Received: 8 October 2023 | Revised: 28 November 2023 | Accepted: 28 January 2024

DOI: 10.4218/etrij.2023-0426

ETRI Journal WILEY

ORIGINAL ARTICLE

Featured Article

Vehicle counting in drone images: An adaptive method with spatial attention and multiscale receptive fields

Yu Liu | **Hang Shen** | **Tianjing Wang** | **Guangwei Bai**

College of Computer and Information Engineering (College of Artificial Intelligence), Nanjing Tech University, Nanjing, China

Correspondence
Hang Shen, College of Computer and Information Engineering (College of Artificial Intelligence), Nanjing Tech University, Nanjing, China.
Email: hshen@njtech.edu.cn

Funding information
This research was supported by the National Natural Science Foundation of China under grants 61502230 and 61501224, the Natural Science Foundation of Jiangsu Province under grant BK20201357, and the Six Talent Peaks Project in Jiangsu Province under grant RJFW-020.

Abstract

We propose an altitude-adaptive vehicle counting method with an attention mechanism and multiscale receptive fields that optimizes the measurement accuracy and inference latency of unmanned aerial vehicle (UAV) images. An attention mechanism is used to aggregate horizontal and vertical feature weights to enhance spatial information and suppress background noise. The UAV flight altitude and shooting depression angle are considered for scale division and image segmentation to avoid acquiring distance measurements. Based on the dilation rate, we introduce a receptive field selection strategy for the trained model to exhibit scale generalization without redundant calculations. A distribution-aware block loss is optimized via k roots to balance the loss of sparse and crowded regions by dividing the density map. Experiments on three authoritative datasets demonstrate that compared with CSRNet, the proposed method improves the mean absolute error by 29.4%–54.0% and mean squared error by 28.6%–41.2% while reducing the inference latency. The proposed method exhibits higher counting accuracy than lightweight models including MCNN and MobileCount.

KEYWORDS
attention mechanism, distribution awareness, multiscale receptive field, UAV imagery, vehicle counting

1 | INTRODUCTION

Embedded vision systems endow unmanned aerial vehicles (UAVs) with environmental perception capabilities for various applications. For instance, object counting is a critical application that leverages vision. UAVs equipped with embedded vision systems show high maneuverability and flexible deployment. Combined with object counting methods, UAVs have been used in crowd counting, animal counting, vehicle counting, and environmental surveys. For example, UAVs can perform in-air traffic measurements and provide guidance for traffic monitoring. Unlike fixed-ground traffic monitoring, UAVs enable wide-area coverage and real-time tracking in modern transportation [1]. However, owing to computational and memory constraints, UAVs often fail to deploy complex deep learning models for real-time object counting. One natural step is to make the model lightweight. Considering the dynamic deployment and varying flight altitude of UAVs [2], UAV image processing faces unique challenges compared with ground image processing:

1) *Dynamic UAV deployment.* Objects in aerial images are often overwhelmed by complex and variable backgrounds, as shown in Figure 1A. In this case, the model attention may diverge from the regions of interest, thus reducing the counting accuracy [3]. Although methods such as binary masks [4] and data augmentation [5] can minimize background interference, they require additional module support, increasing the computational burden. Although contextual information fusion [6] is a feasible approach, it requires strong associations between the object and background (e.g., the same image background for a similar object), failing to adapt to the continuously changing scenarios of aerial images.
2) *Variations in UAV flight altitude and shooting depression angle.* Even objects of comparable scale may exhibit large differences from hundreds to thousands of pixels after imaging [7], as shown in Figure 1B. To handle scale variations, image pyramid models [8] use multilevel downsampled images as inputs to extract multiscale information. Kirillov and others [9] proposed a feature pyramid that considered shallow representational and deep semantic information for multiscale feature extraction aiming to alleviate scale imbalance. Based on few-shot learning, Liu and others [10] and You and others [11] used self-similarity to enhance feature expression. The enhanced features can carry rich semantics extracted from images and capture support–query relationships. However, feature fusion may require enormous computing and memory resources, hindering direct deployment on UAVs. Sample-dependent dynamic neural networks [12] allow to handle variations in image scales and include dynamic network architectures and parameters. Dynamic architectures activate different layers or branches in processing samples of different scales by early exiting [13], layer skipping [14], and multi-expert ensembles [15] to accelerate inference. However, owing to low confidence, the model often fails to perform early exit or layer skipping, which is equivalent to automatically degenerating into a static neural network with full inference. Dynamic parameter-based solutions include parameter modulation [16], weight prediction [17], and dynamic filtering [18]. These methods improve model representation by transforming model parameters, but designing a front-end network to guide parameter modulation requires additional exploration.
3) *Wide-area coverage.* Compared with images collected during ground monitoring, UAV images cover a wider area. Consequently, the variability in target distribution may lead to severe isolated clustering [19], as shown in Figure 1C. Density map estimation is often inaccurate for densely distributed regions, and isolated points in sparsely distributed target regions are easily overlooked. Using Euclidean loss to train the models fails to effectively reduce the impact of distribution differences [20]. Thus, a loss function must be devised to balance the contributions of losses from dense and sparse regions to the total loss.

Considering the limited computing resources in UAVs and aerial image characteristics, we propose a parameter-adaptive vehicle counting method for UAV images based on an attention mechanism and multiscale receptive fields to achieve dual optimization to improve the counting accuracy and inference latency. The main contributions of this study are as follows:

- A spatial-information-enhanced feature extraction network is constructed. It introduces an attention mechanism to aggregate horizontal and vertical features, thereby enhancing the position representation of the object in the UAV image and mitigating background interference.
- During image preprocessing, the UAV altitude and shooting depression angle are used to dynamically divide the scale and segment the image, thus avoiding

(A) (B) (C)

FIGURE 1 Possible characteristics of unmanned aerial vehicle (UAV) images: (A) Changing backgrounds, (B) large scale variations, and (C) uneven distribution.

the need for distance measurements. A switch function allows to adjust convolution dilation parameters upon receiving the split shots. In addition, receptive field selection based on the dilation rate is applied for the trained model to generalize the scale without adding redundant calculations.
- By improving the Euclidean loss via k roots, a distribution-aware block loss is built to balance the proportion of losses in regions with different object distributions in the total loss, thereby reducing adverse effects of object distribution variability on counting.

Experimental results on multiple authoritative datasets confirm that the proposed method substantially reduces the model inference latency while achieving comparable counting accuracy compared with robust static models, including VGG [21] and ResNet [22]. Moreover, compared with state-of-the-art lightweight models, the proposed method shows a higher counting accuracy and similar inference latencies.

2 | PROPOSED COUNTING METHOD

This section describes the design and implementation of the proposed vehicle counting method, whose architecture is shown in Figure 2. It contains three interdependent modules that conform to the following workflow:

- *Multiscale receptive fields* estimate scales, crop the image, and determine the convolution dilation rate.
- *A spatial-information-enhanced feature extraction network* extracts features with spatial information after receiving the cropped images, and parameter-tuned convolutions process the extracted features.
- *A distribution-aware block loss* is applied in backpropagation to balance the loss proportion in a predicted density map.

