applied sciences

MDPI

Article

Learning to Count Crowds from Low-Altitude Aerial Views via Point-Level Supervision and Feature-Adaptive Fusion

Junzhe Mao [1], Lin Nai [1], Jinqi Bai [1], Chang Liu [1] and Liangfeng Xu [2,*]

[1] College of Computer Science, Beijing Information Science and Technology University, Beijing 100192, China
[2] State Post Bureau Safety Supervision Center, Beijing 100091, China
* Correspondence: liangfengxu_sub@126.com

Abstract

Counting small, densely clustered objects from low-altitude aerial views is challenging due to large scale variations, complex backgrounds, and severe occlusion, which often degrade the performance of fully supervised or density-regression methods. To address these issues, we propose a weakly supervised crowd counting framework that leverages point-level supervision and a feature-adaptive fusion strategy to enhance perception under low-altitude aerial views. The network comprises a front-end feature extractor and a back-end fusion module. The front-end adopts the first 13 convolutional layers of VGG16-BN to capture multi-scale semantic features while preserving crucial spatial details. The back-end integrates a Feature-Adaptive Fusion module and a Multi-Scale Feature Aggregation module: the former dynamically adjusts fusion weights across scales to improve robustness to scale variation, and the latter aggregates multi-scale representations to better capture targets in dense, complex scenes. Point-level annotations serve as weak supervision to substantially reduce labeling cost while enabling accurate localization of small individual instances. Experiments on several public datasets, including ShanghaiTech Part A, ShanghaiTech Part B, and UCF_CC_50, demonstrate that our method surpasses existing mainstream approaches, effectively mitigating scale variation, background clutter, and occlusion, and providing an efficient and scalable weakly supervised solution for small-object counting.

Keywords: low-altitude aerial views; clustered small object counting; point-level supervision; weak supervision; feature-adaptive fusion

check for updates

Academic Editor: Suk-Ho Lee

Received: 2 December 2025
Revised: 15 December 2025
Accepted: 15 December 2025
Published: 17 December 2025

Citation: Mao, J.; Nai, L.; Bai, J.; Liu, C.; Xu, L. Learning to Count Crowds from Low-Altitude Aerial Views via Point-Level Supervision and Feature-Adaptive Fusion. *Appl. Sci.* **2025**, *15*, 13211. https://doi.org/10.3390/app152413211

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Clustered small object counting, as one of the important tasks in computer vision, primarily focuses on identifying, locating, and counting dense small objects in complex scenes. In recent years, advancements in deep learning and computer vision have demonstrated the significant value of clustered small object counting in application scenarios such as low-altitude aerial applications, public safety and urban management. Although application scenarios differ, the main challenges faced by clustered small object counting share commonalities: significant target scale variation, complex and strong background interference, severe object density, overlap, and occlusion. To address these issues, this paper proposes a weakly supervised method based on point-level supervision, combined with a feature-adaptive fusion strategy, to achieve accurate counting and localization of clustered small objects for low-altitude aerial views while reducing annotation costs and improving the model's practicality and robustness.

In crowd counting tasks, this method is particularly crucial in scenarios requiring real-time monitoring of large crowds, such as public places, sports events, transportation hubs and other low-altitude aerial applications. High crowd density and rapid dynamic changes, coupled with prevalent occlusion and overlap, make it difficult for traditional density regression or object detection methods to ensure counting accuracy in complex environments. Figure 1 shows typical crowd counting scenes, including concerts, railway stations, commercial streets, and tourist attractions. These areas are characterized by dense populations, high mobility, and potential management difficulties and safety hazards. Real-time monitoring and analysis of crowds using computer vision technology can not only assess crowd density but also provide decision support for public safety management. Accurate counting and analysis help implement effective crowd control measures, ensuring public order and safety.



Figure 1. Example images of crowds in locations such as shopping malls and streets [1].

The task of counting clustered small objects still faces numerous challenges in practical applications, mainly including: (1) Scale variation: Differences in camera angle, shooting distance, and focal length lead to significant differences in target sizes within images; targets closer to the camera appear larger, while those farther away appear smaller, potentially affecting model performance. (2) Complex background interference: Buildings, vegetation, or street signs in the background can easily be confused with target features, causing misidentification. (3) Uneven target distribution: Coexistence of locally dense and sparse areas increases local prediction difficulty, leading to counting errors. (4) Occlusion issues: Partial or complete occlusion between dense small objects makes accurate localization of individual targets difficult. (5) Lack of datasets: Scarcity of high-quality point-level annotated samples limits model training and generalization capability.

To address the aforementioned challenges, the network proposed in this paper uses point-level annotations as supervision signals, belonging to the category of weakly supervised learning, which can significantly reduce annotation costs while achieving precise localization of individual targets. The network structure consists of a front-end feature extraction module and a back-end feature fusion module. The front-end uses the first 13 convolutional layers of VGG16-BN to extract multi-scale deep features, preserving key

spatial information and intermediate layer features. The back-end dynamically adjusts the weights of features at different scales through the Feature Adaptive Fusion module and combines it with the Multi-scale Feature Aggregation module to further integrate multi-level information, thereby enhancing the model's perception capability in dense areas and complex scenes. The proposed method can effectively improve the model's adaptability and counting accuracy for scenes with scale variation, complex backgrounds, and dense occlusion, providing an efficient and scalable solution for small object counting tasks in public safety and urban management.

The main contributions of this paper are as follows: (1) a weakly supervised method that leverages point-level annotations to accurately localize small objects in densely packed and occluded scenes, reducing annotation costs while maintaining high precision; (2) the development of a feature-adaptive fusion strategy that dynamically adjusts the integration of multi-scale features, improving the model's robustness to scale variation, background interference, and occlusion in complex low-altitude aerial environments; and (3) an empirical demonstration of the method's superior performance over existing state-of-the-art approaches, validated across multiple public datasets.

2. Related Work

2.1. Crowd Counting

Crowd counting aims to estimate the number of people in an image, a task that faces many difficulties. With continuous deepening of research, extensive exploration has been conducted, leading to significant improvements in counting accuracy and performance. Existing crowd counting methods can be broadly categorized into traditional methods and deep learning methods.

Traditional crowd counting algorithms, based on different counting principles, can be divided into detection-based methods and regression-based methods. In early research, detection-based computer vision methods were primarily applied to object detection, especially pedestrian detection. These methods typically rely on detecting the entire target or partial body parts to estimate the number of pedestrians in an image. Holistic detection methods usually first preprocess the image, then extract full-body pedestrian features such as edge features [2], Haar wavelet features [3], Histogram of Oriented Gradient (HOG) features [4], and use classifiers like Support Vector Machine (SVM) [5], Random Forest (RF) [6], or Boosting algorithms [7] for target recognition. Such methods can relatively accurately detect pedestrians and estimate the total crowd count, but in dense crowd scenes, their detection performance is significantly affected due to prevalent occlusion.

To address the occlusion problem in dense crowds, researchers proposed detection methods based on partial body parts (such as heads, shoulders, etc.). This method reduces errors caused by full-body occlusion by detecting local features of pedestrians, improving detection accuracy in dense scenes. Li et al. [8] combined HOG features with a foreground segmentation algorithm based on mosaic image difference, improving adaptability to dense scenes by detecting head and shoulder features in crowd areas, achieving good results in dense crowd counting tasks. Zhao et al. [9] proposed a detection method using ellipsoids to simulate 3D human models. This method not only enables pedestrian segmentation and tracking but also infers occlusion between pedestrians, suitable for relatively sparse or mildly occluded scenes. Although detection-based methods perform well in low-density crowds, their effectiveness is limited in high-density or complex backgrounds; when crowd density increases and occlusion intensifies, these methods struggle to cope effectively.

