DOI: 10.23977/jaip.2025.080404
ISSN 2371-8412 Vol. 8 Num. 4

NF-Net: Crowd Counting Based on Near-Far Network and Dynamic Dual Attention Mechanism

Zhang Dongming[1], Tian Xueqing[1,a,*], Zhao Wenhui[1], Guo Yihan[1], Chen Lijia[1]

[1]School of Physics and Electronics, Henan University, Kaifeng, 475004, China
[a]txq@henu.edu.cn
**Corresponding author*

Keywords: Crowd counting; Point localization; Multi-scale feature modelling; Attention mechanism; Deep learning

Abstract: Crowd counting and precise localization in dense scenes are critical tasks in computer vision. Although the point-based prediction framework P2PNet eliminates complex post-processing steps through ensemble prediction, it still faces challenges such as insufficient multi-scale feature extraction and limited ability to perceive key information in complex scenes. To address these issues, this paper improves P2PNet by proposing a model based on a far-near network and a dynamic dual attention mechanism. Specifically, it introduces a far-near adaptive network (FN-Net) and a dynamic dual attention mechanism (DDAM). FN-Net explicitly models continuous scale variations caused by perspective effects by dividing the image into regions based on spatial position and assigning differentiated receptive fields. DDAM focuses on crowded areas through parallel spatial attention and channel attention sub-modules, selects discriminative features, and integrates a dynamic weighted fusion mechanism to adaptively combine the advantages of both attentions. Experiments show that our approach effectively enhances key features while suppressing background noise thus improves crowd counting accuracy.

1. Introduction

Crowd counting in dense scenes is a critical and challenging task in computer vision, with wide-ranging applications in public safety, traffic monitoring, and urban planning. Key technical challenges include significant scale variations, severe occlusions, complex background interference, and highly crowded environments.

Recent advances in deep learning have given rise to two primary paradigms: density map regression methods and detection-based methods. Density map regression approaches estimate total counts but face challenges in achieving precise localization. Detection-based methods predict bounding boxes or points; however, they are sensitive to occlusion and scale variations and often require complex post-processing.

To address these limitations, P2PNet was proposed as a purely point-based prediction framework that simultaneously achieves counting and localization without requiring post-processing. However, it still suffers from limited feature representation and insufficient multi-scale modeling.

This paper improves P2PNet along two core dimensions: feature enhancement and multi-scale modeling. The main contributions are:

(1) We have developed a Far-Near Adaptive Network (FN-Net) that leverages distance information. Partitions the feature map into regions based on distance and assigns differentiated receptive fields to address scale variations caused by perspective.

(2) A dynamic dual attention mechanism is proposed. Integrates spatial and channel attention mechanisms with dynamic fusion to enhance key features and suppress noise.

(3) Extensive experimental validation has been conducted on multiple public datasets. The results demonstrate that it outperforms existing methods in both counting and localization metrics, effectively validating the rationale and effectiveness of each module.

2. Related Work

2.1. Crowd Counting Based on Detection and Localization

Unlike mainstream density map regression methods, the detection- and localization-based paradigm aims to directly predict the position of each pedestrian in an image, typically represented as bounding boxes or points. P2PNet, by introducing a one-to-one matching strategy, effectively eliminates the need for post-processing steps such as non-maximum suppression, thereby simplifying the process and enhancing localization robustness in dense crowds [1]. However, its performance remains constrained by the capacity of the backbone network and the effectiveness of multi-scale modeling.

2.2. Multi-Scale Feature Modeling

There is significant scale variation in crowd images, and effective multi-scale feature modeling is crucial for improving counting and localization performance. Early works, such as the multi-column structure of MCNN [2], the dilated convolutions in CSRNet [3], and methods like the Transformer [4], have been employed to capture global context. Within the point localization framework, efficiently and adaptively fusing multi-scale features to handle continuous scale variations from near to far remains an important research challenge. The FN-Net proposed in this paper offers a novel solution to this problem through distance-aware region partitioning and receptive field allocation.

2.3. Application of Attention Mechanisms in Crowd Counting

Most existing approaches employ only one type of attention mechanism independently, lacking coordinated optimization across both spatial and channel dimensions. This paper proposes a dynamic dual attention mechanism that combines both types and utilizes dynamic fusion to achieve enhanced synergistic effects.

3. Proposed Method

The framework is based on P2PNet and incorporates FN-Net and DDAM. The structure of these modules is illustrated in Figure 1.

