EHNet: An Efficient Hybrid Network for Crowd Counting and Localization

Yuqing Yan[1][0009-0004-7731-0610] and Yirui Wu[✉][0000-0003-3022-3718]

Hohai University, Nanjing Jiangsu 211100, China
2206050101@hhu.edu.cn

Abstract. In recent years, crowd counting and localization have become crucial techniques in computer vision, with applications spanning various domains. The presence of multi-scale crowd distributions within a single image remains a fundamental challenge in crowd counting tasks. To address these challenges, we introduce the Efficient Hybrid Network (EHNet), a novel framework for efficient crowd counting and localization. By reformulating crowd counting into a point regression framework, EHNet leverages the Spatial-Position Attention Module (SPAM) to capture comprehensive spatial contexts and long-range dependencies. Additionally, we develop an Adaptive Feature Aggregation Module (AFAM) to effectively fuse and harmonize multi-scale feature representations. Building upon these, we introduce the Multi-Scale Attentive Decoder (MSAD). Experimental results on four benchmark datasets demonstrate that EHNet achieves competitive performance with reduced computational overhead, outperforming existing methods on ShanghaiTech Part_A, ShanghaiTech Part_B, UCF-CC-50, and UCF-QNRF. Our code is in https://github.com/Kadina17/EHNet.

Keywords: Crowd counting, Crowd localization, Efficient Hybrid Networks.

1   Introduction

Crowd counting, aimed at accurately quantifying large gatherings of individuals, has emerged as a critical research area at the intersection of artificial intelligence and computer vision. In recent years, the applications of crowd counting technology have expanded significantly, encompassing diverse domains such as public safety and urban planning [1], traffic management [2], video surveillance [3], and smart city initiatives [4]. While previous research has established fundamental counting capabilities, significant challenges persist, particularly in scenarios involving high crowd density and high occlusion. Specifically, substantial scale variations arise from two key factors: perspective distortion causes nearby individuals to appear larger than those in distant areas, and mutual occlusion between people creates abrupt scale discontinuities when partially visible body parts are misinterpreted as complete human instances. These phenomena are particularly pronounced in densely packed crowds where overlapping occurs frequently. The combined effects of density-induced perspective changes and occlusion-generated scale ambiguity demand methods with both strong long-range feature

capturing ability to understand global spatial relationships and fine-grained local information extraction capabilities to discern subtle visible cues. Current approaches struggle to simultaneously address three critical requirements: (a) multi-scale perception to handle dramatic intra-scene size differences, (b) occlusion-robust feature learning for partially visible targets, and (c) dynamic adaptation to varying density distributions across different image regions.
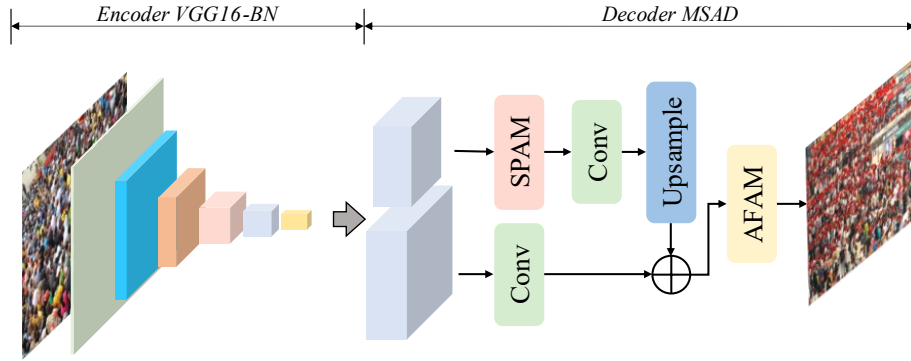

Fig. 1. Overall architecture diagram of our model.

Neural network architectures for crowd counting can be broadly categorized into two main approaches: Convolutional Neural Network (CNN) based approaches [5, 6] and Transformer-based approaches [1, 7]. CNN-based approaches typically take images or video frames as input and generate either crowd density maps or direct count estimates. Notable examples in this category include MCNN [8], CSRNet [9], and FGENet [2]. Although CNNs excel at capturing fine-grained local patterns and hierarchical spatial features, their inherent architectural constraints in modeling long-range dependencies significantly hinder their capability to extract discriminative features for large-scale individuals.