2.1 | Spatial-information-enhanced feature extraction

MobileNetV3 [23] is a lightweight feature extraction network suitable for deployment in devices with limited computing resources. To enhance the spatial information of UAV images, we improve the attention mechanism in MobileNetV3 considering horizontal and vertical weights of extracted features using the architecture shown in Figure 3.

First, input feature map r with dimensions $C \times G \times W$ undergoes horizontal and vertical compression represented by corresponding functions. Compression operations generate channel-wise statistics in the spatial

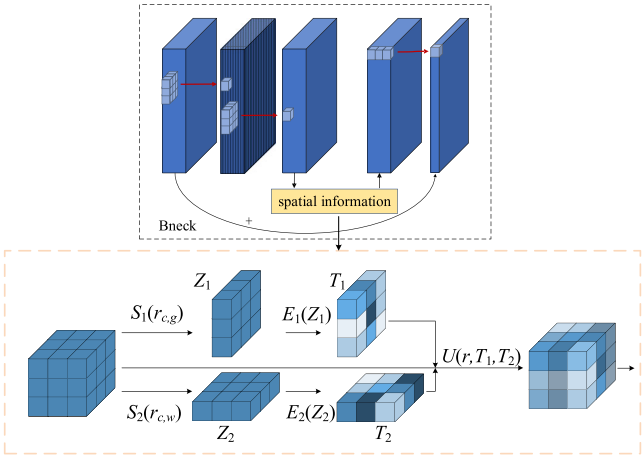

FIGURE 3 Attention enhancement by spatial information.

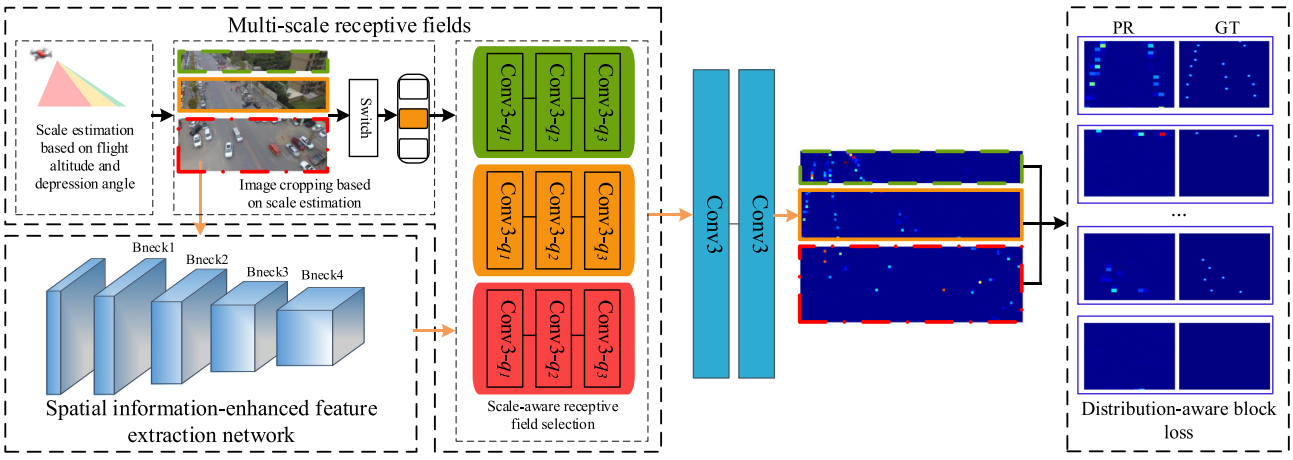

FIGURE 2 Object counting architecture using unmanned aerial vehicle (UAV) images.

dimensions using global average pooling layers. The output of the c-th channel at height g is calculated as follows:

$$z_{c,g} = S_1(r_{c,g}) = \frac{1}{W}\sum_{i=1}^{W} r_{c,g}(i). \quad (1)$$

Similarly, the output of the c-th channel with width w is derived as follows:

$$z_{c,w} = S_2(r_{c,w}) = \frac{1}{G}\sum_{j=1}^{G} r_{c,w}(j). \quad (2)$$

After compression, the generated vertical and horizontal features, Z_1 and Z_2, respectively, are expressed as follows:

$$\begin{cases} Z_1 = \{z_{c,g} | c = 1, 2, \cdots, C, g = 1, 2, \cdots, G\}, \\ Z_2 = \{z_{c,w} | c = 1, 2, \cdots, C, w = 1, 2, \cdots, W\}, \end{cases} \quad (3)$$

where Z_1 and Z_2 have dimensions $C \times G \times 1$ and $C \times 1 \times W$, respectively.

Then, intermediate results Z_1 and Z_2 are fed into activation functions $E_1(\cdot)$ and $E_2(\cdot)$ to generate vertical weights T_1 and horizontal weights T_2 as follows:

$$\begin{cases} T_1 = E_1(Z_1) = \sigma(\delta(F_1(Z_1))), \\ T_2 = E_2(Z_2) = \sigma(\delta(F_2(Z_2))), \end{cases} \quad (4)$$

where $F_1(\cdot)$ and $F_2(\cdot)$ are 1×1 convolution operations, δ represents batch normalization, and $\sigma(\cdot)$ is formulated as follows [24]:

$$\sigma(x) = \begin{cases} 0, & \text{if } x \le -3, \\ \dfrac{x(x+3)}{6}, & \text{if } -3 < x < 3, \\ x, & \text{if } x \ge 3. \end{cases} \quad (5)$$

Unlike traditional activation functions that focus on boosting accuracy, (5) balances the computation time and counting accuracy by enhancing the model nonlinear expression.

Finally, r is reweighted by T_1 and T_2 as follows:

$$U(r, T_1, T_2) = r \times T_1 \times T_2. \quad (6)$$

This aggregates object position information in the next layer.

2.2 | Multiscale receptive field strategy

Unlike a model with a fixed receptive field size that can only adapt to one scale [25], we use a multiscale receptive field to estimate the scale range according to the UAV flight altitude and depression angle, crop the UAV image via scale estimation, and select the convolution dilation rate for scale matching.

2.2.1 | Simple scale estimation

Typically, the scale of an object in an image captured from a UAV is determined by the distance between the object and camera. Scale calculations typically rely on distance measurements. Common methods include binocular cameras [26] and deep learning distance measurements [27], which require high computing power. This section presents the design of a simple scale estimation method to determine the object scale according to the UAV flight altitude and depression angle without requiring a dedicated distance measurement module, thereby reducing the computational burden.

