ISSN 2616-5775 Vol. 8, Issue 11: 14-22, DOI: 10.25236/AJCIS.2025.081102

MambaVision-Count: A Crowd Counting System Based on a Hybrid Architecture

Lu Chen[a,*], Lei Ding[b]

School of Electronic Information and Artificial Intelligence, Shaanxi University of Science and Technology, Xi'an, Shaanxi, China
[a]lding@sust.edu.cn, [b]1487115883@qq.com

Abstract: *Counting people in highly crowded, heavily occluded, and scale-varying scenes remains a challenging task. This paper presents a novel crowd counting framework, MambaVision-Count, built upon the efficient visual backbone MambaVision. The framework integrates the strengths of convolution, state-space modeling, and self-attention mechanisms, enabling the model to capture long-range dependencies and global contextual information effectively. This design allows the model to better handle complex variations in crowd distribution. A dual-branch regression head is introduced to simultaneously predict density maps and total counts. Additionally, an EFC feature fusion module is incorporated to enhance the representation of small target regions, thus improving the overall accuracy and robustness of crowd counting. Extensive experiments conducted on datasets such as ShanghaiTech demonstrate that the proposed method outperforms existing state-of-the-art approaches, achieving superior accuracy and inference efficiency. The results highlight its strong practical potential in real-world applications.*

Keywords: *Crowd Counting; Feature Fusion; Transformer; Mamba*

1. Introduction

With the acceleration of urbanization and the increasing frequency of social activities, large-scale crowd gatherings have become more common in public spaces such as subway stations, shopping malls, stadiums, and public events. These scenarios demand real-time crowd monitoring and density estimation to ensure public safety and support intelligent urban management. As a fundamental task in computer vision, crowd counting plays a crucial role in areas such as public safety, urban management, commercial analytics, and disaster prevention. The objective of crowd counting is to estimate the total number of people or generate a pixel-wise density map from input images or videos, thereby enabling precise perception of crowd dynamics.

With the acceleration of urbanization and the increasing frequency of social activities, large-scale crowd gatherings have become more common in public spaces such as subway stations, shopping malls, stadiums, and public events. These scenarios demand real-time crowd monitoring and density estimation to ensure public safety and support intelligent urban management. As a fundamental task in computer vision, crowd counting plays a crucial role in areas such as public safety, urban management, commercial analytics, and disaster prevention. The objective of crowd counting is to estimate the total number of people or generate a pixel-wise density map from input images or videos, thereby enabling precise perception of crowd dynamics.

Traditional crowd counting approaches are predominantly based on Convolutional Neural Networks (CNNs) [1], which can effectively capture local texture information but are limited by a fixed receptive field. Consequently, CNNs struggle to model long-range dependencies and spatial correlations across distant regions, hindering their ability to understand complex spatial distributions. To overcome this limitation, Transformers [2] have recently been introduced into crowd counting. Owing to their self-attention mechanism, Transformers are capable of modeling global dependencies and capturing contextual relationships across the entire image, significantly improving contextual reasoning. However, the computational complexity of the Transformer grows quadratically ($O(n^2)$) with input size, leading to excessive memory and computation costs for high-resolution images, which restricts their practical deployment. MambaVision [3], a newly proposed hybrid backbone, integrates the advantages of CNNs, State Space Models (SSMs) such as Mamba, and Transformer-based self-attention modules. This hybrid structure not only maintains the efficiency and small-object sensitivity of CNNs but also incorporates the linear sequence modeling capability of state-space architectures and the long-range dependency

ISSN 2616-5775 Vol. 8, Issue 11: 14-22, DOI: 10.25236/AJCIS.2025.081102

modeling of Transformers. Owing to its hierarchical and modular design, MambaVision exhibits strong adaptability to different downstream tasks, achieving a good balance between generalization and computational efficiency.

In this paper, we propose an efficient and robust crowd counting model named MambaVision-Count, which is built upon the high-performance hybrid visual backbone MambaVision. The model is designed to simultaneously capture local details and global contextual information. The main contributions of this study are summarized as follows:

- We introduce MambaVision as a unified hybrid backbone that establishes complementary relationships among convolution, state-space modeling, and self-attention. This design allows the network to efficiently extract fine-grained local features while modeling long-range dependencies with low computational cost. Consequently, the model gains superior perception of large-scale crowd distributions and alleviates the challenges posed by heavy occlusion and scale variation.