In contrast, Transformer-based approaches, inspired by Vision Transformer (ViT) [7], offer an alternative to CNNs in crowd counting by processing image patches to produce density maps or count estimates. While Transformer-based models such as CLTR [1] and LoViTCrowd [10] demonstrate superior capability in modeling long-range spatial dependencies - specifically, the ability to establish contextual relationships between distant crowd regions (e.g., connecting head features in the foreground with corresponding body patterns in the background across 100+ pixels) - they are fundamentally constrained by two critical limitations: First, their patch-based processing leads to inefficient extraction of fine-grained local features crucial for distinguishing partially occluded limbs or small-scale textures in dense crowds. Second, the self-attention mechanism's quadratic computational complexity ($O(n^2)$ for n image patches) creates substantial overhead when processing high-resolution crowd images containing thousands of potential targets. These constraints become particularly problematic in real-world crowd counting where global dependencies refer to three essential capabilities: Understanding perspective-consistent size relationships between near and far

crowd clusters, Maintaining coherent density estimations across occluded areas through cross-region attention, and Adapting to gradual scale transitions from image center to edges in wide-angle surveillance scenarios.

To address this challenge, we propose an Efficient Hybrid Network (EHNet) that achieves global dependency modeling with linear computational complexity while preserving the fine-grained feature extraction capabilities of CNNs. The detailed structure of EHNet is illustrated in **Fig. 1**. This framework combines global contextual modeling with fine-grained feature extraction, specifically designed to address the challenges of multi-scale variations in crowd scenes. Specifically, our approach leverages a Spatial-Position Attention Module (SPAM) attention mechanism to extract comprehensive global dependency, while incorporating fine-grained details through an Adaptive Feature Aggregation Module (AFAM).

Our contributions can be summarized as follows:

— We propose an end-to-end hybrid network named EHNet. EHNet demonstrates superior performance in both accurate crowd counting and precise localization of individuals.
— We propose three innovative modules: the SPAM attention mechanism, which effectively captures global dependency while preserving spatial features; the AFAM module, which enhances the network's capacity to extract and process fine-grained information for improved counting accuracy; and the MSAD, which builds upon the SPAM and AFAM components to further enhance the model's performance.
— We demonstrate the effectiveness of our approach through extensive experiments.

2 Methodology

2.1 Point framework

In point framework, the model directly utilizes the original annotation points in the image as prediction targets, where the annotation points precisely indicate the actual position of each individual. Suppose there are N individuals in an image, the position of the i-th individual is represented by the coordinate pair (x_i, y_i), where $i \in \{1, 2, \cdots, N\}$, with x_i as the horizontal coordinate and y_i as the vertical coordinate.

During the inference process, given an input image, the model first generates a series of anchor points through a feature extraction network, with a stride of 2 pixels. Then, a regression head is used to predict the offset of each anchor point relative to the true individual's position, as well as the corresponding confidence score. These offsets and confidence scores reflect the model's prediction of the position of the individual associated with each anchor point. To ensure prediction reliability, we filter out low-confidence predictions using a threshold τ. Suppose the predicted number of individuals after thresholding is M, the predicted point coordinates are represented by $(\hat{x}_j, \hat{y}_j)$, where $j \in \{1, 2, \cdots, M\}$, denoting the position of the j-th predicted point.

During training, the model needs to match the annotation points with the predicted points to calculate the position loss. Ideally, each predicted point $(\hat{x}_j, \hat{y}_j)$ should be as close as possible to the corresponding annotation point (x_i, y_i), meaning the model