Addressing the limitations of detection-based methods, researchers proposed regression-based methods. The core idea is to directly learn the mapping relationship between image features and the number of people, treating the group to be detected as a whole,

and estimating group density through regression or classification models, thus avoiding the limitations of traditional detection methods. Commonly used features include foreground pixel features, edge features, texture features, Local Binary Patterns (LBP), and Gray Level Co-occurrence Matrix, etc. [10]; regarding regression methods, linear regression, ridge regression, Gaussian process regression, and Bayesian regression, etc. [11] have been widely applied. Compared with detection-based methods, regression methods perform better in dense scenes and demonstrate stronger generalization ability and portability [12]. However, traditional regression methods often do not fully consider spatial information in the image and can only output the total number of people in the scene, unable to depict the spatial distribution of the crowd in detail, which has certain limitations in practical applications.

To address this issue, Lempitsky et al. [13] proposed transforming the crowd counting task into a density map prediction problem, learning the relationship between low-level features and crowd density maps to achieve spatial information learning. Simultaneously, introducing a Gaussian kernel to generate ground truth density maps distributes target weights to adjacent pixels, allowing each pixel in the density map to contain information about surrounding targets, thereby predicting the total number of people through global integration. Compared with traditional methods, density map-based regression algorithms do not require precise detection of individuals, focusing only on group density distribution, performing better in complex scenes, comprehensively utilizing all pixel information in the image, fully considering spatial structure and local details, and significantly improving counting accuracy. Therefore, density map-based training methods have recently become the mainstream approach in crowd counting.

Convolutional Neural Networks (CNNs) [14] have shown significant advantages in counting accuracy and robustness. CNNs learn nonlinear mapping relationships in images, mapping image features to density maps. Depending on the network structure, CNN-based counting methods can be divided into basic CNN architectures, multi-column CNN architectures, and single-column CNN architectures. Fu et al. [15] used a two-layer convolutional network to extract image features, then mapped them to a density map via a fully connected layer, achieving fast estimation without manual features. Due to its simple structure, this method still has limitations when processing high-resolution or complex scenes, thus requiring improvements to the network architecture to enhance performance. Wang et al. [16] proposed a density estimation prediction network architecture based on AlexNet; Walach et al. [17] proposed a CNN Boosting framework combining multiple CNN models, introducing innovative data augmentation methods to improve generalization ability and counting accuracy. Although basic CNN architectures are easy to implement, they lack additional feature information, resulting in limited counting accuracy.

Multi-column CNN architectures capture multi-scale information with different receptive fields through multiple columns of networks with different structures, significantly improving counting accuracy. Zhang et al. [1] proposed the MCNN model using a multi-column convolutional neural network, where each column is assigned convolution kernels of different sizes to process crowds at different scales, enhancing the network's perception ability for regions of different scales. Sam et al. [18] proposed Switch-CNN, which can adaptively adjust the network structure to handle scenes of different densities and scales, combining dense and sparse predictions to obtain local and global features, thereby improving counting accuracy. Sindagi et al. [19] proposed a Contextual Pyramid CNN, analyzing images from multiple scales and introducing contextual information for better understanding of crowd distribution. Guo et al. [20] proposed DADNet, combining dilated convolution, attention mechanisms, and deformable convolution techniques to further improve crowd counting accuracy. However, multi-column CNN architectures still face

issues such as large computational load, information redundancy, and efficiency problems caused by variations in image patch sizes.

Single-column CNN architectures typically deploy a more complex convolutional network, avoiding overly complex network structures while maintaining high counting accuracy and computational efficiency. Li et al. [21] proposed CSRNet, using dilated convolution to expand the receptive field and capture rich contextual information. Cao et al. [22] proposed SANet, aggregating information through multi-scale pyramid down-sampling and up-sampling to improve perception ability for crowd regions of different scales. Liu et al. [23] proposed ADCrowdNet, combining visual attention mechanisms and a multi-scale deformable convolution framework to adaptively optimize the importance of channel feature maps, effectively improving crowd counting performance in complex scenes.

2.2. Point-Supervised Small Object Counting Algorithms

The difficulty in small object counting lies in the tiny target size, complex scenes, and dense distribution characteristics. Point-supervised small object counting methods directly utilize point-level annotation information, achieving more precise localization supervision while reducing annotation costs.

Point-supervised small object counting algorithms have made breakthroughs in recent years. Their core advantage lies in achieving precise localization supervision through point-level annotations while significantly reducing annotation costs. Traditional methods rely on density map regression or detection box generation as intermediate representations. This indirect supervision method can easily lead to loss of localization accuracy for small objects. The P2PNet framework proposed by Song et al. [24] pioneered a point-to-point direct mapping paradigm. Unlike traditional localization methods that rely on intermediate representations (such as density maps or pseudo-boxes) as learning targets, this framework avoids intuitively unreasonable and error-prone practices. Wang et al. [25] proposed DM-Count, which models the similarity between predicted point distribution and ground truth point distribution using optimal transport theory, avoiding the information distortion problem caused by conventional Gaussian kernel blurring. Liu et al. [26] proposed BM-Count, which adopts a two-stage matching strategy combining K-nearest neighbor topology screening and bidirectional soft assignment. This network generates corresponding bounding boxes while locating target center points, transforming the counting task into a detection problem. Its training process uses a bounding box regression loss function to optimize network parameters through backpropagation. Notably, the introduction of semi-supervised learning paradigms has opened new avenues for small object counting.

Addressing the balance between annotation cost and detection accuracy, Chen et al. [27] proposed various weak semi-supervised learning paradigms. The improved method based on the DETR architecture [28], Point DETR, showed significant advantages on the MS-COCO dataset: when using only 20% of fully annotated data, its detection accuracy reached 33.3 AP, an improvement of 2.0 AP over the FCOS baseline. This method expands the feature representation space by introducing a point encoder, achieving a performance improvement of over 10 percentage points on the AR series of metrics.

Although point-level-supervised small object counting algorithms have achieved good performance, they suffer from decreased counting accuracy in clustered environments and under severe occlusion. Furthermore, existing models fail to fully utilize multi-scale feature information, leading to missed detections in counting clustered small objects. Research on counting clustered small objects for other subjects is still insufficient. Therefore, work focused on counting clustered small objects warrants in-depth study.

2.3. Crowd Counting Datasets

To address the crowd counting problem in complex scenes, researchers have proposed several datasets, as shown in Table 1, such as ShanghaiTech Part A [1], ShanghaiTech Part B [1], and UCF_CC_50 [29]. These datasets mostly originate from publicly available internet images or public surveillance videos, covering scenes of different densities, with large sample sizes and precise annotations. The datasets consider factors such as weather, viewpoint, lighting, density, and resolution, and have high annotation accuracy, becoming important research resources in the field of crowd counting and promoting the optimization and accuracy improvement of counting models.

Table 1. Comparison of different crowd counting datasets.