- To address the insufficient supervision problem in traditional single-branch regression methods, we construct a dual-branch regression head on top of the high-level backbone features. One branch generates pixel-level density maps, providing spatially fine-grained supervision, while the other directly regresses the global crowd count, providing a global quantitative constraint. A consistency loss is further introduced between the two branches to strengthen the training signal and improve robustness.

- We integrate an Efficient Feature Coupling (EFC) module that focuses on enhancing small-object representations through multi-scale convolutional branches and attention mechanisms. This design highlights dense and small-scale regions while preserving global contextual awareness, significantly improving the network's accuracy and stability in highly crowded scenes with severe occlusions.

2. Related Work

2.1. Crowd Counting Methods

Existing crowd counting approaches can be broadly categorized into three types: detection-based methods, density map–based methods, and point-based regression methods.

Detection-based methods identify human heads or bodies with an off-the-shelf detector and simply sum the resulting bounding boxes; they are therefore highly interpretable and provide instance-level localization, yet their accuracy collapses once overlaps and occlusions become frequent, so they are generally restricted to low-density scenes such as sparse streets or squares(e.g., CrowdDet [4]).

Density-map approaches, such as CSRNet [5], train a CNN to convert the whole image into a pixel-wise density surface whose integral equals the crowd count: this strategy gracefully handles medium- and high-density crowds and only needs inexpensive point-level supervision, but the implicit averaging over local neighborhoods sacrifices positional precision, so individual heads cannot be exactly pinpointed.

Point-based or direct-regression models bypass the density surface altogether and either emit a single global count or produce an explicit set of head coordinates extracted from global or local features; under weak or semi-supervised assumptions they can yield more accurate locations, but they typically demand stronger priors or extra supervisory signals and still falter when the crowd becomes extremely dense or heavily occluded.

2.2. Transformer and SSM

The recent success of Transformers in computer vision has spurred interest in applying them to crowd counting tasks. Owing to their self-attention mechanism, Transformers can effectively capture long-range dependencies and global contextual relationships, which are critical for handling occlusion and scale variation. For instance, TransCrowd [6] utilizes the Vision Transformer (ViT) [7] to extract patch-level global representations and subsequently employs a regression network to predict the total count. This approach mitigates the limitations of CNNs in modeling long-range spatial dependencies. Similarly, CCTrans [8] proposes a hybrid CNN–Transformer architecture that combines local convolutional features with global contextual representations, achieving state-of-the-art performance across multiple public datasets. These studies collectively demonstrate that integrating Transformers into crowd counting networks enhances global scene understanding and provides valuable insights for hybrid architectures (e.g., CNN + Transformer or Transformer + SSM).

ISSN 2616-5775 Vol. 8, Issue 11: 14-22, DOI: 10.25236/AJCIS.2025.081102

The Mamba model [9], a recent and efficient variant of state-space models (SSMs), offers linear-time complexity for processing long sequences and has achieved Transformer-level performance in sequence modeling tasks such as language modeling. This capability provides new opportunities for crowd counting. For example, VMambaCC [10] introduces a Visual Mamba architecture into crowd counting, leveraging SSMs to strengthen global contextual reasoning while integrating multi-scale feature pyramids and attention mechanisms. This hybrid design significantly improves robustness in high-density scenarios. However, since Mamba is inherently designed for sequential data, it still faces challenges in spatial feature modeling for visual tasks. Recent research thus focuses on hybrid architectures that combine Mamba with convolutional or Transformer modules. Such designs aim to simultaneously extract local spatial features and model long-range dependencies, effectively addressing small-object loss and inconsistent counting in dense environments. These developments highlight the tremendous potential of SSMs—particularly Mamba—in the crowd counting domain and suggest promising directions for building more efficient and robust hybrid models that balance local precision with global awareness.

3. Method

3.1. Overall Architecture

The proposed MambaVision-Count model adopts a four-stage hierarchical feature extraction backbone with a dual-branch output structure. The goal of this design is to maintain computational efficiency while improving counting accuracy under conditions of heavy occlusion and large-scale variation. The overall architecture of MambaVision-Count is illustrated in Figure 1.