should accurately predict and align with the annotated point's location. This matching process can be formalized as a bipartite matching problem, where one set of nodes consists of the annotation points and the other set consists of the predicted points. The Hungarian algorithm is used to optimize the matching between the predicted and annotated points, ensuring that the cost of matching, typically represented as the distance between the annotation and predicted points, is minimized.
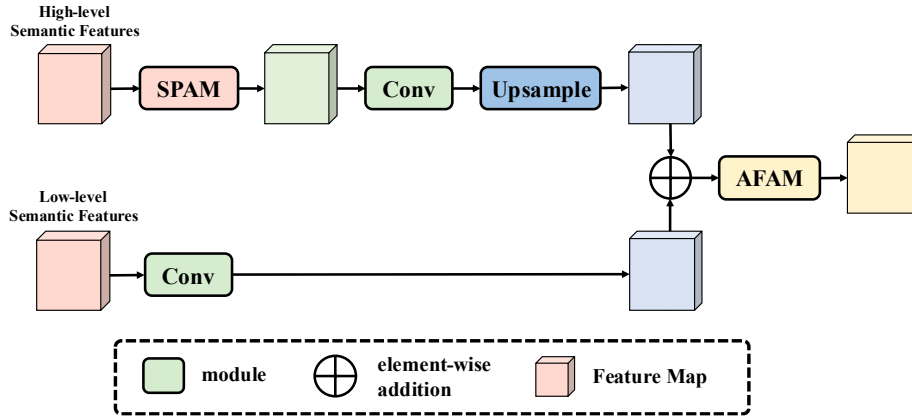

Fig. 2. The architecture of MSAD.

2.2 Efficient hybrid network

This model design adopts an encoder-decoder architecture, primarily aimed at crowd counting and localization tasks. In this architecture, the model first uses a pre-trained VGG16-bn [11] as the encoder to extract multi-scale features from the input image.

The extracted multi-scale features are then fed into a module called the Multi-Scale Attentive Decoder. The purpose of this decoder is to progressively decode the feature maps, further refining and enhancing the fine-grained information of individuals in the crowd. The output of the decoding process mainly consists of two components: offset values and confidence scores. The offset values are used to predict the displacement of each anchor point relative to the center of the actual individual, while the confidence scores indicate the likelihood that the anchor point has detected an individual. By combining these two outputs, the model can not only effectively estimate the total number of individuals in the image but also accurately locate each individual in dense crowd scenes.

2.3 Multi-Scale Attentive Decoder

Our model decoder consists of two main modules: Spatial-Position Attention Module (SPAM) and Adaptive Feature Aggregation Module (AFAM), shown in **Fig. 2**.
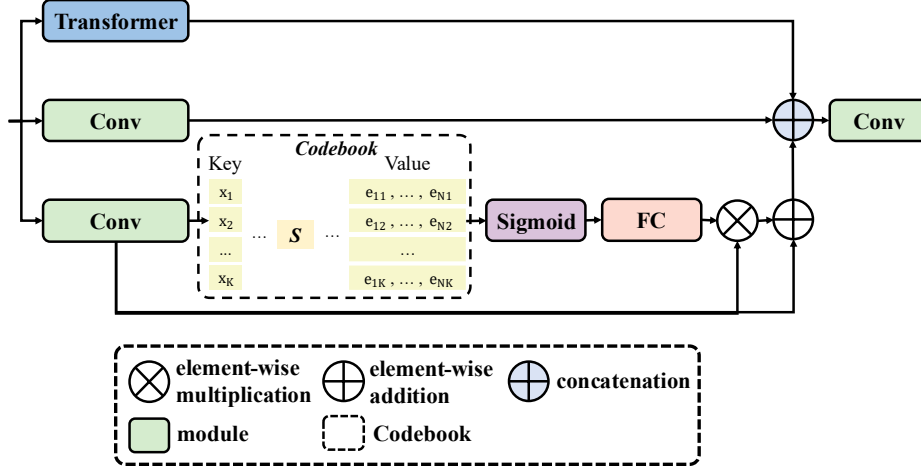
Fig. 3. The architecture of SPAM.

Spatial-Position Attention Module (SPAM). SPAM is a novel three-branch architecture that synergizes Transformer Block, convolution operations, and Cipherbook Module, as illustrated in **Fig. 3**. This design enables comprehensive multi-scale feature learning, effectively capturing both spatial dependencies and position-sensitive patterns in dense crowd scenes.

Each branch in SPAM serves a distinct yet complementary purpose: the Transformer Block models long-range dependencies for global context understanding, the convolution branch processes local spatial features with position sensitivity, and the Cipherbook Module captures explicit spatial patterns through a learned dictionary of feature representations. Through this synergistic combination, SPAM enhances the model's ability to accurately count and locate individuals in diverse crowd scenarios.