3.1. Far-Near Adaptive Network (FN-Net)

To explicitly address scale variations caused by perspective, we propose the FN-Net module, which assigns different receptive fields to objects based on their spatial positions within the image.

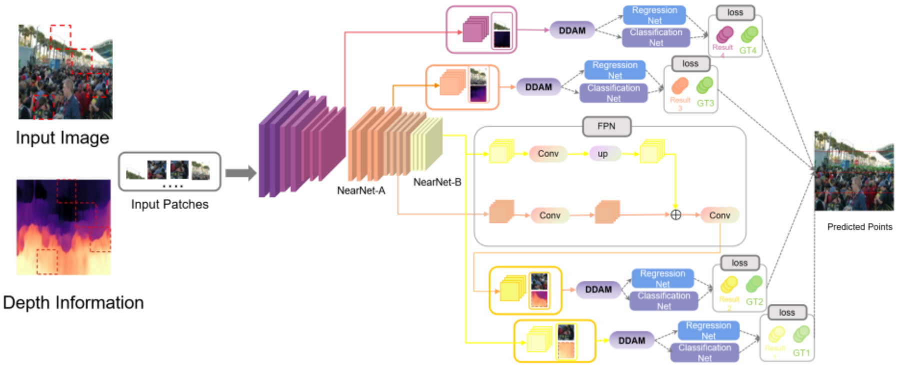


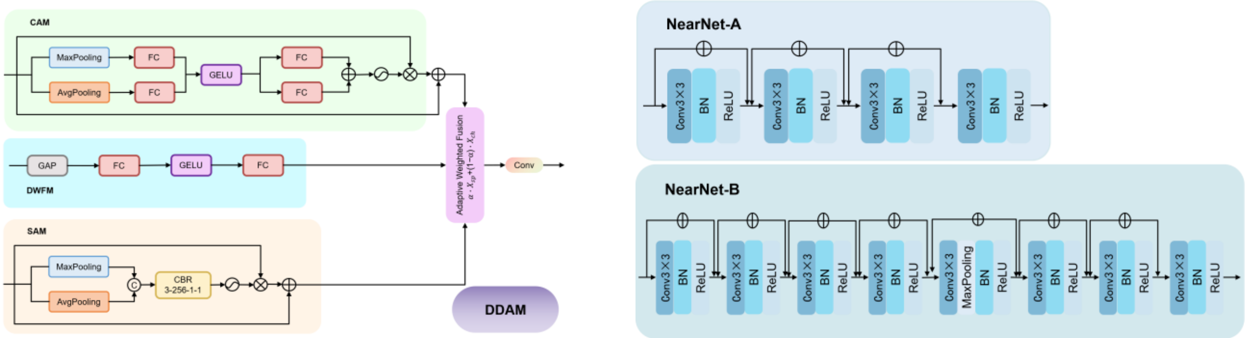
Figure 1: Structural overview of the FN-Net proposed in this paper.

FN-Net utilizes VGG_16_bn to extract deep features, followed by the integration of NearNet-A and NearNet-B modules. Different convolutional layers are stacked to assign smaller receptive fields to high-density populations for capturing fine details, and larger receptive fields to low-density populations to capture broader contextual information.

Assume we have the feature map output by the backbone network $F_s \in \mathbb{R}^{H_f \times W_f \times C_f}$. We divide it evenly into four regions along the height dimension R_1, R_2, R_3, R_4. For each region R_i, we apply a convolutional network branch of a specific depth for NearNet processing. The detailed architecture of NearNet is shown in Figure 2(b). After processing through their respective branches, all regions are reassembled into the original spatial layout to form the enhanced feature map F_s'.

3.2. Dual Dynamic Attention Mechanism

To enhance the model's ability to focus on key features of the crowd while suppressing background noise, we designed a Dynamic Dual Attention Module (DDAM), the structure of which is shown in Figure 2(a). This module integrates a Spatial Attention Submodule (SAM) and a Channel Attention Submodule (CAM) in parallel and performs adaptive integration through a Dynamic Weighted Fusion Mechanism (DWFM).



(a) Dynamic Dual Attention Module.

(b) NearNet-A and NearNet-B modules.

Figure 2: Separate descriptions for the DDAM module and the NearNet module.