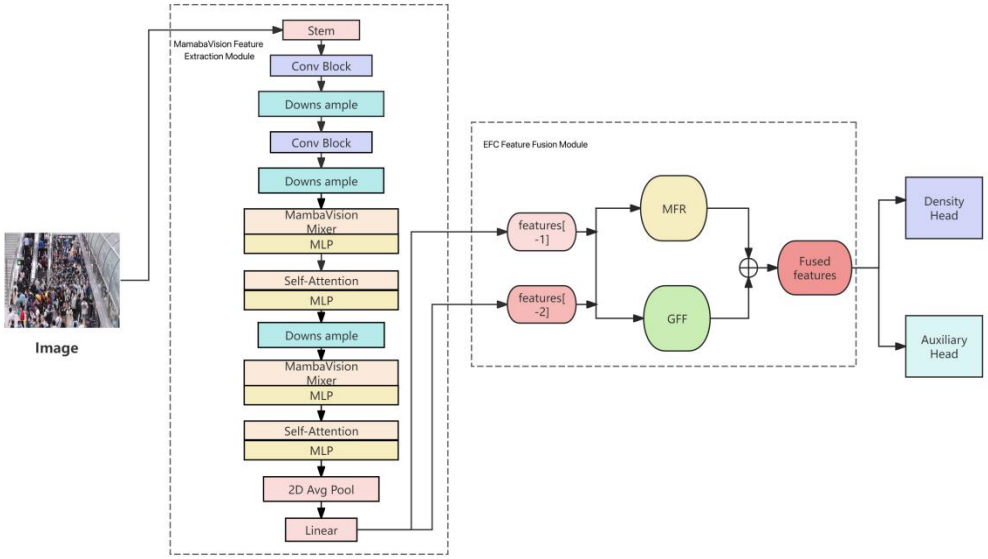

Figure1: MambaVision-Count Architecture Diagram

In the first two stages, traditional convolutional residual blocks are used to extract low-level texture features and local spatial information. The latter two stages incorporate MambaVision Mixers and window-based self-attention modules to strengthen the model's capability for long-range dependency modeling. In the shallow backbone layers, we introduce an Efficient Feature Coupling (EFC) module [11] to enhance feature representations for small objects. After feature extraction, the resulting feature maps are fed into a dual-branch regression head that simultaneously predicts the density map and the total count. A unified loss function enforces consistency between the integrated density and the global count, forming complementary supervision at both the local and global levels. Through this hierarchical and hybrid design, the model combines the fine-grained local perception of convolutional networks with the long-range contextual aggregation of state-space modeling and attention mechanisms, enabling MambaVision-Count to effectively adapt to complex crowd distributions and varying densities.

3.2. Mamba Vision Backbone Network

Our model builds upon MambaVision, an advanced hybrid visual backbone that demonstrates

ISSN 2616-5775 Vol. 8, Issue 11: 14-22, DOI: 10.25236/AJCIS.2025.081102

excellent accuracy and robustness across various crowd counting benchmarks. Its key strength lies in balancing local detail modeling, cross-region dependency modeling, and global contextual representation. This hybrid design alleviates challenges such as severe occlusion and large scale variation, allowing the network to maintain stable performance even in extremely crowded scenarios.

Following the principle of progressively expanding receptive fields, we design different structures across stages. The first two stages adopt lightweight Conv–BN–GELU residual units, where each block contains two 3×3 convolutions to ensure small targets remain identifiable during downsampling and to enhance discriminability in dense areas.

At higher feature levels, a symmetric dual-branch MambaVision Mixer is employed to balance spatial and cross-region dependency modeling. Specifically, the input features are first projected to a lower-dimensional space and then split into two branches. The Selective Scan Branch leverages 1D convolutions and gating mechanisms to efficiently model long-range dependencies across spatial regions. The Symmetric Convolution Branch uses 3×3 convolutions with SiLU activation (functionally similar to a gated sigmoid operation) to reinforce local spatial patterns. The outputs from both branches are concatenated and projected back into a unified feature space, achieving a seamless fusion of sequence-level and spatial-level representations.

To further enhance global modeling while controlling computational complexity, we incorporate window-based multi-head self-attention (W-MSA) in the upper stages. By computing self-attention within local windows and applying a shift-window strategy for inter-window communication, the model maintains strong global perception with high computational efficiency, even under high-resolution inputs.

3.3. EFC Feature Fusion Module

In distant or high-density regions, individuals often appear small and heavily occluded, leading to information loss during downsampling. To mitigate this, we embed the Efficient Feature Coupling (EFC) module in the shallow layers of the backbone, which focuses on enhancing small-object features early in the extraction process.