The three-branch architecture operates simultaneously on different aspects of the input features, enabling SPAM to process global context, local details, and spatial structures in parallel. This parallel processing strategy not only improves computational efficiency but also provides a more comprehensive feature representation, where each branch's unique strengths contribute to the overall performance in crowd counting and localization tasks.

Adaptive Feature Aggregation Module (AFAM). AFAM is an efficient feature aggregation network that combines channel-wise attention with multi-scale feature learning, as shown in **Fig. 4**. The module integrates parallel convolutions with RepConv [12] for feature extraction, while employing both channel and spatial attention mechanisms to dynamically emphasize informative features. This dual-attention design, coupled with the inference-optimized RepConv structure, enables AFAM to achieve both computational efficiency and effective feature representation.
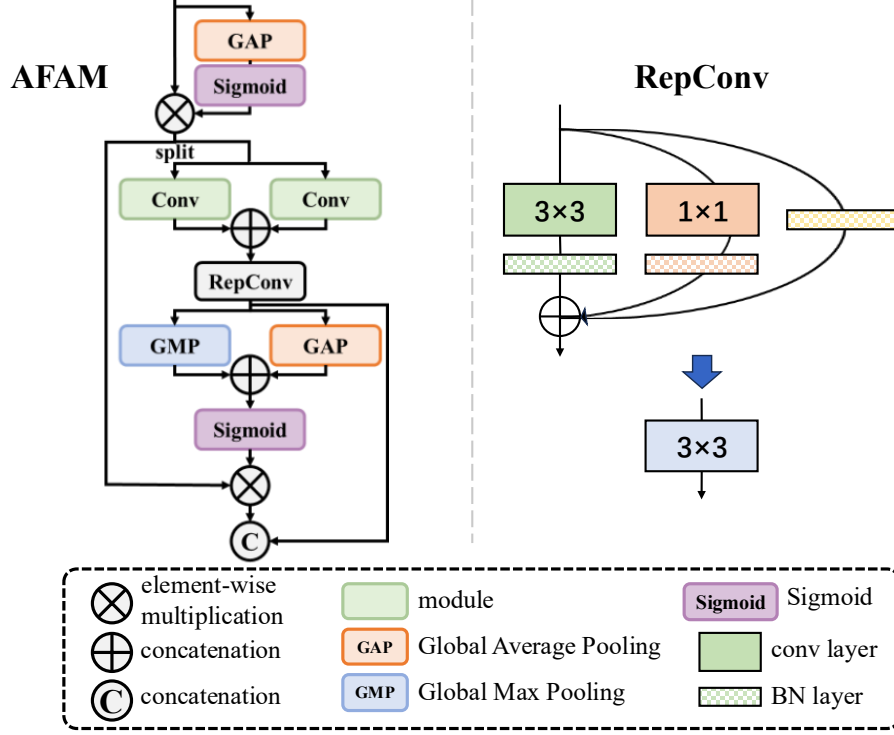
Fig. 4. The architecture of AFAM.

The module first compresses spatial information to obtain channel descriptors, which are transformed into weights through a Sigmoid function to enhance important information across different channels.

The weighted features are then split into two paths. The first path processes features through two parallel convolution layers, whose outputs are concatenated and fed into a RepConv module. As illustrated in **Fig. 4**, RepConv consists of parallel 3×3 and 1×1 convolutions with BN layers, which can be fused into a single 3×3 convolution during inference. The second path applies both Global Max Pooling (GMP) and Global Average Pooling (GAP) operations, combines their outputs through addition, and generates spatial attention weights via a Sigmoid function. These weights are applied to the features through element-wise multiplication.

Finally, the module concatenates the outputs from both paths to produce the final aggregated features, effectively combining multi-scale spatial information with attention-enhanced representations.

Table 1. Comparison of different approaches on four datasets.