SAM. The purpose is to highlight densely populated areas in the image. Given the input feature map X, we first obtain X_{max} and X_{avg}. After concatenating these, the CBR(Convolution-BatchNorm-ReLU) is applied to construct the spatial attention weight map G_{sp}. The final spatial enhanced feature is X_{sp} obtained by element-wise multiplication of the original features with the attention weight map:

$$G_{\mathrm{sp}} = \sigma\left(\mathrm{Conv}_{7\times7}\left(\mathrm{Concat}(X_{\mathrm{max}}, X_{\mathrm{avg}})\right)\right) \tag{1}$$

$$\mathrm{X}_{\mathrm{sp}} = \mathrm{X} \odot \mathrm{G}_{\mathrm{sp}} \tag{2}$$

Where σ represents the sigmoid function.

CAM. This module models the dependencies between channels and filters out feature channels that are more discriminative for counting tasks. It performs global average pooling on the input feature map X to obtain channel statistics vectors z. Subsequently, a bottleneck structure containing two fully connected layers, generates channel attention weight vectors $G_{\mathrm{ch}} \in \mathbb{R}^{1\times1\times C}$:

$$G_{\mathrm{ch}} = \sigma\left(\mathrm{W}_2 \cdot \mathrm{GELU}(\mathrm{W}_1 \cdot z)\right) \tag{3}$$

Where $\mathrm{W}_1 \in \mathbb{R}^{C/r\times C}$, $\mathrm{W}_2 \in \mathbb{R}^{C\times C/r}$ are learnable parameters, and r is the reduction rate. The channel enhancement features X_{ch} are calculated as follows:

$$X_{\mathrm{ch}} = X \odot G_{\mathrm{ch}} \tag{4}$$

After extracting the spatial enhancement features X_{sp} and the channel enhancement features X_{ch}, this paper does not simply add them together. Instead, it introduces a Dynamic Weighted Fusion Mechanism (DWFM) to adaptively balance their contributions.

Specifically, we first X perform global average pooling on the input features to obtain a global context vector. Then, this vector is passed through a fully connected layer with a GELU activation function, which maps it to a scalar weight α in the interval [0,1]. This weight represents the importance of spatial attention in the current input. The final feature fusion is achieved through the weighted sum of α and $1-\alpha$.

$$X_{fusion} = \alpha \cdot X_{sp} + (1-\alpha) \cdot X_{ch} \tag{5}$$

This design allows the module to dynamically decide whether to emphasize spatial location focus or channel feature filtering based on the content of the image's content, thereby achieving a more intelligent and robust feature enhancement.

3.3. Loss Function

We adopt the loss function design of P2PNet, which consists of two components: regression loss $\mathcal{L}_{\mathrm{reg}}$and classification loss $\mathcal{L}_{\mathrm{cls}}$. We determine the optimal binary matching between the predicted points and the ground truth points using the Hungarian matching algorithm.

For the real point set g and the predicted point set p, find the optimal permutation $\hat{\sigma}$ that minimizes the total matching cost:

$$\hat{\sigma} = \arg\min_{\sigma\in\mathfrak{S}_N} \sum_{i=1}^{N} \left[\lambda_{\mathrm{cls}}\mathcal{L}_{\mathrm{cls}}\left(p_{\sigma(i)}, g_i\right) + \lambda_{\mathrm{reg}}\mathcal{L}_{\mathrm{reg}}\left(p_{\sigma(i)}, g_i\right)\right] \tag{6}$$

Where $\mathcal{L}_{\mathrm{cls}}$ is the focus loss, $\mathcal{L}_{\mathrm{reg}}$is the $L1$ loss, λ_{cls} and λ_{reg} are the balancing weight.

After determining the optimal matching, the total loss is calculated as follows:

$$\mathcal{L} = \sum_{i=1}^{N} \left[\lambda_{\mathrm{cls}}\mathcal{L}_{\mathrm{cls}}\left(p_{\hat{\sigma}(i)}, g_i\right) + \lambda_{\mathrm{reg}}\mathcal{L}_{\mathrm{reg}}\left(p_{\hat{\sigma}(i)}, g_i\right)\right] \tag{7}$$

4. Experiments

4.1. Dataset

We evaluate the proposed method on three publicly available crowd counting datasets: ShanghaiTech Part A (SHT A), ShanghaiTech Part B (SHT B), and UCF_CC_50. Comprehensive details of these datasets are presented in Table 1.