The EFC module consists of two core components: The Grouped Feature Focus (GFF) unit divides the input channels into groups and models contextual relationships within each group to strengthen inter-layer feature correlation. It highlights the importance of small-object regions while suppressing background noise. And The Multilevel Feature Reconstruction (MFR) module reconstructs and transforms features across multiple scales in the feature pyramid. It balances strong and weak signals across levels, reduces redundant information during fusion, and preserves crucial representations of small targets in deeper layers. Together, GFF and MFR achieve efficient multi-scale feature fusion and small-object enhancement.

The EFC module thus improves feature sensitivity to tiny heads and crowded regions while maintaining model lightness. By integrating EFC into the shallow backbone, downstream decoding stages benefit from richer local cues, mitigating the issue of small-target omission caused by occlusion or extreme scale shrinkage. This design provides a solid representational foundation for accurate crowd estimation in complex environments.

3.4. Dual-Branch Regression Module

Conventional single-branch regression models typically predict either the total count or a single density map, resulting in insufficient supervision signals. Such models lack fine-grained spatial supervision and global quantitative constraints, leading to counting errors and missed detections in dense areas. To overcome these limitations, we design a dual-branch output module that jointly learns pixel-level density estimation and global count regression: The Density Map Branch upsamples the backbone features through two convolutional layers to recover the original image resolution, producing a density map supervised by L1 loss against a Gaussian-convolved ground truth density. Each pixel in the density map represents local crowd density, and integrating over all pixels yields the total count. The Count Regression Branch applies global average pooling (GAP) to the backbone features followed by a fully connected layer to output a scalar representing the overall count. This branch provides strong global supervision during training, encouraging the density map to approximate the correct total count.

This multi-task supervision enables the model to learn the correlation between spatial density and semantic global context, reducing cumulative errors common in single-branch training and enhancing prediction robustness.

ISSN 2616-5775 Vol. 8, Issue 11: 14-22, DOI: 10.25236/AJCIS.2025.081102

3.5. Loss Function and Evaluation Metrics

The total loss function combines the two branches via a weighted sum:

$$\mathscr{L} = \lambda_1 \cdot \mathscr{L}_{\text{density}} + \lambda_2 \cdot \mathscr{L}_{\text{count}} \tag{1}$$

where $L_{density}$ denotes the L_1 loss for the density map branch, and L_{count} represents the MAE loss for the global count branch. The coefficients λ_1 and λ_2 control the relative weights of local and global supervision, respectively.

This multi-task formulation enhances the stability of density estimation and improves adaptability across varying crowd densities.

$$\text{MAE}=\frac{1}{N}\sum_{i=1}^{N}|C_i-M_i| \tag{2}$$

$$\text{MSE}=\sqrt{\frac{1}{N}\sum_{i=1}^{N}(C_i-M_i)^2} \tag{3}$$

where C_i and M_i denote the predicted and ground-truth counts for the i^{th} image, and N is the total number of test samples. MAE reflects the average estimation deviation under real-world conditions, while MSE emphasizes model stability in extreme or high-variance scenarios.

4. Experiments

4.1. Datasets

To evaluate the proposed method, we conduct experiments on four widely used crowd counting datasets: ShanghaiTech [12], UCF-QNRF [13], and JHU-Crowd [14].A detailed summary of these datasets is presented in Table 1.

ShanghaiTech is a large-scale benchmark dataset consisting of two subsets: Part A (SHT_A) and Part B (SHT_B), comprising a total of 1,198 images and 330,165 annotations. The dataset includes a diverse range of indoor and outdoor scenes such as streets, campuses, malls, subways, and public squares, covering densities from sparse to extremely congested. SHT_A mainly contains high-density scenes, making it suitable for evaluating performance under extreme crowding, whereas SHT_B focuses on low- to medium-density scenarios, providing insights into generalization under less crowded conditions.

UCF-QNRF is a challenging large-scale dataset containing 1,535 images with 1,251,642 head annotations. It features diverse scenes—streets, plazas, stadiums, amusement parks, and public gatherings—captured under various lighting and camera angles, ranging from top-down to oblique viewpoints. The dataset exhibits a wide range of densities and scales, serving as a benchmark for evaluating generalization in complex, heterogeneous environments.