Considering the point where a UAV is projected onto the ground as the origin, the ground as the abscissa, and the direction from the intersection of the upline-of-sight and ground to the origin as the positive direction of the abscissa, we establish a Cartesian coordinate system for scale estimation, as illustrated in Figure 4, where the scales in the image differ at different UAV altitudes. The altitude range is divided into three levels to facilitate processing. For a drone hovering at altitude h, d_1 and d_2 represent low- and high-altitude thresholds, respectively. Hence, $h \le d_1$ ($h \ge d_2$) indicates that the drone captures images at a low (high) altitude. If $d_1 \le h \le d_2$, the UAV is at a medium flight altitude. The coordinates of ground objects with distances to d_1 and d_2 from the UAV are $(\pm e_1, 0)$ and $(\pm e_2, 0)$, respectively, where $e_1 = \sqrt{d_1^2 - h^2}$ and $e_2 = \sqrt{d_2^2 - h^2}$.

Consider Figure 4B. The coordinates of the downline-of-sight and upline-of-sight intersections with the ground are $(\tau_1, 0)$ and $(\tau_2, 0)$, respectively, with

$$\tau_1 = \begin{cases} -\sqrt{\left(\frac{h}{\cos\left(90^\circ - \frac{\alpha}{2} - \beta\right)}\right)^2 - h^2}, & \text{if } 90^\circ - \frac{\alpha}{2} - \beta < 0 \\ \sqrt{\left(\frac{h}{\cos\left(90^\circ - \frac{\alpha}{2} - \beta\right)}\right)^2 - h^2}, & \text{otherwise} \end{cases} \quad (7)$$

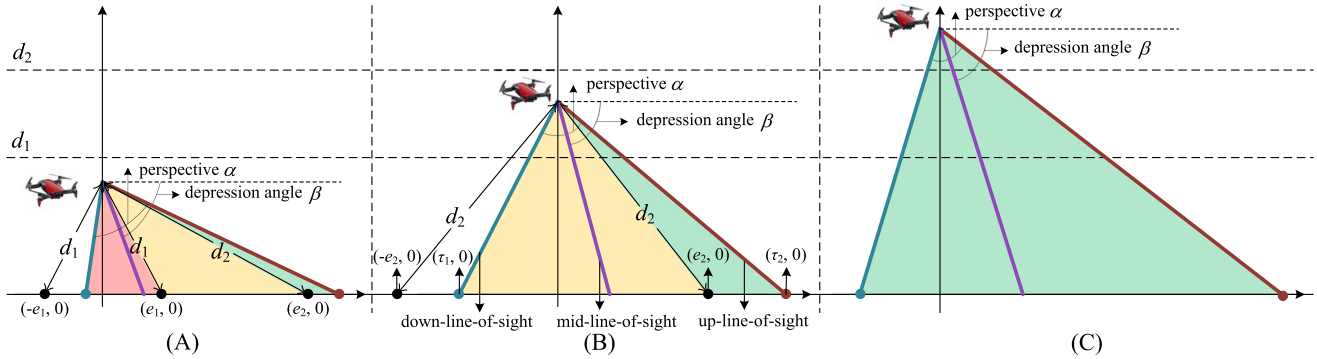

FIGURE 4 Scale distribution under varying unmanned aerial vehicle (UAV) flight altitude and depression angle.

TABLE 1 Scale estimation rules according to UAV flight altitude and depression angle.

Flight altitude level	Range of h	Range of x	Scale	Angle range
Low	$(0, d_1]$	$(\tau_1, \min\{e_1, \tau_2\})$	Small	Red in Figure 4A
		$(\max\{e_1, \tau_1\}, \min\{e_2, \tau_2\})$	Medium	Yellow in Figure 4A
		$(\max\{e_2, \tau_1\}, \tau_2)$	Large	Green in Figure 4A
Medium	(d_1, d_2)	$(\max\{-e_2, \tau_1\}, \min\{e_2, \tau_2\})$	Medium	Yellow in Figure 4B
		$(\max\{e_2, \tau_1\}, \tau_2)$	Large	Green in Figure 4B
High	$[d_2, +\infty)$	(τ_1, τ_2)	Large	Green in Figure 4C

Abbreviation: UAV, unmanned aerial vehicle.

and

$$\tau_2 = \sqrt{\left(\frac{h}{\cos\left(90^\circ + \frac{\alpha}{2} - \beta\right)}\right)^2 - h^2}. \tag{8}$$

In (7) and (8), α and β represent the perspective and depression angle, respectively.

For an object located at $(x, 0)$, the UAV can calculate its scale according to h and β. The calculation rules for different cases are listed in Table 1. Consider a UAV flying at $h \in (0, d_1)$. The images captured by the UAV may contain small, medium, and large scales, which correspond to the red, yellow, and green areas, respectively, in Figure 4A.

2.2.2 | Image cropping by scale estimation

Consider Figure 4A, which illustrates all the possible scales, to explain image segmentation according to the estimated scale. The proportions of the red, yellow, and green areas at angle α are used to calculate the ratios for small-, medium-, and large-scale UAV images, respectively. Given flight altitude h and depression angle β, the values of these proportions are listed in Table 2.

For clarity, Figure 5 shows the angle ranges and ground reference point coordinates at different scales corresponding to Figure 4A, where $h \in (0, d_1)$, $\tau_1 \in (-e_1, e_1)$, and $\tau_2 \in (e_2, +\infty)$. If an object in an image falls within (τ_1, e_1), it is considered a small-scale object. Similarly, objects falling in (e_1, e_2) and (e_2, τ_2) are considered as medium- and large-scale objects, respectively. Let ω_1, ω_2, and ω_3 denote the proportions for the small, medium, and large scales, respectively. Based on the angle at each scale range, we have

$$\begin{cases} \omega_1 = \dfrac{\arccos\dfrac{h}{d_1} - 90^\circ + \dfrac{\alpha}{2} + \beta}{\alpha}, \\ \omega_2 = \dfrac{\arccos\dfrac{h}{d_2} - \arccos\dfrac{h}{d_1}}{\alpha}, \\ \omega_3 = \dfrac{90^\circ - \arccos\dfrac{h}{d_2} - \beta + \dfrac{\alpha}{2}}{\alpha}. \end{cases} \tag{9}$$

Using (10), the UAV image is cropped, and the image patches are input to a feature extraction network and switch function for adjusting the convolution parameters.

TABLE 2 Image cropping rules under varying UAV flight altitude and depression angle.