4.2. Evaluation Indicators

We use the Mean Absolute Error (MAE) and Mean Squared Error (MSE) to measure counting accuracy and robustness. Additionally, to comprehensively evaluate the model's localization and counting performance, we use Normalized Average Precision (nAP) as an evaluation metric for positioning accuracy. The results are reported at different thresholds (δ = 0.05, 0.25, 0.50).

$$MAE = \frac{1}{N}\sum_{i=1}^{N}\left|C_i - C_i^{GT}\right| \tag{8}$$

$$MSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left|C_i - C_i^{GT}\right|^2} \tag{9}$$

4.3. Implementation Details

The model is implemented using PyTorch , starting with an initial learning rate of 1e-4. The batch size is set to 16, and training continues for a total of 1000 epochs.

4.4. Comparison of Counting Performance

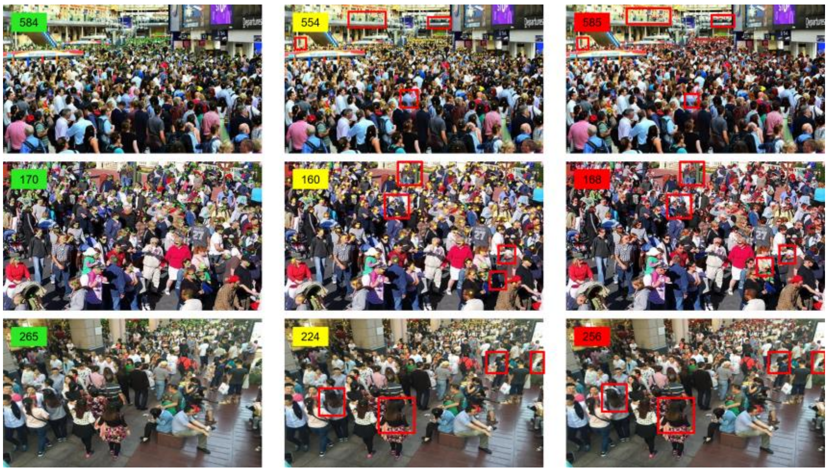


Figure 3: Experimental result annotation map. Green represents the ground truth annotation, yellow represents the P2PNet network's predicted annotation, and red represents the FN-Net network's predicted annotation.

To visually demonstrate the effectiveness of the proposed method, Figure 3 presents the visual counting results of FN-Net on the ShanghaiTech and UCF_CC_50 datasets, compared them with the results of the ground truth and the original P2PNet. Table 2 and Table 3 present the experimental results in comparison with other leading methods.

Table 1: Statistical information of the dataset.

Dataset	**#Numberofimages**	**#Training /Test**	**Total**	**Min**	**Average**	**Max**
SHT_A	482	300/182	241,677	33	501.4	3139
SHT_B	716	400/316	88,488	9	123.6	578
UCF_CC_50	50	-	63974	94	1,280	4543

Table 2: Comparison of state-of-the-art methods on the ShanghaiTech datasets, with the best results highlighted in bold.

method	**way**	**Shanghai Tech Part A**		**Shanghai Tech Part B**	
		MAE	MSE	MAE	MSE
SDANet	AAAI 2020	63.6	101.8	7.8	10.2
GL[5]	CVPR2021	61.3	95.4	7.3	11.7
P2Pnet[1]	CVPR2021	58.8	97.47	7.1	11.3
CCTrans[6]	CVPR2022	64.4	95.4	7.3	11.5
PET[7]	ICCV2023	49.3	78.8	6.2	**9.7**
HSNet[8]	EAAI2024	54.4	89.7	6.8	10.8
Ours		**48.89**	**76.51**	**6. 18**	**9.7**

Table 3: Comparison of state-of-the-art methods on the UCF_CC_50 dataset, with the best results highlighted in bold.

method	**way**	**UCF_CC_50**	
		MAE	MSE
ASNet	CVPR2020	174.8	251.6
P2Pnet[1]	CVPR2021	181.6	249.39
CCTrans[6]	CVPR2022	245.0	343.6
DDC[9]	CVPR 2023	157.12	220.59
M2PLNet[10]	ICME 2024	123.3	**185.14**
EHNet[11]错误!未找到引用源。	ArXiv 2025	136.2	211.37
Ours		**112.7**	209.33

4.5. Ablation Experiment

To verify the effectiveness of each module proposed in this paper, we conducted ablation experiments on the ShanghaiTech Part A dataset. The results are presented in Table 4.