Dataset Name	Image Number	Annotation Count	Average Resolution	Average Count	Train Set	Test Set
ShanghaiTech Part A	482	241,677	589 × 868	510	300	182
ShanghaiTech Part B	716	88,488	768 × 1024	124	400	316
UCF_CC_50	50	63,974	2101 × 2888	1280	40	10

2.3.1. UCF_CC_50

The UCF_CC_50 dataset focuses on high-density crowd counting tasks, containing 50 high-resolution grayscale images with an average resolution of 2101 × 2888, sourced from stadiums, concerts, and large public events. Each image contains between 94 and 4543 point annotations, with an average of 1280 annotated targets, totaling 63,974 point annotations. High-density annotations help models better understand crowd distribution patterns and improve prediction accuracy. Due to the limited sample size, 5-fold cross-validation is typically used in research, with 4 folds used as the training set and 1 fold as the test set each time, cycling 5 times, and calculating average performance metrics to obtain reliable evaluation results. The dataset can be accessed at [https://www.crcv.ucf.edu/data/ucf-cc-50/], last accessed on [1 December 2025].

2.3.2. Shanghai Tech Part A

The ShanghaiTech Part A dataset consists of 482 images, with 300 used for training and 182 for testing. The images are sourced from the internet, covering diverse scenes with high density. The number of head annotations per image ranges from 33 to 3139, with a high average count, making it suitable for testing crowd counting algorithms in high-density scenes. The large scene variation and diverse viewpoints in this dataset require models to have strong adaptability to cope with complex perspectives and occlusion situations. Its high density and complex scene characteristics make it an important resource for evaluating and developing efficient crowd counting algorithms, providing data support for practical applications. The dataset can be accessed at [https://pan.baidu.com/s/1xJnhmJbwPdnNKBM1K6F1Cg?pwd=iga3], last accessed on [1 December 2025].

2.3.3. Shanghai Tech Part B

The ShanghaiTech Part B dataset contains 716 images, with 400 used for training and 316 for testing. The images are mainly from Shanghai city streets, with low crowd density. The number of head annotations ranges from 9 to 578, distributed relatively evenly, making it suitable for evaluating crowd counting algorithms in low-density scenes. This dataset has less scene variation and small perspective changes, with relatively consistent crowd appearances, making the detection task relatively simple. It can be used to validate the

robustness and accuracy of models in sparse crowd scenes. The dataset can be accessed at [https://pan.baidu.com/s/1xJnhmJbwPdnNKBM1K6F1Cg?pwd=iga3], last accessed on [1 December 2025].

Existing crowd counting and small object counting methods still face several challenges in dense or clustered environments: traditional detection methods are prone to inaccurate counting in high-density scenes due to occlusion; regression-based density map methods, while considering spatial distribution, have limited localization accuracy for small objects; point-supervised small object counting methods still suffer from missed detections in clustered and multi-scale environments. Addressing the above issues, the work in this paper has the following innovations:

- Joint modeling of density and location for clustered small objects: Unlike methods relying solely on global density regression or single-point detection, this paper proposes a strategy combining density maps and point-level supervision, effectively addressing occlusion and dense distribution issues while maintaining small object localization accuracy.
- Multi-scale feature fusion strategy: By fusing image features at different scales, the network's perception ability for small objects of varying sizes and uneven distributions is enhanced, reducing missed detections and duplicate counting.
- Efficient training and annotation cost optimization: While maintaining high accuracy, it combines point-level annotation with semi-supervised strategies, effectively reducing annotation costs and improving the model's generalization ability on limited datasets.

3. Crowd Counting Based on Point-Level Supervision and Feature-Adaptive Fusion

In the task of counting small clustered objects, traditional methods often rely on single-scale feature extraction, ignoring the dynamic changes in target scales, which constrains the model's generalization. Meanwhile, occlusion and background noise in dense target areas often interfere with target localization and identification, posing significant challenges for accurate counting. Therefore, how to extract effective target features under various interfering factors becomes a key issue in this field. To reduce annotation costs and improve model performance in point-level annotated scenarios, this paper utilizes a weakly supervised strategy based on point supervision, directly using point-level annotations for supervision, enabling the model to achieve precise localization and counting of small objects without relying on dense bounding boxes or density map generation. This strategy not only reduces expensive annotation workload but also enhances the model's generalization ability in dense and occluded environments. On this basis, this paper proposes a crowd small object counting method based on feature-adaptive fusion. This method combines a front-end backbone network with a back-end multi-scale feature fusion module, effectively enhancing the model's different scales performance and complex backgrounds. The fusion weights of features from different branches were adjusted by the Feature Adaptive Fusion module dynamically, achieving optimized integration of multi-scale information and enhancing the model's robustness to small object scale variations; the back-end Multi-scale Feature Aggregation module expands the field of reception, enabling the model to count small objects more accurately when processing dense areas. Furthermore, combined with a customized loss function, the model can pay more attention to the spatial distribution of counting errors during training, thereby further improving overall counting accuracy and robustness.

3.1. Overall Network Architecture

This paper proposes a weakly supervised counting model for small clustered objects based on feature-adaptive fusion. The network aims to improve traditional counting models that cannot fully utilize multi-scale feature information in densely occluded areas. The model includes a feature extraction module, a feature adaptive fusion module, and a multi-scale feature aggregation module, achieving more accurate counting of crowd small objects. Its overall structure is shown in Figure 2.

In the feature extraction module, the first thirteen convolutional layers of VGG16-BN [30] are used to extract deep semantic features of the targets, while preserving key spatial detail information and intermediate layer feature information, providing multi-level feature support for subsequent modules. We selected VGG16-BN due to its simplicity and effectiveness in extracting multi-scale features while preserving spatial details, which is crucial for counting small objects in crowded scenes. In contrast, more complex architectures like CSRNet or SwinViT may introduce unnecessary computational overhead, especially when dealing with point-level supervision. Secondly, the Feature Adaptive Fusion module dynamically adjusts the weight relationships between features at different scales, fully utilizing the low-level features' spatial resolution advantages and the high-level features' semantic information advantages, thereby achieving efficient fusion of multi-scale features.

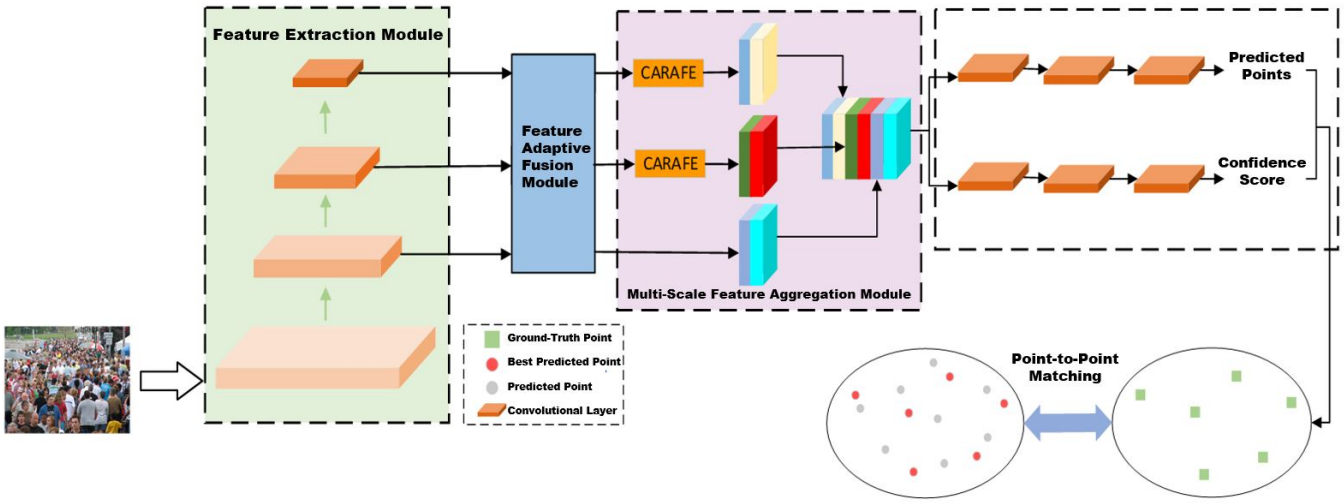


Figure 2. Overall network architecture diagram.