Approaches	Venue	ShanghaiTech				UCF_CC_50		UCF-QNRF	
		Part A		Part B					
		MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE
MCNN [8]	CVPR'16	110.2	-	26.4	-	377.6	-	-	-
CAN [13]	CVPR'19	62.8	101.8	7.7	12.7	212.2	243.7	107.0	183.0
ASNet [14]	CVPR'20	57.78	90.13	-	-	174.84	251.63	91.59	159.71
SUA [15]	ICCV'21	68.5	121.9	14.1	20.6	-	-	130.3	226.3
GauNet [16]	CVPR'22	54.8	89.1	<u>6.2</u>	9.9	186.3	256.5	81.6	153.7
HDNet [17]	ICME'22	53.4	89.9	-	-	-	-	83.2	148.3
CLTR [1]	ECCV'23	56.9	95.2	6.5	10.6	-	-	85.8	141.3
DDC [18]	CVPR'23	52.87	85.62	6.08	<u>9.61</u>	157.12	220.59	**65.79**	**126.53**
CHS-Net [19]	ICASSP'23	59.2	97.8	7.1	12.1	-	-	83.4	144.9
FGENet [2]	MMM'23	<u>51.66</u>	<u>85.0</u>	6.34	10.53	<u>142.56</u>	<u>215.87</u>	82.1	143.76
STEERER [20]	ICCV'23	54.5	86.9	**5.8**	**8.5**	-	-	79.5	144.3
VMambaCC [21]	ArXiv'24	51.87	**81.3**	7.48	12.47	-	-	88.42	144.73
Gramformer [22]	AAAI'24	54.7	87.1	-	-	-	-	<u>76.7</u>	<u>129.5</u>
M2PLNet [23]	ICME'24	**50.86**	89.86	-	-	**123.3**	**185.14**	-	-
Ours		52.29	81.01	6.64	10.46	136.2	211.37	78.25	132.63

3   Experiments

3.1   Datasets and implementation details

Datasets. To comprehensively evaluate our model's performance, we utilized four representative crowd counting datasets: ShanghaiTech Part_A, ShanghaiTech Part_B, UCF_CC_50, and UCF-QNRF. These datasets cover a wide range of crowd densities and scenarios, providing a robust testing framework. ShanghaiTech Part_A contains 482 high-density images (average 501 individuals per image), while Part_B includes 716 lower-density street scenes (average 123 individuals). UCF_CC_50, despite having only 50 images, presents extreme variations in crowd counts (94-4,543 individuals). UCF-QNRF, a large-scale dataset, comprises 1,535 high-resolution images with over 1.3 million annotated individuals across diverse settings.

Implementation details. The experiments were conducted on a computing platform equipped with an RTX 4060 TI GPU with 16GB of memory. To enhance the model's generalization ability and robustness, we implemented data augmentation techniques, including random flipping and random cropping. For optimization, we employed the Adam optimizer with an initial learning rate of 1e-4.

Table 2. Efficiency of different approaches.

Approaches	Parameters		Computation		Memory Usage		Self Time	
	Total	Trainable	FLOPs	Average Inference Time	GPU	CPU	CPU	CUDA
MCNN [8]	133,705	133,705	2.692G	1.88ms	33.59 MB	1994.00 MB	11.692ms	11.777ms
CSRNet [9]	16,263,489	16,263,489	41.445G	5.53ms	74.12 MB	1996.07 MB	6.089ms	8.128ms
CAN [13]	16,530,113	16,530,113	41.868G	6.61ms	82.57 MB	2333.90 MB	12.297ms	12.306ms
ASNet [14]	30,398,087	30,398,087	62.745G	9.34ms	0.57 MB	2294.65 MB	28.600ms	28.623ms
CLTR [1]	-	-	-	-	OOM	-	-	-
FGENet [2]	183,271,139	183,271,139	64.647G	22.14ms	280.14 MB	2413.40 MB	74.955ms	74.964ms
STEERER [20]	39,296	39,296	3.969G	1.34ms	61.83 MB	2334.20 MB	4.787ms	5.466ms
Gramformer [22]	-	-	60.9G	12.6ms	0.57 MB	2286.12 MB	-	-
M2PLNet [23]	508,460,865	508,460,865	234.103G	59.99ms	369.08 MB	2420.55 MB	178.544ms	178.531ms
Ours	**2,515,008**	**2,515,008**	**133.871G**	**11.08ms**	**161.57 MB**	**2287.59 MB**	**.314ms**	**6.510ms**

3.2　Experiment

Counting experiment. Experimental results in Table. 1 demonstrate our method's strong performance across four benchmark datasets. On ShanghaiTech Part_A, we achieve state-of-the-art results with MAE/MSE of 50.29/81.01, outperforming recent approaches like M2PLNet and VMambaCC. While STEERER leads on ShanghaiTech Part_B, our method shows competitive performance with MAE/MSE of 6.64/10.46. For UCF_CC_50, we achieve the second-best results (136.2/211.37) after FGENet. On UCF-QNRF, our approach maintains competitive performance (78.25/132.63) compared to leading methods DDC and Gramformer. These results demonstrate our method's strong generalization ability across diverse crowd counting scenarios.