Flight altitude and depression angle			Proportions of three scales		
Range of h	**Range of τ_1**	**Range of τ_2**	**Small scale (ω_1)**	**Medium scale (ω_2)**	**Large scale (ω_3)**
$(0, d_1]$	$(-e_1, e_1)$	$(-e_1, e_1)$	1	0	0
	(e_1, e_2)	(e_1, e_2)	0	1	0
	$(e_2, +\infty)$	$(e_2, +\infty)$	0	0	1
	$(-e_1, e_1)$	(e_1, e_2)	$\left(\arccos\frac{h}{d_1} - 90^\circ + \frac{\alpha}{2} + \beta\right)/\alpha$	$\left(90^\circ + \frac{\alpha}{2} - \beta - \arccos\frac{h}{d_1}\right)/\alpha$	0
	$(-e_1, e_1)$	$(e_2, +\infty)$	$\left(\arccos\frac{h}{d_1} - 90^\circ + \frac{\alpha}{2} + \beta\right)/\alpha$	$\left(\arccos\frac{h}{d_2} - \arccos\frac{h}{d_1}\right)/\alpha$	$\left(90^\circ - \arccos\frac{h}{d_2} - \beta + \frac{\alpha}{2}\right)/\alpha$
	(e_1, e_2)	$(e_2, +\infty)$	0	$\left(\arccos\frac{h}{d_2} - 90^\circ + \frac{\alpha}{2} + \beta\right)/\alpha$	$\left(90^\circ + \frac{\alpha}{2} - \beta - \arccos\frac{h}{d_1}\right)/\alpha$
(d_1, d_2)	$(-e_2, e_2)$	$(-e_2, e_2)$	0	1	0
	$(e_2, +\infty)$	$(e_2, +\infty)$	0	0	1
	$(-e_2, e_2)$	$(e_2, +\infty)$	0	$\left(\arccos\frac{h}{d_2} - 90^\circ + \frac{\alpha}{2} + \beta\right)/\alpha$	$\left(90^\circ + \frac{\alpha}{2} - \beta - \arccos\frac{h}{d_1}\right)/\alpha$
$[d_2, \infty)$	$(-\infty, +\infty)$	$(-\infty, +\infty)$	0	0	1

Abbreviation: UAV, unmanned aerial vehicle.

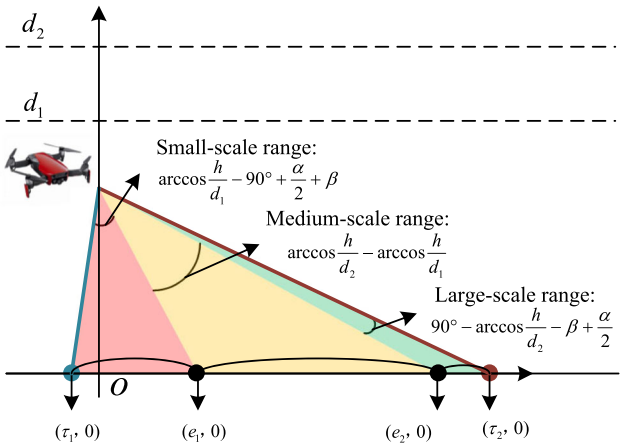

FIGURE 5 Angles for different scales.

$$\begin{cases} \omega_1 = \dfrac{\arccos\dfrac{h}{d_1} - 90^\circ + \dfrac{\alpha}{2} + \beta}{\alpha}, \\ \omega_2 = \dfrac{\arccos\dfrac{h}{d_2} - \arccos\dfrac{h}{d_1}}{\alpha}, \\ \omega_3 = \dfrac{90^\circ - \arccos\dfrac{h}{d_2} - \beta + \dfrac{\alpha}{2}}{\alpha}. \end{cases} \tag{10}$$

2.2.3 | Scale-aware receptive field selection

A switch function allows to determine the convolution dilation rate based on the estimated scale. Accordingly, we introduce a scale-aware selection strategy to select the optimal receptive field.

Three adjustable dilated convolutions are stacked to obtain multiscale receptive fields. Their dilation rates are represented as triplets $Q = (q_1, q_2, q_3)$, where the three elements are required to use every pixel and avoid the grid effect [28]. In the block for scale-aware receptive field selection shown in Figure 2, multiple values for Q are set, with the receptive fields that cover small, medium, and large scales being indicated in red, yellow, and green, respectively. When the scale of an input image is identified, the switch function guides the convolution to adjust the dilation rate accordingly.

Dynamically selecting the receptive fields allows the model to focus on objects at a specific scale and ensures that training data at all scales can be learned. The trained parameters have known scale-invariant features, enabling the model to accurately detect different scale instances of the same class. In addition, using the switch function to adjust the model parameters allows to maintain the single-column model structure. Compared with multicolumn structures [29], the number of model parameters is reduced, and compared with multibranch structures [19], redundant convolutions are omitted in mismatching scale branches.

Depending on the UAV flight altitude and depression angle, scale estimation uses a scale range for image cropping. Receptive field selection adjusts the `dilation rate to adapt to scale variations, thereby expanding the receptive field coverage without increasing the model size.

2.3 | Distribution-aware block loss

We establish a distribution-aware block loss function to handle fluctuations in the spatial distribution of an object.

Figure 6 shows a diagram of the loss calculation. First, the predicted and ground-truth density maps are evenly divided into m blocks, $A_i (i \in 1, 2, ..., m)$, to establish block-level constraints and roughly distinguish dense and sparse regions. Then, the Euclidean distance loss and counting loss of $A_i (i \in 1, 2, ..., m)$ are set to the $1/k$ power, thereby suppressing blocks with dense objects while emphasizing those with sparse objects. Finally, the reweighted losses of $A_i (i \in 1, 2, ..., m)$ are summed to obtain the final loss of the density maps for model training.

Let $F(X_{i,j};\Theta)$ and $F_{i,j}^{GT}$ denote the predicted and ground-truth density maps of the j-th block in image i, respectively, and $V_{i,j}^{PR}$ and $V_{i,j}^{GT}$ denote the predicted and ground-truth counts of the j-th block in image i, respectively. The improved loss function is expressed as

$$L(\Theta) = \frac{1}{2N}\sum_{i=1}^{N}\sum_{j=1}^{m}\sqrt[k]{\left\|F(X_{i,j};\Theta) - F_{i,j}^{GT}\right\|_2^2 + \lambda\left(V_{i,j}^{PR} - V_{i,j}^{GT}\right)}, \tag{11}$$

where N is the number of samples, Θ represents the training parameters, and λ is the weight of the counting loss. The optimal value of k is determined experimentally, as detailed in Section 3.

3 | EXPERIMENTS AND RESULTS

We conducted evaluations implemented in PyTorch using the Adam optimizer and a learning rate of 1×10^{-5}. The model parameters were randomly initialized from a Gaussian distribution with zero mean and standard deviation of 0.01. We also set λ in (11) to 0.1,

We adopted the mean absolute error (MAE) and mean squared error (MSE) as evaluation metrics to determine the accuracy and robustness of vehicle counting. The metrics are defined as

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}\left|V_i^{PR} - V_i^{GT}\right|, \tag{12}$$

$$\text{MSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(V_i^{PR} - V_i^{GT}\right)^2}, \tag{13}$$

where N is the number of images and V_i^{GT} and V_i^{PR} are the ground-truth and predicted counts from image i, respectively.

Table 3 lists the network parameters used for feature extraction. Conv2d denotes a 3×3 convolutional layer. Each Bneck describes an operation with n identical layers, stride s, channel expansion ratio t, output channel c, and application or not of spatial-aware attention. Considering parameters of commercial UAVs (e.g., Da-Jiang Innovations—DJI models), we set perspective angle α to 82°. Nevertheless, we listed some possible flight altitudes and depression angles and assigned suitable parameters to each image by observation.