In the module of Multi-scale Feature Aggregation, the model further enhances the perception ability for target areas of different densities through a specific aggregation strategy, ensuring excellent performance in both dense and sparse scenes. Simultaneously, the matching process between predicted points and ground truth points using the Hungarian algorithm improves the model's localization and counting accuracy for crowd small objects.

The crowd small object counting network directly accepts point annotations as its learning target and provides the precise location of individuals within the small objects, rather than merely counting the number of individuals. The individual's location is typically represented by the center point of the head, with an optional confidence score. Specifically, given an image containing N targets, the head center point of the i-th target is denoted by $p_i = (x_i, y_i)$, located at (x_i, y_i). The set of center points for all targets can be represented as $P = p_i \mid i \in \{1, \ldots, N\}$. The model is trained to output two sets $\hat{P} = \{\hat{p}_j \mid j \in \{1, \ldots, M\}\}$ and $\hat{C} = \{\hat{c}_j \mid j \in \{1, \ldots, M\}\}$, where M is the number of predicted targets, and $\hat{c}_j$ is

the confidence score of the predicted point $\hat{p}_j$. Generally, it can be assumed that $\hat{p}_j$ is the prediction of the ground truth point p_i. The model's goal is to ensure that the distance between $\hat{p}_j$ and p_i is as close as possible, and the score $\hat{c}_j$ is sufficiently high. Simultaneously, the predicted number of individuals M should be sufficiently close to the true number of small objects N. In short, the new framework can simultaneously achieve small object counting and individual localization.

3.2. Feature Adaptive Fusion Module

To achieve precise fusion of multi-scale features in crowd small object detection scenes, this paper designs a Feature Adaptive Fusion module. This module generates feature fusion weights and simultaneously utilizes a content-aware upsampling mechanism to achieve alignment and efficient integration of cross-scale features. The specific design of the AFF is shown in Figure 3.

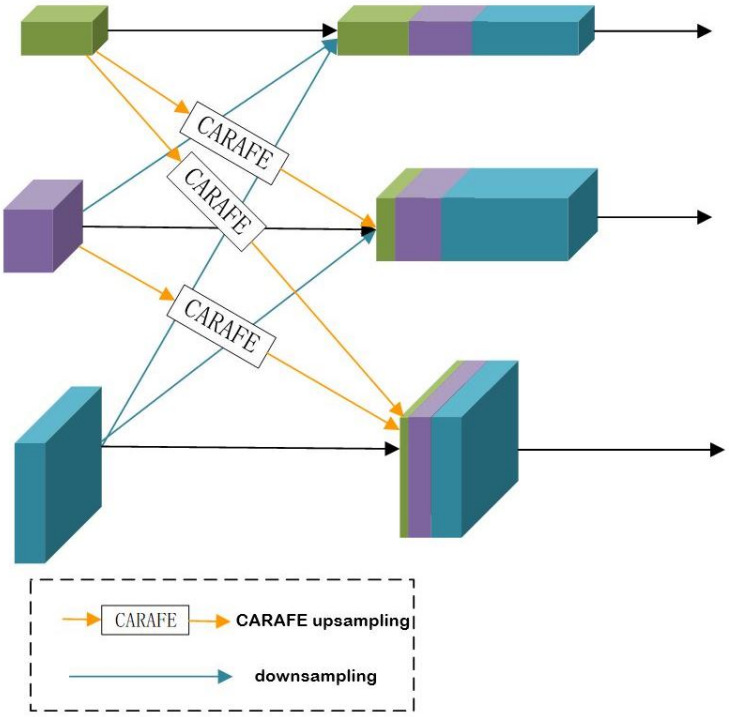


Figure 3. Structure of the Feature Adaptive Fusion module.

The input to the AFF includes feature maps from three scales X_1, X_2, X_3, representing low-level, mid-level, and high-level features, respectively. Taking the X_1 branch as an example, the dimensions of the X_2 and X_3 branches are aligned to X_1. Here, CARAFE (ContentAware ReAssembly of FEatures) content-aware dynamic upsampling is adopted. CARAFE is a content-aware dynamic upsampling method. X_2 and X_3 are upsampled by factors of 2 and 4, respectively. Dynamic fusion weights are generated for each aligned branch feature map. X_1, X_2, X_3 are first multiplied by the weight parameters α, β, γ and then summed. The formula for obtaining the new fused feature Y is as follows:

$$Y = \alpha_l \cdot X_{1 \to l} + \beta_l \cdot X_{2 \to l} + \gamma_l \cdot X_{3 \to l}, \tag{1}$$

Here, the aligned feature maps from different scales are passed through a 1×1 convolutional layer to compute the weight parameters α, β, γ. The parameters α, β, and γ are then concatenated and passed through a softmax function so that their ranges are within $[0, 1]$ and their sum is 1. The same applies to other branches.

This weighted summation process fully utilizes the information from multi-scale features while highlighting the features more important for the final detection task through dynamic weight allocation. This fusion method ensures efficient processing of the crowd small object detection task subsequently, especially performing prominently in complex scenes or when target scales are diverse.

3.3. Multi-Scale Feature Aggregation Module

Small objects are often distributed across feature maps of different scales, and features from a single scale are difficult to comprehensively represent the diversity of small objects. To this end, this paper designs an efficient module of Multi-scale Feature Aggregation, as shown in Figure 4, thereby improving the accuracy and robustness of small object detection.

The backbone network generates feature maps with different resolutions and semantic depths through layer-by-layer feature extraction. This paper selects the low-level, mid-level, and high-level features of the network, denoted as C_3, C_4, and C_5, respectively. Among them, C_3 is the low-level feature with the highest resolution, containing more spatial detail information; C_4 and C_5 are mid-level and high-level features, with progressively lower resolutions but stronger semantic information. To effectively fuse these features, it is necessary to upsample the lower-resolution C_4 and C_5 to align with the spatial resolution of C_3.

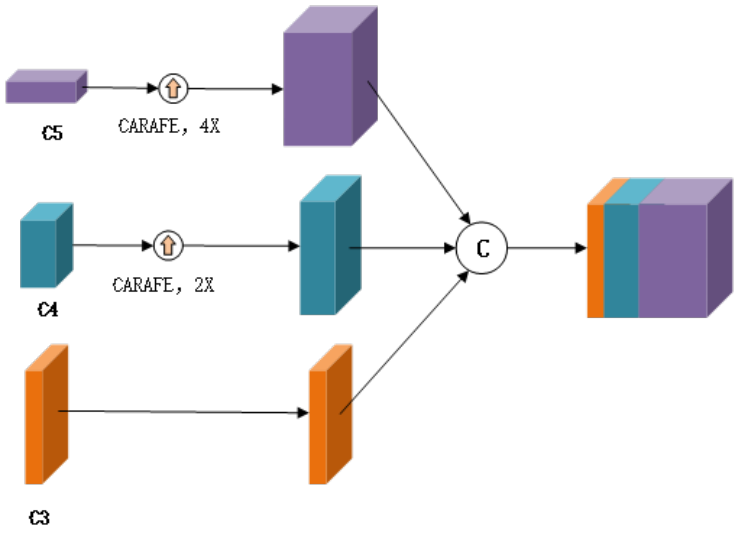


Figure 4. Multi-scale Feature Aggregation module's structure.