Efficiency experiment. Our approach demonstrates significant advantages across multiple dimensions, shown in Table. 2. The parameters are substantially lower than that of other approaches, indicating a more lightweight model. This is highly beneficial for deployment in resource-constrained environments, while also enhancing efficiency during both training and inference processes. Although our FLOPs is relatively higher than other approaches, the model achieves outstanding performance through reduced inference time and highly efficient resource utilization, suggesting superior computational efficiency. Memory usage on both GPU and CPU remains within a reasonable range, making the model well-suited for practical engineering deployment. While the self CUDA time is higher compared to some lightweight approaches, such as STEERER, the self CPU time demonstrates significantly greater computational acceleration efficiency compared to approaches like ASNet and CAN. Overall, our approach exhibits notable advantages in terms of parameters, memory usage, and self CPU time. It is a lightweight, high-performance, and computationally efficient crowd counting approach, making it particularly suitable for real-time applications and deployment scenarios.

Table 3. Localization experimental results on SHT_A

Approaches	P(%)	R(%)	F(%)
LCFCN [24]	43.30	26.00	32.50
Method in [25]	34.90	20.70	25.90
LSC-CNN [26]	33.40	31.90	32.60
TopoCount [27]	41.70	40.60	41.10
CLTR [1]	43.60	42.70	43.20
P2PNet [3]	46.00	41.50	39.80
Ours	**52.78**	**51.53**	**52.13**

Table 4. Model ablation experiment on SHT_A

Module Name	MAE	MSE
BaseLine	54.02	82.59
+SPAM	51.28	81.93
+AFAM	50.97	82.64
Ours	**50.29**	**81.01**

Localization experiment. Experimental results demonstrate that our approach achieves state-of-the-art performance in crowd localization on the SHT_A dataset, as shown in Table. 3. Using Hungarian algorithm for point matching with a distance threshold of 4 pixels to determine positive and negative samples, our method achieves 52.78% precision, 51.53% recall, and 52.13% F1 score. This represents a significant improvement of approximately 9 percentage points across all metrics compared to the previous best method P2PNet. The balanced improvement in both precision and recall metrics demonstrates our method's robust performance in accurate crowd localization.
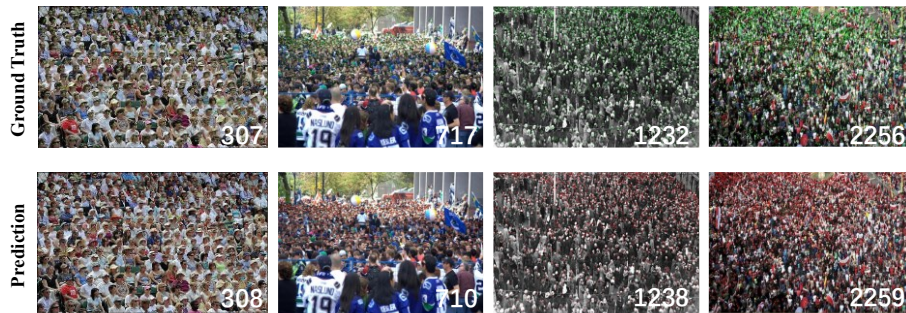

Fig. 5. The predicted results of our EHNet. The white numbers denote to the ground truth or prediction number.