Comparative experiments were conducted on the VisDrone2019 Vehicle dataset [19] to determine the optimal value of k for the distribution-aware block loss. As shown in Figure 7, the MAE and MSE were minimal at $k = 4$. This k value was used in the subsequent experiments. As shown in Figure 8, we conducted experiments on the VisDrone2019 Vehicle dataset to obtain optimal Q settings for the various scales. The MAE and MSE were minimal for $Q = (1, 1, 1), (1, 2, 3), (3, 4, 5)$.

Table 4 lists the object categorization into scales according to pixel size based on the definitions in [25]. In addition, d_1 and d_2 were set to 40 m and 100 m such that the scale of the objects directly below the UAV

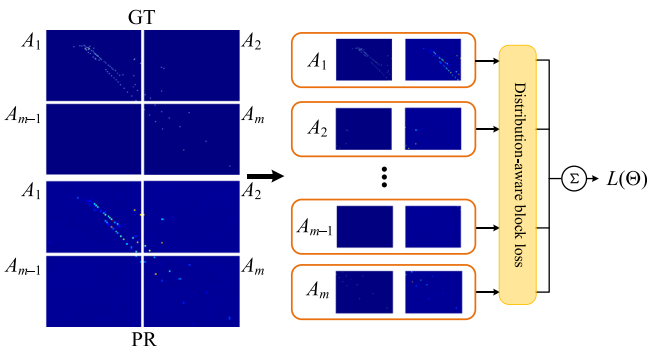

FIGURE 6 Architecture to calculate distribution-aware block loss.

TABLE 3 Structure of feature extraction network.

Operator	c	t	n	s	Attention
Image	3	–	–	–	–
Conv2d	16	–	1	1	–
MaxPool	16	–	1	2	–
Bneck1	48	1	4	1	✓
Bneck2	64	6	4	2	✓
Bneck3	128	6	4	1	✓
Bneck4	160	6	4	2	✓

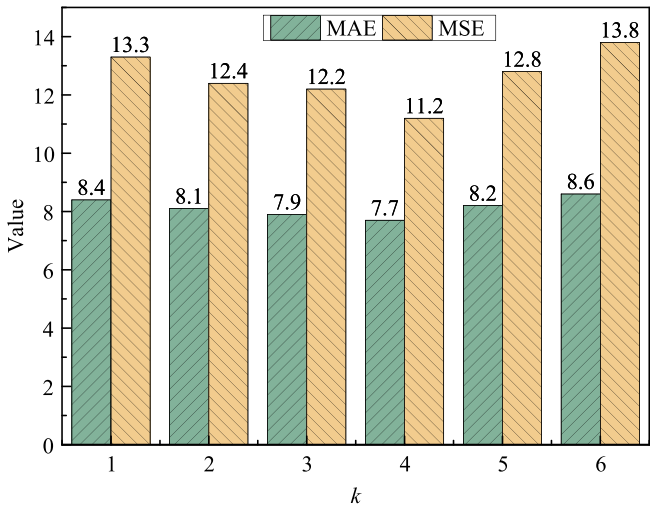

FIGURE 7 Mean absolute error (MAE) and mean squared error (MSE) for different values of k.

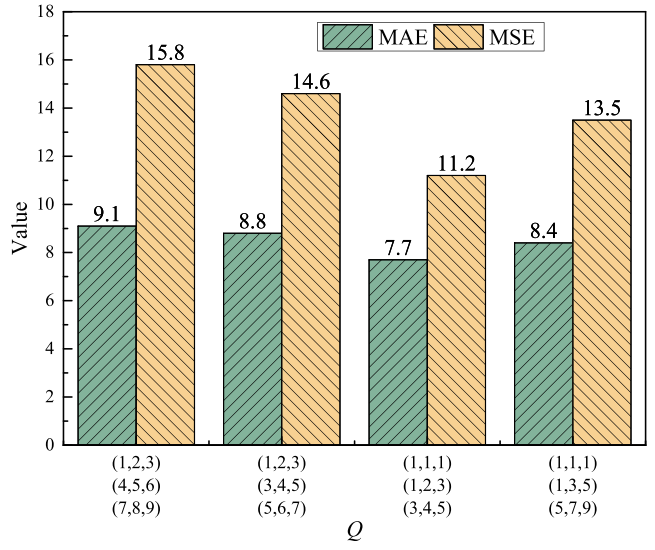

FIGURE 8 Mean absolute error (MAE) and mean squared error (MSE) for different values of Q at (large, medium, small) scales.

TABLE 4 Q values according to scale.

Size (pixels)	Scale	Q
$<32 \times 32$	Large	(1, 1, 1)
$32 \times 32 - 96 \times 96$	Medium	(1, 2, 3)
$>96 \times 96$	Small	(3, 4, 5)

conformed to the above definition when the UAV altitude changed. Three sets of values were configured as optional parameters for the dilation rate to evaluate receptive field selection. Consecutive dilation rates allowed the model to smoothly adapt to scale variations.

Comparative experiments were conducted on multiple authoritative datasets using ablation studies to demonstrate the contribution of each module to the proposed method.

3.1 | Results on different datasets

CARPK and PUCPR+ [30] are vehicle counting datasets containing images of nearly 90 000 cars captured using UAVs from different parking lots. The maximum and minimum numbers of cars per image are 331 and 0 in crowded and sparse scenes, respectively.

Table 5 compares the proposed method with other approaches on the CARPK and PUCPR+ datasets. Compared with the LMSFFNet lightweight multiscale fusion model, the proposed method reduces the MAE and MSE by 10.5% and 2.9% on PUCPR+, respectively, with a lower model complexity. This confirms that the multiscale receptive-field strategy alleviates the negative impact of scale mismatch under slight scale variations. Although the MAE is reduced by 5.1% on CARPK, the MSE of the proposed method is inferior to that of LMSFFNet, indicating inevitable fluctuations in the model predictions when interference exists. HLCNN outperforms our method in terms of MAE and MSE. Theoretically, the model volume positively correlates with the counting accuracy, but a complex model cannot be directly deployed on a UAV. Instead, we aim to achieve high-precision counting under model size constraints.

Predicted density maps for samples from the CARPK dataset are shown in Figure 9A,B, with accurate results being achieved for crowded and sparse scenarios. In Figure 9C, vehicles occluded by shadows are not fully recovered in the predicted density map, which partially explains the slightly higher MSE of our method compared with LMSFFNet. Predictions on samples from the PUCPR+ dataset shown in Figure 10 exhibit similar patterns.

The VisDrone2019 Vehicle dataset contains cars, trucks, and busses for vehicle counting and covers severe occlusions and truncations. UAVs captured the images at varying altitudes and depression angles for wide coverage in this dataset.