The CARAFE content-aware dynamic upsampling method is used for resolution alignment of C_4 and C_5. CARAFE is a content-aware dynamic upsampling method whose main characteristic is dynamically generating convolution kernels based on local context, thereby reconstructing features with higher accuracy, and having better detail preservation capabilities compared to ordinary bilinear interpolation or transposed convolution. CARAFE uses context-based learnable reorganization kernels, effectively avoiding the blurring effect and loss of detail information present in traditional upsampling methods. In this way, the feature maps of all branches are aligned in spatial resolution, laying the foundation for subsequent multi-scale feature fusion.

3.4. Design of Loss Function

The appropriate loss function design is crucial for improving model performance during the training of deep learning models. The model employs a combination of two loss

functions: cross-entropy loss for the classification task and Euclidean loss for the regression task. The classification loss L_{cls} is used to measure the model's classification performance, calculated as:

$$L_{cls} = -\frac{1}{M}\left[\sum_{i=1}^{N} \ln \widehat{c}_{\partial(i)} + \lambda_1 \sum_{i=N+1}^{M} \ln(1 - \hat{c}_{\partial(i)})\right], \quad (2)$$

where M is the number of samples, $\hat{c}_{\partial(i)}$ is the confidence score of the i-th predicted point. The cross-entropy loss helps the model optimize its classification performance. The regression loss L_{loc} is used to measure the model's regression error, formulated as:

$$L_{loc} = \frac{1}{N}\sum_{i=1}^{N} \|p_i - \hat{p}_i\|_2^2 \quad (3)$$

where N is the number of samples, p_i is the i-th ground truth point, and $\hat{p}_i$ is the best matching predicted point for the i-th ground truth point. The Euclidean loss optimizes the model's performance in the regression task by calculating the squared difference between predicted points and ground truth points. The final loss function L_{total} is a liner combination of these two losses:

$$L_{total} = L_{cls} + \lambda_2 L_{loc} \quad (4)$$

where λ_2 is a weight term balancing the impact of the regression loss, and λ_1 is the weight factor for negative predicted points. By appropriately adjusting these weight factors, the model's performance in classification and regression tasks can be balanced, thereby improving overall performance. This design combining cross-entropy and Euclidean loss can simultaneously optimize the model's classification and regression performance, helping it perform more excellently in small object counting tasks.

In summary, this paper proposes a crowd small object counting model based on feature-adaptive fusion. Through the feature extraction module, feature adaptive fusion module, and multi-scale feature aggregation module, it enhances the perception ability of dense areas and effectively improves counting accuracy. By precisely locating the head center points of targets instead of relying on traditional bounding boxes or other coarse annotation methods, this model provides more refined information for subsequent small object counting, especially suitable for environments with dense occlusion and uneven target distribution.

4. Experimental Results and Analysis

4.1. Experimental Setup

4.1.1. Experimental Environment and Parameter Settings

All experiments were conducted on an Ubuntu system (22.04) with an Nvidia 3090 GPU, implemented using the PyTorch 2.0.0 framework. During the data loading phase, the batch size of the DataLoader was set to 8. The Adam algorithm was used as the gradient optimization method to optimize the model training process. During the training phase, the number of training epochs was set to 2100. To adjust the learning rate and improve training efficiency, a Cosine Annealing Learning Rate Scheduler was used, setting the learning rate decay period to 300, gradually reducing the learning rate from the initial value until reaching the minimum value. This scheduler avoids premature convergence by gradually decreasing the learning rate. The model's complexity is characterized by 45 million parameters and requires 12.5 GFLOPs for inference. The model achieves an inference speed of 10 FPS on a GeForce RTX 3090 GPU.

4.1.2. Data Augmentation

In the task of counting crowd small objects, data augmentation is an important means to improve model performance, especially to alleviate insufficient training data and enhance generalization capability. Addressing the challenges of dense targets, large scale variations, and complex backgrounds, this paper adopts a comprehensive augmentation strategy including random scaling, random cropping, and random flipping to expand training data diversity and improve model adaptability. The random scaling factor was set to [0.7, 1.3], the crop size was set to 128 × 128 pixels, and random flipping was applied horizontally. Ground truth point coordinates are adjusted proportionally, and images are processed using bilinear interpolation to ensure detail integrity. Random cropping generates local area images to strengthen local feature learning simultaneously adjusting the positions of target density points, expanding training samples, and improving recognition capability for local crowd targets. Random flipping increases sample diversity through horizontal mirroring and synchronously adjusts target coordinates, enabling the model to adapt to target distributions from different viewpoints and postures, improving counting accuracy and robustness.

4.2. Comparative Experimental Results and Analysis

4.2.1. Experimental Results on ShanghaiTech Part A Dataset

Table 2 summarizes the experimental results of the proposed method on the ShanghaiTech Part A dataset, comparing its performance with several leading crowd counting algorithms in recent years. The table uses bold to indicate the optimal results and underlining to indicate the suboptimal results. It can be seen that the proposed method shows significant advantages on this dataset, where the Mean Absolute Error (MAE) reaches 51.52, which is 1.22 lower than the P2PNet algorithm, making it the best result among all methods; the Mean Squared Error (MSE) is 81.04, achieving a best performance. This indicates that the proposed method outperforms most existing algorithms in terms of accurate counting and error reduction.

Table 2. Experimental results on the ShanghaiTech Part A dataset.

Model	Year	MAE	MSE
CSRNet [21]	2018	68.2	115.0
CAN [31]	2019	62.3	100.0
DM-Count [25]	2020	59.7	95.7
P2PNet [24]	2021	52.74	85.06
GauNet [32]	2022	54.8	89.1
CLTR [33]	2022	56.9	95.2
FGENet [34]	2024	<u>51.66</u>	85.0
VMambaCC [35]	2024	51.87	<u>81.3</u>
Proposed Method		**51.52**	**81.04**

For better comparison of the actual results of the models, a visual comparison of the ground truth, the P2PNet model, and the proposed model was conducted. Three images were randomly selected from the ShanghaiTech Part A dataset for comparison, visually presenting the results. Figure 5 shows the ground truth on the left, the P2PNet model in the middle, and the proposed model on the right.

It can be clearly seen from the figure that the proposed model has significantly improved counting accuracy compared to P2PNet. Specifically, the difference between the ground truth shown on the left and the prediction results on the right is small, indicating that the proposed model can better approximate the real crowd distribution, especially in clustered areas and complex scenes, where its counting accuracy is significantly better

than the P2PNet model. In the prediction results of the P2PNet model, there are large errors in some high-density areas, especially where crowds are densely gathered or severely occluded, where the P2PNet model tends to overestimate or underestimate. The proposed method effectively reduces these errors and can more accurately capture density changes and local features of the crowd, especially between complex backgrounds and crowded targets, stably outputting more accurate count values.



Figure 5. Visual comparison results on the ShanghaiTech Part A dataset.