Table 4: Ablation experiment of DDAM on the ShanghaiTech Part A dataset.

method	**MAE**	**MSE**
P2PNet	58.8	97.47
P2PNet+DDAM	56.5	95.0
P2PNet+FN-Net	57.2	96.4
P2PNet+DDAM+ FN-Net	56.35	94.8

In addition, we further evaluated the contribution of FN-Net to localization performance using the nAP metric on the UCF_CC_50 dataset. The results are shown in Table 5.

Table 5: Comparison of localization performance (nAP) of FN-Net on the UCF_CC_50 dataset.

nAP	UCF_CC_50	
	P2PNet	NF-Net
δ=0.50	39.87%	50.3%

δ=0.25	17.68%	20.1%
δ=0.05	1.29%	0.94%

The ablation experiment results demonstrate that each module proposed in this paper effectively contributes to performance improvement. Specifically, DDAM, through the dynamic fusion of spatial and channel attention, enables the model to intelligently focus on discriminative regions and features, which is key to improving counting accuracy. Meanwhile, FN-Net, by explicitly modeling scale changes caused by perspective, provides a more adaptive multi-scale features for the point localization heads, thus achieving significant progress in nAP. These results collectively demonstrate that our method, through the synergistic effect of Feature Augmentation (DDAM) and Multi-Scale Structural Design (FN-Net), fundamentally enhances the baseline model's ability to address the core challenges of dense crowds.

5. Conclusion

This paper systematically analyzes the limitations of the P2PNet crowd counting and localization framework in feature representation and multi-scale modeling, and proposes two key improvements: the Dual Attention Module (DDAM) and the Far-Near Adaptive Network (FN-Net). FN-Net addresses perspective scale variations, while DDAM enhances feature discriminability through dynamic attention fusion. Experimental results demonstrate that this method consistently improves counting and localization performance across multiple datasets.

Acknowledgements

Supported by the Henan Province Science and Technology Research Project (232102211037), Open Research Laboratory Project of Henan University(20252304056).

References

[1] Song, Q., Wang, C., Wang, Y., Zhang, Y. and Zhang, C. (2021) Rethinking Counting and Localization in Crowds: A Purely Point-Based Framework. Proceedings of the IEEE/CVF International Conference on Computer Vision, 1, 1-10.

[2] Zhang, Y., Zhou, D., Chen, S., Gao, S. and Ma, Y. (2016) Single-Image Crowd Counting via Multi-Column Convolutional Neural Network. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1, 1-9.

[3] Li, Y., Zhang, X. and Chen, D. (2018) CSRNet: Dilated Convolutional Neural Networks for Understanding the Highly Congested Scenes. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1, 1-9.

[4] Liang, D., Chen, X., Xu, W., Zhou, Y. and Bai, X. (2022) An End-to-End Transformer Model for Crowd Localization. European Conference on Computer Vision, 1, 1-16.

[5] Liu, W., Salzmann, M. and Fua, P. (2021) GL: A Generic Framework for Learning Object Counting from Point Supervision. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 1, 1-10.

[6] Liang, D., Xu, W., Bai, X. and Zhou, Y. (2022) CCTrans: Simplifying and Improving Crowd Counting with Transformer. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 1, 1-10.

[7] Ma, Z., Wei, X., Hong, X. and Gong, Y. (2023) PET: A Purely Point-Based End-to-End Transformer for Crowd Counting. Proceedings of the IEEE/CVF International Conference on Computer Vision, 1, 1-10.

[8] Zhang, L., Li, H., Chen, X. and Wang, Y. (2024) HSNet: Hierarchical Scale Network for Crowd Counting. Engineering Applications of Artificial Intelligence, 128, 107589.

[9] Wang, Y., Zhang, J., Zhang, Y., Wang, L. and Wang, C. (2023) DDC: A Dual-branch Dilated Convolutional Network for Crowd Counting. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 1, 1-10.

[10] Wang, Q., Li, J., Zhang, Y., Zhou, Y. and Yang, J. (2024) M2PLNet: Multi-scale Multi-patch Learning Network for Crowd Counting. IEEE International Conference on Multimedia and Expo, 1, 1-6.

[11] Wang, Z., Liu, Y., Chen, Z., Li, M. and Zhang, T. (2025) EHNet: Enhancing Crowd Counting with A Scale-Aware Hierarchical Network. arXiv preprint, arXiv:2503.12061, 1-10.