Table 6 lists the results for the proposed and state-of-the-art methods on the VisDrone2019 Vehicle dataset. The MAE of the proposed method is slightly higher than that of the top-performing VCNet. Nevertheless, owing to depth-wise separable convolutions in MobileNetV3, fewer parameters and computations are involved in the proposed method than if using conventional convolutions. With a comparable model complexity, the

TABLE 5 Counting performance comparison on CARPK and PUCPR+ datasets.

Method	No. parameters ($\times 10^6$)	GFlops[a]	CARPK		PUCPR+	
			MAE	MSE	MAE	MSE
CSRNet [29]	16.3	325.3	11.5	13.3	8.7	10.2
LPN [30]	16.2	325.3	23.8	36.8	22.8	24.5
PSGCNet [31]	27.5	385.7	8.1	10.5	5.2	7.4
HLCNN [32]	14.8	308.1	2.1	3.0	2.5	3.4
MCNN [33]	0.1	21.2	39.1	43.3	21.9	29.5
LMSFFNet [34]	4.5	14.9	7.0	9.0	4.5	6.2
Proposed	1.6	16.1	6.6	9.5	4.0	6.0

Abbreviations: MAE, mean absolute error; MSE, mean squared error.
[a]GFlops, millions of floating-point operations.

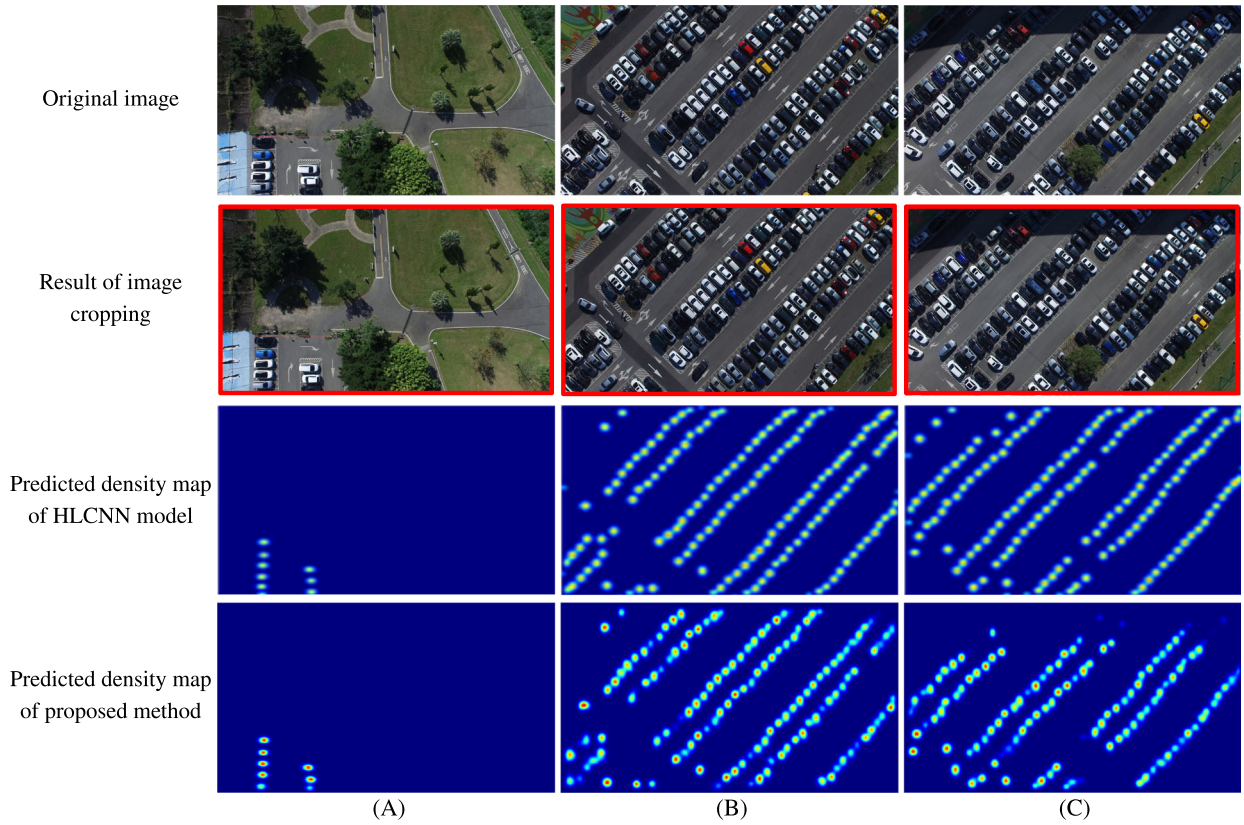

FIGURE 9 Visualization analysis on three samples (columns A–C) from CARPK dataset (first row, original image; second row, result of image cropping; third row, predicted density map of HLCNN model; fourth row, predicted density map of proposed method).

proposed method reduces the MAE and MSE by 13.5% and 18.9% compared with MobileCount, respectively. The computational complexity and counting accuracy comparison between the complex and lightweight models in Tables 5 and 6 confirm that the proposed method suitably balances the model size and counting accuracy.

Figure 11 shows predictions of samples from the VisDrone2019 Vehicle dataset. In Figure 11A, the attention module suppresses distracting backgrounds, and the vehicles are accurately localized. Figure 11B shows extreme scale variations that are effectively handled by dividing the image into blocks and dynamically adjusting the expansion ratios for optimal receptive fields to generate high-quality density maps. In Figure 11C, the predicted density map accurately reflects the sparse distribution despite a severe density discrepancy, thereby validating the proposed loss function.