4.2.2. Experimental Results on ShanghaiTech Part B

On the ShanghaiTech Part B dataset, the proposed method achieved a Mean Absolute Error (MAE) of 6.78, slightly inferior to the optimal model (gap of 0.58), and a Mean Squared Error (MSE) of 9.68, as shown in Table 3. This result ranks among the top in current mainstream crowd counting algorithms, indicating that the proposed method can relatively accurately estimate crowd numbers in crowd counting tasks.

Table 3. Experimental results on the ShanghaiTech Part B.

Model	Year	MAE	MSE
CSRNet [21]	2018	10.6	16.0
CAN [31]	2019	7.8	12.2
DM-Count [25]	2020	7.4	11.8
P2PNet [24]	2021	<u>6.25</u>	<u>9.9</u>
GauNet [32]	2022	**6.2**	<u>9.9</u>
CLTR [33]	2022	6.5	10.6
FGENet [34]	2024	6.34	10.53
VMambaCC [35]	2024	7.48	12.47
Proposed Method		6.78	**9.68**

A visual comparison of the ground truth, the P2PNet model, and the proposed model was conducted. Three images were randomly selected from the ShanghaiTech Part B dataset for comparison, visually presenting the results. Figure 6 shows the ground truth on the left, the P2PNet model in the middle, and the proposed model on the right.



Figure 6. Results comparison on the ShanghaiTech Part B dataset.

It can be clearly seen from the comparison figures that the network model proposed in this paper performs more accurately in predictions compared to the P2PNet model. Especially when processing the ShanghaiTech Part B dataset, although the crowd density distribution in this dataset varies greatly and the scenes are complex, the proposed model can still relatively accurately capture the spatial distribution characteristics of the crowd. Although in terms of conventional evaluation metrics such as MAE and MSE, the performance of the proposed model on the ShanghaiTech Part B dataset did not completely surpass the P2PNet model, its results are still very close, with a small gap. Besides traditional quantitative metrics, the actual detection effect is particularly important for the practical application of the model. Judging from the actual detection results, the proposed model demonstrates more excellent performance in various complex scenes, especially in high-density crowds and under severe target occlusion. Compared to the P2PNet model, the proposed model's predictions are closer to the ground truth and can more accurately predict crowd numbers. The advantage of this model is especially evident in dense areas; the proposed model can better handle scenes with dense crowds and limited visual information, and the error between the generated counting results and the actual values is significantly smaller.

4.2.3. Experimental Results on UCF_CC_50

Table 4 summarizes the experimental results of the proposed method on the UCF_CC_50 dataset, providing a detailed performance comparison with several excellent crowd counting algorithms in recent years. For ease of analysis, the table uses bold to indicate the optimal results and underlining to indicate the suboptimal results.

Table 4. Experimental results on the UCF_CC_50 dataset.

Model	Year	MAE	MSE
CSRNet [21]	2018	215.4	296.4
CAN [31]	2019	212.2	243.7
DM-Count [25]	2020	211.0	291.5
P2PNet [24]	2021	172.7	256.1
GauNet [32]	2022	186.3	256.5
CLTR [33]	2024	142.56	215.87
FGENet [34]	2024	<u>64.34</u>	**78.71**
Proposed Method		**64.2**	<u>103.62</u>

To more intuitively validate the superiority of our model, a visual comparison of the prediction results was conducted. In the comparison figures, as shown in Figure 7, the left side shows the GT, the middle shows the prediction results of P2PNet model, and the right side shows the prediction results of the proposed model. It can be clearly seen from the figures that the predicted values of the proposed model are closer to the ground truth of the images, especially in dense areas and complex scenes, where it can more accurately capture the distribution characteristics of the crowd.

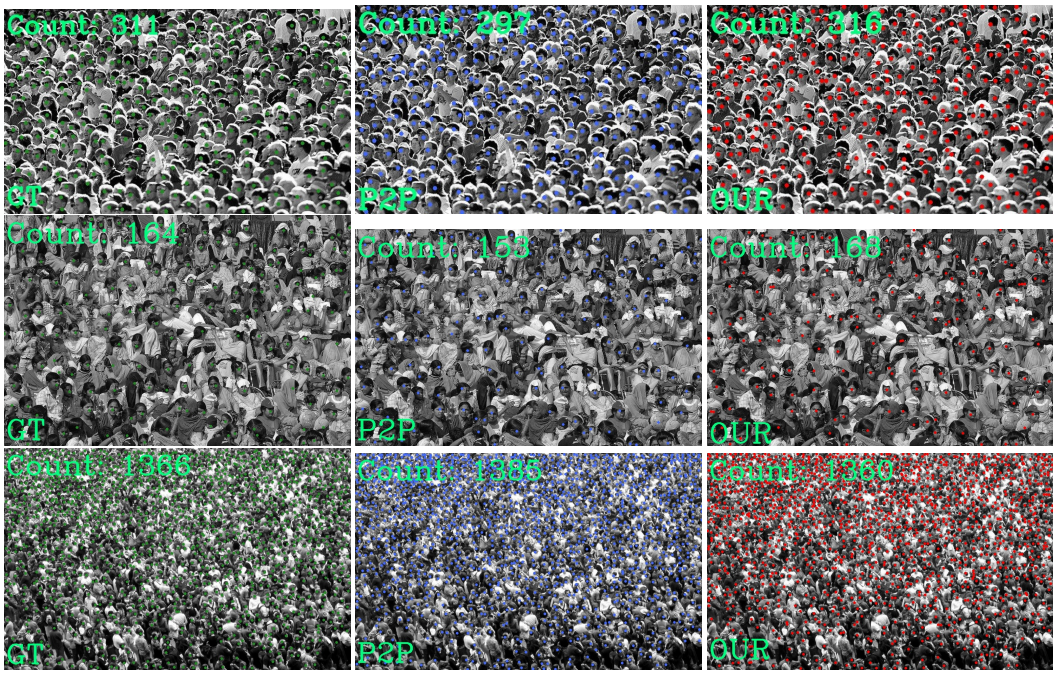


Figure 7. Results comparison on the UCF_CC_50.

The UCF_CC_50 dataset is a complex and challenging benchmark dataset for crowd counting. The crowd distribution in its images is complex and highly dense, making it very suitable for evaluating algorithm performance in the task of counting crowd small objects. This dataset covers crowd scenes ranging from sparse to extremely dense, possessing wide applicability and high difficulty, and can effectively test the generalization ability and robustness of models in different scenes. It can be clearly seen from the comparison figures that the proposed model's prediction results are closer to the ground truth. Compared to the P2PNet model, the proposed model shows higher accuracy in both detailing dense areas and capturing targets in sparse areas. Furthermore, the proposed model's performance is more stable in complex scenes, able to better adapt to changes in crowd distribution, demonstrating stronger generalization ability.

4.3. Ablation Experiments and Analysis

The Feature Adaptive Dynamic Fusion module is referred to as the ADFF module, the module of Multi-scale Feature Aggregation is referred to as Fusion, and the Content-Aware Upsampling module is referred to as CARAFE. The ablation experiments were divided into five groups for comparison: P2PNet, P2PNet+ADFF, P2PNet+Fusion, P2PNet+CARAFE+Fusion, P2PNet+ADFF+CARAFE+Fusion. The results for the ShanghaiTech Part A dataset are shown in Table 5. And we set the learning rate to be 0.0001, and the batch size to be 8.

Table 5. Ablation experiment results on ShanghaiTech Part A.

