W. Cheng, (2025) *"Crowd counting with WiFi sensing based on iterative attentional feature fusion"* **Computer Communications *241***: 108245. DOI: 10.1016/j.comcom.2025.108245.

[23] R. Ma, Y. Hou, C. Li, H. Jia, and X. Xie, (2025) *"Scene-adaptive unsupervised crowd counting for video surveillance"* **IEEE Transactions on Circuits and Systems for Video Technology *35***(7): 6910–6925. DOI: 10.1109/TCSVT.2025.3540850.