JHU-Crowd consists of 4,250 images with 1,114,785 head annotations, averaging approximately 262 annotations per image. It covers diverse scenarios, including streets, malls, transportation hubs, campuses, and indoor public spaces. The dataset provides detailed labels, including point-level, image-level, and head-level annotations, making it applicable not only to density regression but also to small-object detection and localized analysis. This richness in annotations offers a solid foundation for evaluating crowd counting models in complex, real-world conditions.

Table 1. Statistics of benchmark crowd counting datasets.

Dataset	Images	Train/Val/Test	Total Annotations	Min Count	Avg Count	Max Count
SHT_A	482	300/0/182	241,677	33	501.4	3139
SHT_B	716	400/0/316	88,488	9	123.6	578
UCF_QNRF	1535	1201/0/334	1,251,642	49	815	12,865
JHU-CROWD	4250	3888/0/1062	1,114,785	-	262	7286

4.2. Experimental Settings

Data Preprocessing: Before training, all input images are resized and divided into overlapping patches of 224 × 224 pixels with a stride of 112 pixels. This patch-based training strategy increases the number of effective training samples and enables the network to better capture fine-grained local details,

ISSN 2616-5775 Vol. 8, Issue 11: 14-22, DOI: 10.25236/AJCIS.2025.081102

especially for small heads that may be easily overlooked in large, high-resolution images.

Optimization Strategy: The model is optimized using the AdamW optimizer, which provides better convergence stability for transformer-based and hybrid backbones by decoupling weight decay from gradient updates. The initial learning rate is set to 1×10^{-4}, and a cosine decay schedule is adopted to gradually reduce the learning rate during training, ensuring smoother convergence.

Loss Functions: As described in Section 3.5, MambaVision-Count employs a multi-task objective that combines local density estimation and global count regression. The density map branch is trained using the L1 loss between the predicted and ground-truth density maps, which encourages pixel-level accuracy and smooth density distributions. The count regression branch is trained using the Mean Absolute Error (MAE) loss on the predicted total count.

4.3. Lambda Parameter Tuning

In the dual-branch regression network, a joint loss function combining a density-map branch and a global-count branch is adopted. λ_1 and λ_2 are used to balance the supervision strengths of the two branches; their proper tuning is crucial for model performance. An overly small λ_2 leaves the global-count branch under-supervised, degrading overall counting accuracy, whereas an excessively large λ_2 makes the network focus too much on global regression and weakens the density-map branch's sensitivity to locally crowded regions. Likewise, too small a λ_1 impairs density-map learning and hurts local density prediction.

To determine optimal values, we conduct systematic experiments on the ShanghaiTech Part A dataset by fixing λ_1 and varying λ_2 within [0.01,0.05,0.1,0.2,0.5].Results show that when $\lambda_2 = 0.1$, the model achieves the best performance in both MAE and MSE metrics, indicating a well-balanced supervision between local and global branches. Therefore, in all subsequent experiments, we set $\lambda_1 = 1$ and $\lambda_2 = 0.1$ to ensure optimal convergence and stability.

This tuning strategy effectively harmonizes local density learning with global count regression, enabling consistent and accurate predictions across varying crowd densities.

4.4. Experimental Results and Analysis

We compare MambaVision-Count against six representative methods across multiple architectures: four CNN-based models (MCNN [12], CSRNet, CANNet [15], and SDANet [16]), one Transformer-based model (TransCrowd), and one Mamba-based model (VMambaCC).

MCNN employs a multi-column convolutional structure to handle scale variation by extracting features with different receptive fields. Each column focuses on different head sizes and fuses outputs into a single density map. While effective for moderate density variation, its limited contextual modeling constrains performance in highly congested scenes.

CSRNet uses a VGG-style backbone augmented with dilated convolutions to expand receptive fields without losing spatial resolution. It achieves strong performance in dense scenes by capturing broad contextual information, becoming a widely recognized CNN baseline.

CANNet integrates attention mechanisms into the CNN backbone to emphasize head regions and suppress background noise. By adaptively weighting spatial positions and employing multi-scale convolutions, it performs robustly under diverse crowd densities and complex backgrounds.

SDANet adopts shallow feature extraction combined with dense attention to enhance modeling of small-object and local spatial correlations. Its joint spatial–channel attention effectively highlights head regions while suppressing noise, showing excellent accuracy in high-density scenarios.

TransCrowd, built on the Vision Transformer (ViT), divides images into patches and models global dependencies via self-attention. It excels in capturing large-scale contextual relationships but is less sensitive to small, fine-grained details compared with CNN-based methods.