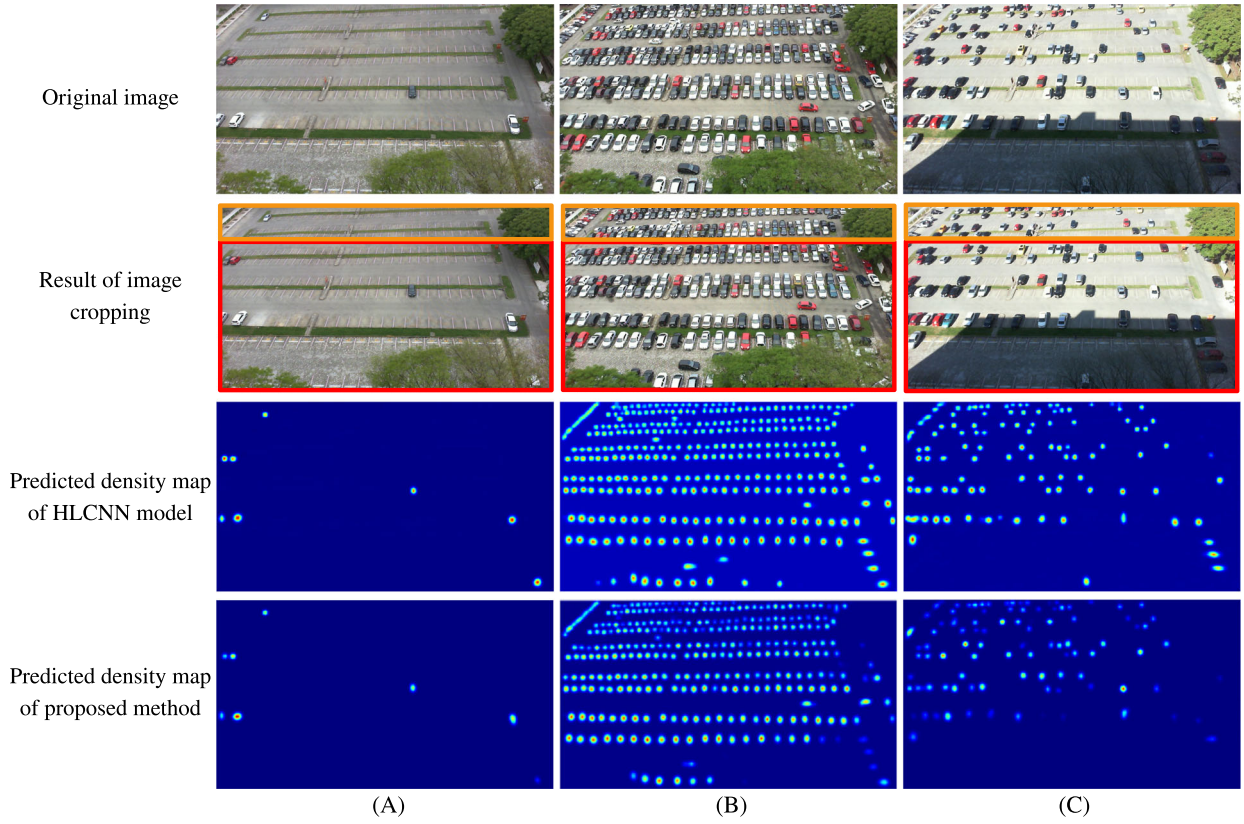

FIGURE 10 Visualization analysis on three samples (columns A–C) from PUCPR+ dataset (first row, original image; second row, result of image cropping; third row, predicted density map of HLCNN model; fourth row, predicted density map of proposed method).

TABLE 6 Counting performance comparison on VisDrone2019 Vehicle dataset.

Method	No. parameters ($\times 10^6$)	GFlops	MAE	MSE
CSRNet [29]	16.3	325.3	10.9	16.6
SACANet [19]	17.6	396.7	8.6	12.9
VCNet [35]	30.0	495.0	2.8	5.7
MCNN [33]	0.1	21.2	14.9	21.6
MobileCount [36]	3.4	12.3	8.9	13.8
Proposed	1.6	16.1	7.7	11.2

Abbreviations: MAE, mean absolute error; MSE, mean squared error.

3.2 | Ablation experiments

Ablation experiments were conducted on VisDrone2019 Vehicle and CARPK to analyze the contribution of each component to prediction. As listed in Table 7, substantial improvements in MAE and MSE (33.6% and 38.1% lower for the proposed method than for the MobileNetV3 baseline, respectively) validate the effectiveness of our complete method. Removing the spatially enhanced attention increases the MAE and MSE by 15.2% and 18.2%, respectively. Excluding the multiscale receptive field module leads to even higher MAE and MSE (by 40.3% and 45.5%, respectively). Without the object-distribution-aware block loss, the MAE and MSE increase by 6.5% and 7.1%, respectively. Similar patterns are observed for the CARPK dataset.

The ablation results confirm the superiority and effectiveness of the proposed method and its components. Multiscale receptive fields notably contribute to the model by endowing it with scale awareness and adaptive tuning.

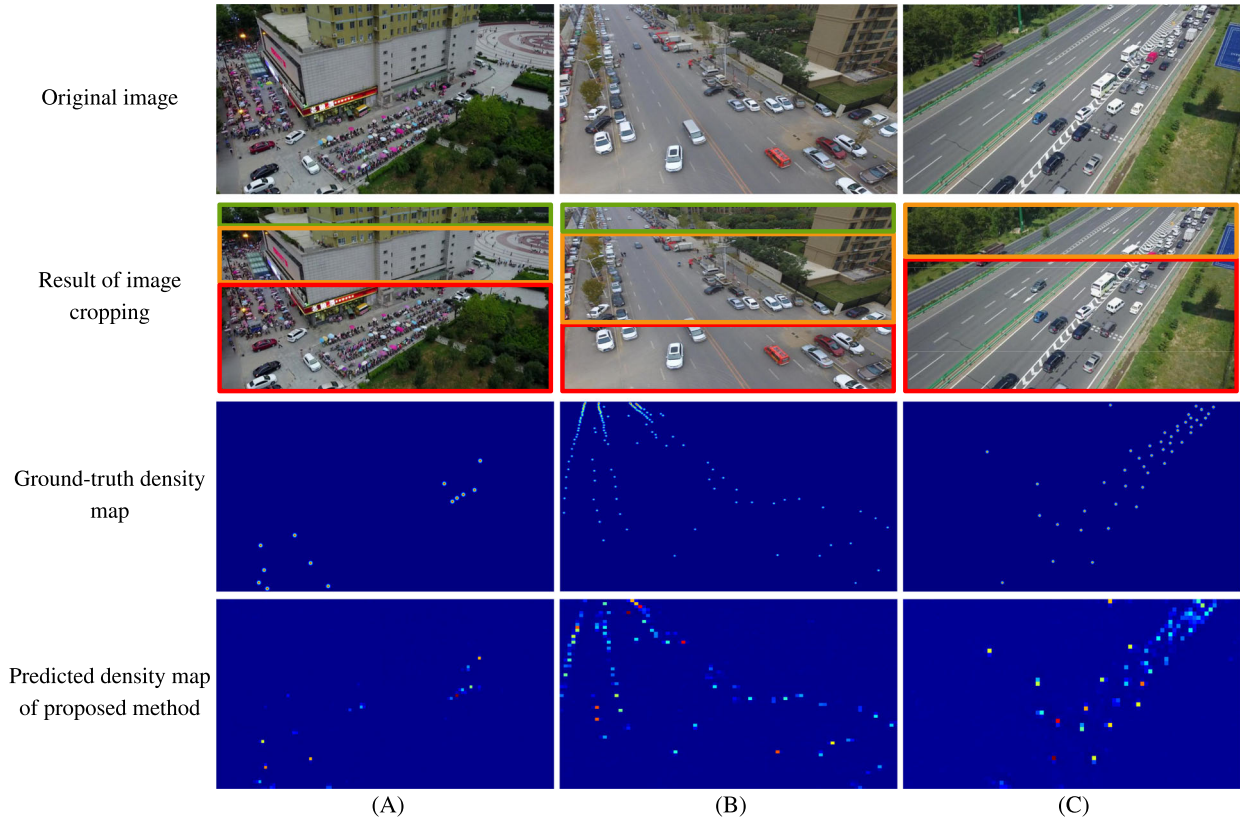

FIGURE 11 Visualization analysis on three samples (columns A–C) from VisDrone2019 Vehicle dataset (first row, original image; second row, result of image cropping; third row, ground-truth density map; fourth row, predicted density map of proposed method).