Method	MAE	MSE
P2PNet	52.74	85.06
P2PNet+ADFF	53.77	82.42
P2PNet+Fusion	52.19	81.99
P2PNet+CARAFE+Fusion	52.02	81.81
P2Pnet+ADFF+CARAFE+Fusion	51.52	81.64
Improvement	1.22 (2.3%)	3.42 (4.02%)

From the results on the ShanghaiTech Part A dataset in Table 5, it can be seen that for the baseline network model based on P2PNet, after adding the module of Feature Adaptive Dynamic Fusion and the module of Multi-scale Feature Aggregation, the model integrates different scales' feature information, improving the detection ability for small targets in dense and occluded scenes, and the values of MAE and MSE are reduced. After adding CARAFE, MAE and MSE are further reduced, effectively lowering the error in small object counting. The ablation experimental results on the ShanghaiTech Part A dataset verify the effectiveness of each module.

4.4. Visual Comparison

The advantage of the proposed method is reflected not only in the overall counting accuracy but also in the handling of small-scale, locally dense areas. In these areas, the P2PNet model counts inaccurately, with phenomena of missed localization and redundant localization, while the proposed model improves this issue by introducing a more effective feature fusion mechanism and multi-scale processing, making the prediction results more refined and accurate. This accuracy demonstrated in local areas, especially when facing highly dense crowds, indicates the advantage of the proposed method in the task of counting crowd small objects. However, in extremely dense areas, the model may overestimate or miss certain small objects due to occlusion or similar visual features in the background.

For a clearer comparison of the results, local enlargements of the test images are shown in Figure 8. The left side shows the test results of P2PNet, and the right side shows the test results of the proposed model. It can be seen that for the same scene, the P2PNet method has missed localization phenomena in dense scenes and severely occluded scenes, as well as redundant localization phenomena. From the figures on the right, it can be seen that the proposed method can accurately count and localize the issues existing in P2PNet. Comprehensively evaluating the metrics and the actual counting and localization results, the proposed method demonstrates superior counting and localization performance for scenes of different densities with severe occlusion on different datasets.

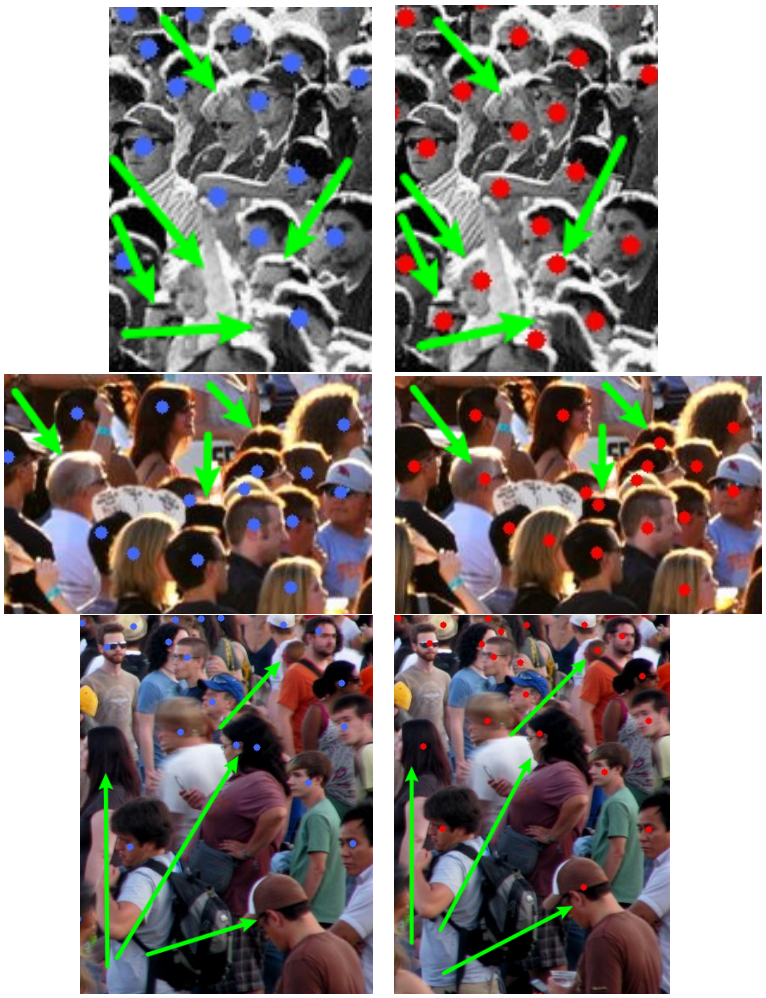


Figure 8. Redundant localization and missed points in occluded scenes.

Through this visual comparison, the performance differences between the two models in practical tasks are not only intuitively demonstrated, but the effectiveness and superiority of the proposed method in the task of counting crowd small objects are further verified. By optimizing the model architecture and introducing more efficient feature learning mechanisms, the proposed method significantly improves the model's robustness and accuracy, especially when processing complex crowd distributions, providing more reliable and precise prediction results. The visual comparison provides strong evidence for the practical application potential of the proposed method, demonstrating its wide applicability in real-world scenarios.

5. Discussion

This paper proposes a novel method for crowd small object counting based on feature-adaptive fusion, aimed at addressing the challenges of target scale variation, background interference, and target occlusion in low-altitude aerial views. The proposed network architecture consists of a front-end backbone network and a back-end feature fusion module. Experimental results on multiple public datasets show that the proposed method achieves excellent generalization and robustness across a range of crowd densities, outperforming state-of-the-art crowd counting algorithms in terms of both accuracy and stability. The method not only enhances small object counting accuracy but also provides robust feature support for subsequent tasks, such as behavior recognition and target tracking.

However, the current approach has some limitations. The model's performance in handling extreme occlusions and highly dynamic environments still requires improvement.

Specifically, accurately counting small objects in rapidly changing or high-density scenes remains challenging. Additionally, while the method performs well on static image data, its real-time application to video or dynamic scenes requires further refinement. Existing methods mainly focus on static image analysis, and applying them effectively to real-time video scenarios necessitates additional research and optimization. Moreover, despite the improvement in accuracy, the model's computational demands may limit its scalability in resource-constrained environments, such as drones or mobile devices.

6. Conclusions

This paper introduces a novel approach for small object counting in crowd scenes within low-altitude aerial views, utilizing a feature-adaptive fusion strategy. The method effectively addresses challenges like scale variation, background interference, and occlusion. Experimental results confirm its superiority over existing methods, demonstrating its potential for robust small object counting in complex environments.

Future work will focus on improving the model's robustness and real-time performance in dynamic environments. By incorporating temporal information and optimizing deep learning network structures, we aim to significantly enhance performance. Additionally, further research into efficient post-processing algorithms will increase the model's applicability in real-time video scenarios.

Another key direction for future research is the development of lightweight models suited for deployment in low-latency, low-power environments, such as drones and mobile terminals. These models would enable efficient, real-time small object counting across a wide range of practical applications, including intelligent surveillance and autonomous driving. The integration of such lightweight models into resource-constrained environments will support the widespread adoption of this technology in low-altitude aerial views and other dynamic real-world settings.

Author Contributions: Conceptualization, J.M.; methodology, J.M. and C.L.; software, L.N. and J.B.; formal analysis, C.L.; resources, L.X. and J.B.; writing—original draft preparation, J. M.; writing—review and editing, C.L.; visualization, L.N. and J.B.; supervision, L.X.; project administration, L.X. and C.L. All authors have read and agreed to the published version of the manuscript.