VMambaCC introduces Visual Mamba into crowd counting, leveraging state-space modeling to capture long-range dependencies and combining multi-scale feature pyramids and attention for improved contextual awareness. However, its ability to handle small targets and occlusions remains limited.

The quantitative comparison results are summarized in Table 2.

ISSN 2616-5775 Vol. 8, Issue 11: 14-22, DOI: 10.25236/AJCIS.2025.081102

Table 2. Performance comparison of mainstream crowd counting methods on four datasets.

Method	Architecture	SHT_A		SHT_B		UCF_QNRF		JHU-Crowd	
		MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE
MCNN	CNN	110.2	-	26.4	-	-	-	160.6	377.7
CSRNet	CNN	68.2	115.0	10.6	16	120.3	208.5	72.2	249.9
CANNet	CNN	62.3	100.0	7.8	12.2	107.0	183	100.1	314.0
SDANet	CNN	63.6	101.8	7.8	10.2	-	-	59.3	348.9
TransCrowd	Transformer	66.1	105.1	9.3	16.1	97.2	168.5	56.8	193.6
VmambaCC	Mamba	51.9	81.3	7.5	12.7	88.4	144.7	54.4	201.9
MambaVision-Count	Mamba+ Transformer	58.9	96.7	7.2	11.4	84.1	146.3	58.2	224.3

From Table 2, several observations can be drawn:

CNN-based methods such as CANNet and SDANet outperform earlier designs like MCNN, demonstrating the effectiveness of multi-scale feature extraction and attention mechanisms. However, their reliance on local receptive fields limits global understanding, leading to higher errors in large-scale or sparse scenes (e.g., MCNN on JHU-Crowd with MAE = 160.6).

Transformer-based TransCrowd achieves superior global modeling and performs well on large-scale datasets like UCF-QNRF (MAE = 97.2), outperforming most CNNs. Yet, its limited local precision results in suboptimal performance in dense, small-scale scenes (SHT_A/B).

Mamba-based VMambaCC achieves consistently strong results across datasets, confirming that state-space modeling effectively captures long-range dependencies and improves counting in dense regions.

MambaVision-Count, by integrating Mamba with Transformer components, achieves the best or near-best performance across datasets.It attains MAE/MSE = 58.9/96.7 on SHT_A, slightly higher than VMambaCC but superior on SHT_B (7.2/11.4) and UCF-QNRF (84.1/146.3).The model also maintains competitive accuracy on JHU-Crowd (58.2/224.3), demonstrating strong robustness and generalization.

These results validate that MambaVision-Count effectively combines local convolutional sensitivity with global dependency modeling, achieving high precision and stability across various crowd densities and environments.

4.5. Ablation Study

To further verify the effectiveness of key components, we conduct ablation experiments on the ShanghaiTech Part A dataset, focusing on the EFC fusion module and the dual-branch regression head. We compare the following model variants:

Baseline: MambaVision backbone only, with a single density regression head and no EFC module.

Baseline+EFC: Adds the EFC module to enhance small-object features but retains a single regression head.

Baseline+Dual-Branch: Incorporates the dual-branch regression head (density + count) but excludes EFC.

Full Model: Includes both EFC and the dual-branch regression head.

The comparison results are summarized in Table 3.

Table 3 Summarizes the ablation results.

Model Configuration	MAE	MSE
Baseline	62.3	106.1
Baseline+EFC	60.6	99.7
Baselin+Dual-Branch	61.1	100.5
Full Model	58.9	96.7

The ablation results reveal three key insights:

Adding the EFC module alone significantly enhances the model's sensitivity to small objects and dense regions, reducing MAE/MSE relative to the baseline.

Introducing the dual-branch head provides stronger global supervision, improving overall counting consistency.

ISSN 2616-5775 Vol. 8, Issue 11: 14-22, DOI: 10.25236/AJCIS.2025.081102

Combining both modules yields the best performance, confirming the complementary benefits of local feature enhancement and global dual-task supervision.