3.3 Ablation experiment

The ablation study of model demonstrates that the complete model configuration (Ours) achieved the best performance across both MAE and MSE, shown in Table. 4. Adding the SPAM module reduced the MAE to 51.28, a decrease of 2.74, and slightly improved the MSE to 81.93. Incorporating the AFAM module further reduced the MAE to 50.97, showing a 3.05 improvement compared to the baseline, while the MSE increased slightly to 82.64. Combining all modules together achieved the lowest MAE of 50.29 and the lowest MSE of 81.01. Compared to the baseline, the MAE decreased by approximately 3.73, and the MSE improved by 1.58. The superior performance of ours highlights that our full model configuration achieves the most accurate and robust results, outperforming the baseline and other partial configurations.

Table 5. AFAM ablation experiment on SHT_A

Module Name	MAE	MSE
Conv	55.46	92.27
ELAN [28]	54.91	86.36
CSPLayer [29]	53.78	87.48
AFAM	**50.29**	**81.01**

The AFAM module achieved the best performance on both metrics, with an MAE of 50.29 and an MSE of 81.01, shown in Table. 5. Compared to ELAN and CSPLayer, AFAM reduced the MAE by approximately 4.62 and 3.49 respectively, and decreased the MSE by 5.35 and 6.47 respectively. These results demonstrate that the AFAM module has significant advantages in improving counting accuracy and reducing errors, confirming its effectiveness in crowd counting tasks.

Table 6. SPAM ablation experiment on SHT_A

Exp. No.	Module Name	Configuration	MAE	MSE
1	Conv	Baseline	54.02	82.59
2	+Transformer	Conv + Transformer	54.00	84.32
3	+CodeBook	Conv + CodeBook	52.46	81.58
4	**+Transformer & CodeBook**	**Conv + Transformer + CodeBook**	**50.29**	**81.01**

This ablation study compared the performance of different module combinations, shown in Table. 6. The results show that the complete model configuration (Conv + Transformer + CodeBook) achieved the best performance in both MAE and MSE, with values of 50.29 and 81.01 respectively. Compared to the baseline model, MAE decreased by approximately 3.73, while MSE reduced by 1.58. Adding either the Transformer or CodeBook module alone led to performance improvements, but the combination of both yielded the best results. This indicates that the Transformer and CodeBook modules have a synergistic effect in improving crowd counting accuracy, confirming the effectiveness of model design.

4   Conclusion

In this work, we propose a novel end-to-end hybrid architecture. To address the multi-scale issue caused by image perspective effects, we introduce SPAM, AFAM and MSAD. Through extensive experiments, we demonstrate the effectiveness of our approach in crowd localization and counting, achieving impressive results across multiple datasets.

References

1. Dingkang Liang, Wei Xu, and Xiang Bai, "An End-to-End Transformer Model for Crowd Localization," in ECCV, 2022.
2. Hao-Yuan Ma, Li Zhang, and Xiang-Yi Wei, "Fgenet: Fine-grained extraction network for congested crowd counting," in MMM, 2024.
3. Qingyu Song, Changan Wang, Zhengkai Jiang, et al., " Rethinking Counting and Localization in Crowds:A Purely Point-Based Framework," in ICCV, 2021.
4. Ao Luo, Fan Yang, Xin Li, Dong Nie, Zhicheng Jiao, Shangchen Zhou, and Hong Cheng, " Hybrid graph neural networks for crowd counting," in AAAI, 2020.
5. Feng Dai, Hao Liu, Yike Ma, Xi Zhang, and Qiang Zhao, "Dense scale network for crowd counting," in ICMR, 2021.
6. Rahul Rama Varior, Bing Shuai, Joe Tighe, and Davide Modolo, "Scale-aware attention network for crowd counting," in CoRR, 2019.
7. Alexander Kolesnikov, Alexey Dosovitskiy, Dirk Weissenborn, Georg Heigold, Jakob Uszkoreit, Lucas Beyer, Matthias Minderer, Mostafa Dehghani, Neil Houlsby, Sylvain Gelly, Thomas Unterthiner, and Xiaohua Zhai, "An image is worth 16x16 words: Transformers for image recognition at scale," in ICLR, 2021.
8. Yingying Zhang, Desen Zhou, Siqin Chen, Shenghua Gao, and Yi Ma, "Single-Image Crowd Counting via Multi-Column Convolutional Neural Network," in CVPR, 2016.
9. Yuhong Li, Xiaofan Zhang, and Deming Chen, "CSRNet: Dilated Convolutional Neural Networks for Understanding the Highly Congested Scenes," in CVPR, 2018.
10. Nguyen Hoang Tran, Ta Duc Huy, Soan Thi Minh Duong, et al., "Improving local features with relevant spatial information by vision transformer for crowd counting," in BMVC, 2022.
11. Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," in ICLR, 2015.
12. Xiaohan Ding, Xiangyu Zhang, Ningning Ma, Jungong Han, Guiguang Ding, and Jian Sun, "Repvgg: Making vgg-style convnets great again," in CVPR, 2021.
13. Weizhe Liu, Mathieu Salzmann, and Pascal Fua, "Context-Aware Crowd Counting," in CVPR, 2019.
14. Xiaoheng Jiang, Li Zhang, Mingliang Xu, Tianzhu Zhang, Pei Lv, Bing Zhou, Xin Yang, and Yanwei Pang, "Attention Scaling for Crowd Counting," in CVPR, 2020.
15. Yanda Meng, Hongrun Zhang, Yitian Zhao, Xiaoyun Yang, Xuesheng Qian, Xiaowei Huang, and Yalin Zheng, "Spatial Uncertainty-Aware Semi-Supervised Crowd Counting," in ICCV, 2021.
16. Zhi-Qi Cheng, Qi Dai, Hong Li, Jingkuan Song, Xiao Wu, and Alexander G Hauptmann, "Rethinking Spatial Invariance of Convolutional Networks for Object Counting," in CVPR, 2022.