Funding: This study is supported by QinXin Talents Cultivation Program: Beijing Information Science and Technology University, No. QXTCPB202105.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data that support the findings of this study are available from the corresponding author, upon reasonable request.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Zhang, Y.; Zhou, D.; Chen, S.; Gao, S.; Ma, Y. Single-image crowd counting via multi-column convolutional neural network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 589–597.
2. Wu, B.; Nevatia, R. Detection of multiple, partially occluded humans in a single image by bayesian combination of edgelet part detectors. In Proceedings of the Tenth IEEE International Conference on Computer Vision (ICCV'05), Beijing, China, 17–20 October 2005; Volume 1, pp. 90–97.
3. Jones, M.J.; Snow, D. Pedestrian detection using boosted features over many frames. In Proceedings of the 2008 19th International Conference on Pattern Recognition, Tampa, FL, USA, 8–11 December 2008; pp. 1–4.
4. Dalal, N.; Triggs, B. Histograms of oriented gradients for human detection. In Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), San Diego, CA, USA, 20–26 June 2005; pp. 886–893.

5. Gao, C.; Liu, J.; Feng, Q.; Lv, J. People-flow counting in complex environments by combining depth and color information. *Multimed. Tools Appl.* **2016**, *75*, 9315–9331. [CrossRef]
6. Pham, V.Q.; Kozakaya, T.; Yamaguchi, O.; Okada, R. Count forest: Co-voting uncertain number of targets using random forest for crowd density estimation. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 3253–3261.
7. Viola, V.; Snow, D. Detecting pedestrians using patterns of motion and appearance. In Proceedings of the Ninth IEEE International Conference on Computer Vision, Nice, France, 13–16 October 2003; Volume 2, pp. 734–741.
8. Li, M.; Zhang, Z.; Huang, K.; Tan, T. Estimating the number of people in crowded scenes by mid based foreground segmentation and head-shoulder detection. In Proceedings of the 2008 19th International Conference on Pattern Recognition, Tampa, FL, USA, 8–11 December 2008; pp. 1–4.
9. Zhao, T.; Nevatia, R.; Wu, B. Segmentation and tracking of multiple humans in crowded environments. *IEEE Trans. Pattern Anal. Mach. Intell.* **2008**, *30*, 1198–1211. [CrossRef] [PubMed]
10. Ryan, D.; Denman, S.; Sridharan, S.; Fookes, C. An evaluation of crowd counting methods, features and regression models. *Comput. Vis. Image Underst.* **2015**, *130*, 1–17. [CrossRef]
11. Paragios, N.; Ramesh, V. A MRF-based approach for real-time subway monitoring. In Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition CVPR 2001, Kauai, HI, USA, 8–14 December 2001; p. I-I.
12. Chan, A.B.; Liang, Z.-S.J.; Vasconcelos, N. Privacy preserving crowd monitoring: Counting people without people models or tracking. In Proceedings of the 2008 IEEE Conference on Computer Vision and Pattern Recognition, Anchorage, AK, USA, 23–28 June 2008; pp. 1–7.
13. Lempitsky, V.; Zisserman, A. Learning to count objects in images. *Adv. Neural Inf. Process. Syst.* **2010**, *23*.
14. LeCun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE* **1998**, *86*, 2278–2324. [CrossRef]
15. Fu, M.; Xu, P.; Li, X.; Liu, Q.; Ye, M.; Zhu, C. Fast crowd density estimation with convolutional neural networks. *Eng. Appl. Artif. Intell.* **2015**, *43*, 81–88. [CrossRef]
16. Wang, C.; Zhang, H.; Yang, L.; Liu, S.; Cao, X. Deep people counting in extremely dense crowds. In Proceedings of the 23rd ACM International Conference on Multimedia, Brisbane, Australia, 26–30 October 2015; pp. 1299–1302.
17. Walach, E.; Wolf, L. Learning to count with cnn boosting. In Proceedings of the Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016; Proceedings, Part II 14; Springer: Berlin/Heidelberg, Germany, 2016; pp. 660–676.
18. Babu Sam, D.; Surya, S.; Venkatesh Babu, R. Switching convolutional neural network for crowd counting. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 5744–5752.
19. Sindagi, V.A.; Patel, V.M. Generating high-quality crowd density maps using contextual pyramid cnns. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 1861–1870.
20. Guo, D.; Li, K.; Zha, Z.-J.; Wang, M. Dadnet: Dilated-attention-deformable convnet for crowd counting. In Proceedings of the 27th ACM International Conference on Multimedia, Nice, France, 21–25 October 2019; pp. 1823–1832.
21. Li, Y.; Zhang, X.; Chen, D. Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 1091–1100.
22. Cao, X.; Wang, Z.; Zhao, Y.; Su, F. Scale aggregation network for accurate and efficient crowd counting. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 734–750.
23. Liu, N.; Long, Y.; Zou, C.; Niu, Q.; Pan, L.; Wu, H. Adcrowdnet: An attention-injective deformable convolutional network for crowd understanding. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 3225–3234.
24. Song, Q.; Wang, C.; Jiang, Z.; Wang, Y.; Tai, Y.; Wang, C.; Li, J.; Huang, F.; Wu, Y. Rethinking counting and localization in crowds: A purely point-based framework. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 3365–3374.
25. Wang, B.; Liu, H.; Samaras, D.; Nguyen, M.H. Distribution matching for crowd counting. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 1595–1607.
26. Liu, H.; Zhao, Q.; Ma, Y.; Chong, M.O.; Zhang, Y. Bipartite Matching for Crowd Counting with Point Supervision. In Proceedings of the 30th International Joint Conference on Artificial Intelligence (IJCAI), Virtual, 19–27 August 2021; pp. 860–866.
27. Chen, L.; Yang, T.; Zhang, X.; Zhang, W.; Sun, J. Points as queries: Weakly semi-supervised object detection by points. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual, 19–25 June 2021; pp. 8823–8832.
28. Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; Dai, J. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv* **2020**, arXiv:2010.04159.

29. Idrees, H.; Saleemi, I.; Seibert, C.; Shah, M. Multi-source multi-scale counting in extremely dense crowd images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, 23–28 June 2013; pp. 2547–2554.
30. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* 2014, arXiv:1409.1556.
31. Liu, W.; Salzmann, M.; Fua, P. Context-aware crowd counting. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 5099–5108.
32. Cheng, Z.-Q.; Dai, Q.; Li, H.; Song, J.; Wu, X.; Hauptmann, A.G. Rethinking spatial invariance of convolutional networks for object counting. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 19638–19648.
33. Liang, D.; Xu, W.; Bai, X. An end-to-end transformer model for crowd localization. In Proceedings of the European Conference on Computer Vision (ECCV), Tel Aviv, Israel, 23–27 October 2022; Springer: Cham, Switzerland, 2022; pp. 38–54.
34. Ma, H.-Y.; Zhang, L.; Wei, X.-Y. Fgenet: Fine-grained extraction network for congested crowd counting. In Proceedings of the International Conference on Multimedia Modeling (MMM), Amsterdam, The Netherlands, 29 January–2 February 2024; Springer: Cham, Switzerland, 2024; pp. 43–56.
35. Ma, H.-Y.; Zhang, L.; Shi, S. VMambaCC: A Visual State Space Model for Crowd Counting. *arXiv* **2024**, arXiv:2405.03978. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.