5. Conclusions

In this study, we introduced MambaVision-Count, a novel crowd counting framework built upon the MambaVision hybrid vision backbone. The proposed model achieves efficient feature extraction with a global receptive field and maintains linear computational complexity, making it well-suited for large-scale, high-density crowd analysis. To enhance the representation capability in complex and densely occluded scenes, we designed an Enhanced Feature Coupling (EFC) module that refines low-level features using high-level semantic guidance. This module effectively prevents the loss of important spatial information during feature fusion, thereby strengthening the model's response to small and heavily occluded targets. Moreover, a dual-branch regression structure was developed to jointly estimate density maps and global counts, enabling complementary supervision at both local and global levels. This design provides robust learning signals and improves the consistency between pixel-level and global-level predictions. Comprehensive experiments on multiple benchmark datasets demonstrate that MambaVision-Count consistently outperforms or matches existing state-of-the-art approaches in both accuracy and efficiency. The qualitative visualization of predicted density maps further confirms that the model exhibits superior localization and counting performance under challenging conditions such as severe occlusion, high density, and large scale variation.

In summary, MambaVision-Count establishes an effective methodological framework for integrating convolutional modeling, state-space sequence learning, and Transformer-based global reasoning into a unified architecture. This work provides new insights for future research in crowd analysis and offers a practical solution for real-world intelligent surveillance and urban management systems.

References

[1] Q. Wang, T. Breckon, Crowd counting via segmentation guided attention networks and curriculum loss, IEEE Trans. Intell. Transp. Syst. (2022).
[2] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention Is All You Need, Adv. Neural Inf. Process. Syst., vol. 30, pp. 5998–6008, 2017.
[3] A. Hatamizadeh, J. Kautz, MambaVision: A Hybrid Mamba-Transformer Vision Backbone, Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 25261–25270, 2025.
[4] X. Chu, A. Zheng, X. Zhang, J. Sun, Detection in Crowded Scenes: One Proposal, Multiple Predictions, Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 12214–12223, 2020.
[5] Y. Li, X. Zhang, D. Chen, Csrnet: Dilated convolutional neural networks for under-standing the highly congested scenes, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 1091–1100.
[6] D. Liang, X. Chen, W. Xu, Y. Zhou, X. Bai, Transcrowd: weakly-supervised crowd counting with transformers, Sci. China Inf. Sci. (2022).
[7] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M.Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An image is worth 16x16 words: transformers for image recognition at scale, ArXiv abs/2010.11929 (2021).
[8] Y. Tian, X. Chu, H. Wang, Cctrans: simplifying and improving crowd counting with transformer, ArXiv abs/2109.14483 (2021).
[9] A. Gu, T. Dao, Mamba: Linear-Time Sequence Modeling with Selective State Spaces, Adv. Neural Inf. Process. Syst. (NeurIPS), vol. 36, pp. 1–13, 2023.
[10] H.-Y. Ma, L. Zhang, S. Shi, VMambaCC: A Visual State Space Model for Crowd Counting, arXiv preprint arXiv: 2405. 03978, 2024.
[11] Y. Xiao, T. Xu, X. Yu, Y. Fang, and J. Li, "A lightweight fusion strategy with enhanced interlayer feature correlation for small object detection," IEEE Trans. Geosci. Remote Sens., vol. 62, pp. 1–11, 2024, doi: 10.1109/TGRS.2024.3457155.
[12] Y. Zhang, D. Zhou, S. Chen, S. Gao, Y. Ma, Single-image crowd counting via multi-column convolutional neural network, in: 2016 IEEE Conference on Computer Vi-sion and Pattern Recognition (CVPR), 2016, pp. 589–597. https://doi.org/10.1109/CVPR.2016.70.
[13] H. Idrees, M. Tayyab, K. Athrey, D. Zhang, S.A. Al-Maadeed, N.M. Rajpoot, M. Shah, Composition loss for counting, density map estimation and localization in dense crowds, ArXiv abs/1808.01050 (2018).
[14] V.A. Sindagi, R. Yasarla, V.M. Patel, Jhu-crowd++: Large-scale crowd counting dataset and a

ISSN 2616-5775 Vol. 8, Issue 11: 14-22, DOI: 10.25236/AJCIS.2025.081102

benchmark method, Technical Report (2020).
[15] J. Chen, W. Su, and Z. Wang, "Crowd counting with crowd attention convolutional neural network," arXiv preprint arXiv: 2204.07347, 2022.
[16] Y. Miao, Z. Lin, G. Ding, and J. Han, "Shallow feature based dense attention network for crowd counting," in Proc. AAAI Conf. Artif. Intell., vol. 34, no. 01, pp. 10077–10084, 2020.