17. Chenliang Gu, Changan Wang, Bin-Bin Gao, Jun Liu, and Tianliang Zhang, "Hdnet: A hierarchically decoupled network for crowd counting," in IEEE International Conference on Multimedia and Expo, ICME 2022, Taipei, Taiwan, July 18-22, 2022. 2022, pp. 1--6, IEEE.

18. Yasiru Ranasinghe, Nithin Gopalakrishnan Nair, Wele Gedara Chaminda Bandara, and Vishal M Patel, "Diffuse-Denoise-Count: Accurate Crowd-Counting with Diffusion Models," 2023.

19. Mingliang Dai, Zhizhong Huang, Jiaqi Gao, Hongming Shan, and Junping Zhang, "Cross-Head Supervision for Crowd Counting with Noisy Annotations," in ICASSP, 2023.

20. Tao Han, Lei Bai, Lingbo Liu, and Wanli Ouyang, "STEERER: resolving scale variations for counting and localization via selective inheritance learning," in ICCV, 2023.

21. Hao-Yuan Ma, Li Zhang, and Shuai Shi, "Vmambacc: A visual state space model for crowd counting," in CoRR, 2024.

22. Hui Lin, Zhiheng Ma, Xiaopeng Hong, Qinnan Shangguan, and Deyu Meng, "Gramformer: Learning crowd counting via graph-modulated transformer," in AAAI, 2024.

23. Hao-Yuan Ma and Li Zhang, "Multi-head multi-scale pixel localization network for crowd counting with highly dense and small-scale samples," in ICME, 2024.

24. Issam H. Laradji, Negar Rostamzadeh, Pedro O. Pinheiro, David V\'azquez, and Mark Schmidt, "Where are the blobs: Counting by localization with point supervision," in ECCV, 2018.

25. Javier Ribera, David Guera, Yuhao Chen, and Edward J. Delp, "Locating objects without bounding boxes," in CVPR, 2019.

26. Deepak Babu Sam, Skand Vishwanath Peri, Mukuntha Narayanan Sundararaman, Amogh Kamath, and R. Venkatesh Babu, "Locate, size, and count: Accurately resolving people in dense crowds via detection," in TPAMI, 2021.

27. Shahira Abousamra, Minh Hoai, Dimitris Samaras, and Chao Chen, "Localization in the crowd with topological constraints," in AAAI, 2021.

28. Xindong Zhang, Hui Zeng, Shi Guo, and Lei Zhang, "Efficient long-range attention network for image super-resolution," in ECCV, 2022.

29. Chien-Yao Wang, Hong-Yuan Mark Liao, Yueh-Hua Wu, Ping-Yang Chen, Jun-Wei Hsieh, and I-Hau Yeh, "CSPNet: A New Backbone that can Enhance Learning Capability of CNN," in CVPR, 2020.