TABLE 7 Comparison of results from ablation experiments.

	CARPK		VisDrone2019 vehicle	
Method	**MAE**	**MSE**	**MAE**	**MSE**
MobileNet V3	10.1	14.3	11.6	18.1
W/o spatial information-enhanced attention	7.6	11.1	8.7	13.7
W/o receptive field selection	9.5	13.1	10.8	16.3
W/o distribution-aware block loss	7.1	10.4	8.2	12.0
Proposed	6.6	9.5	7.7	11.2

Abbreviations: MAE, mean absolute error; MSE, mean squared error.

4 | CONCLUSION

We present a vehicle counting method using UAV images based on an attention mechanism and multiscale receptive fields. Our method improves the counting accuracy of UAV images and reduces the computational burden through receptive field adaptation. In addition, the proposed method achieves a counting accuracy close to that of powerful static models while substantially reducing the model inference latency. Compared with state-of-the-art lightweight models, the proposed method counts more accurately and has a similar inference latency. The scenario described in this study is reproducible and general, and our method is not limited by the UAV flight altitude and shooting depression angle. In fact, the proposed method is scalable and can be applied to object detection in large-scale changing scenes. Nevertheless, darkness, lousy weather, and other conditions inevitably aggravate missing counts, which will become the focus of future research.

CONFLICT OF INTEREST STATEMENT

The authors declare that there are no conflicts of interest.

ORCID

Hang Shen https://orcid.org/0000-0002-8804-2787

REFERENCES

1. H. Shen, Y. Tian, T. Wang, and G. Bai, *Slicing-based task offloading in space-air-ground integrated vehicular networks*, IEEE Trans. Mobile Comput. **23** (2024), no. 5, 4009–4024.
2. H. Shen, Q. Ye, W. Zhuang, W. Shi, G. Bai, and G. Yang, *Drone-small-cell-assisted resource slicing for 5G uplink radio access networks*, IEEE Trans. Vehic. Technol. **70** (2021), no. 7, 7071–7086.
3. J. Wei, S. Wang, and Q. Huang, *F^3Net: fusion, feedback and focus for salient object detection*, Proc. AAAI Conf. Artif. Intel. **34** (2020), no. 7, 12321–12328.
4. X. Jiang, L. Zhang, M. Xu, T. Zhang, P. Lv, B. Zhou, X. Yang, and Y. Pang, *Attention scaling for crowd counting*, (Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA), 2020, pp. 4706–4715.
5. C. Duan, Z. Wei, C. Zhang, S. Qu, and H. Wang, *Coarse-grained density map guided object detection in aerial images*, (Proc. of the IEEE/CVF International Conference on Computer Vision, Montreal, Canada), 2021, pp. 2789–2798.
6. W. Liu, M. Salzmann, and P. Fua, *Context-aware crowd counting*, (Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA), 2019, pp. 5099–5108.
7. G. Gao, Q. Liu, and Y. Wang, *Counting from sky: a large-scale data set for remote sensing object counting and a benchmark method*, IEEE Trans. Geosci. Remote Sens. **59** (2020), no. 5, 3642–3655.
8. M. Najibi, B. Singh, and L. S. Davis, *Autofocus: efficient multi-scale inference*, (Proc. of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea), 2019, pp. 9745–9755.
9. A. Kirillov, R. Girshick, K. He, and P. Dollár, *Panoptic feature pyramid networks*, (Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA), 2019, pp. 6399–6408.
10. C. Liu, Y. Zhong, A. Zisserman, and W. Xie, *CounTR: Transformer-based generalised visual counting*, arXiv preprint, 2022, DOI 10.48550/arXiv.2208.13721.
11. Z. You, K. Yang, W. Luo, X. Lu, L. Cui, and X. Le, *Few-shot object counting with similarity-aware feature enhancement*, (Proc. of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA), 2023, pp. 6315–6324.
12. Y. Han, G. Huang, S. Song, L. Yang, H. Wang, and Y. Wang, *Dynamic neural networks: a survey*, IEEE Trans. Pattern Anal. Machine Intel. **44** (2021), no. 11, 7436–7456.
13. G. Huang, D. Chen, T. Li, F. Wu, L. Van Der Maaten, and K. Q. Weinberger, *Multi-scale dense networks for resource efficient image classification*, arXiv preprint, 2017, DOI 10.48550/arXiv.1703.09844.
14. A. Veit and S. Belongie, *Convolutional networks with adaptive inference graphs*, (Proc. of the European Conference on Computer Vision (ECCV), Munich, Germany), 2018, pp. 3–18.
15. D. Eigen, M. Ranzato, and I. Sutskever, *Learning factored representations in a deep mixture of experts*, arXiv preprint, 2013, DOI 10.48550/arXiv.1312.4314.
16. Y. Chen, X. Dai, M. Liu, D. Chen, L. Yuan, and Z. Liu, *Dynamic convolution: attention over convolution kernels*, (Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA), 2020, pp. 11030–11039.
17. M. Denil, B. Shakibi, L. Dinh, M. Ranzato, and N. De Freitas, *Predicting parameters in deep learning*, Adv. Neural Inform. Process. Syst. **26** (2013), 2148–2156.
18. J. Hu, L. Shen, and G. Sun, *Squeeze-and-excitation networks*, (Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake Sity, UT, USA), 2018, pp. 7132–7141.
19. H. Bai, S. Wen, and S.-H. Gary Chan, *Crowd counting on images with scale variation and isolated clusters*, (Proc. of the IEEE/CVF International Conference on Computer Vision Workshops, Seoul, Republic of Korea), 2019, DOI 10.1109/ICCVW.2019.00009.
20. Z.-Q. Cheng, J.-X. Li, Q. Dai, X. Wu, and A. G. Hauptmann, *Learning spatial awareness to improve crowd counting*, (Proc. of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea), 2019, pp. 6152–6161.
21. K. Simonyan and A. Zisserman, *Very deep convolutional networks for large-scale image recognition*, arXiv preprint, 2014, DOI 10.48550/arXiv.1409.1556.
22. K. He, X. Zhang, S. Ren, and J. Sun, *Deep residual learning for image recognition*, (Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA), 2016, pp. 770–778.
23. A. Howard, M. Sandler, G. Chu, L.-C. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan, and Q. V. Le, *Searching for MobileNetV3*, (Proc. of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea), 2019, pp. 1314–1324.
24. P. Ramachandran, B. Zoph, and Q. V. Le, *Searching for activation functions*, arXiv preprint, 2017, DOI 10.48550/arXiv.1710.05941.
25. Y. Li, Y. Chen, N. Wang, and Z. Zhang, *Scale-aware trident networks for object detection*, (Proc. of the IEEE/CVF International Conference on Computer Vision, Seouul, Republic of Korea), 2019, pp. 6054–6063.
26. X. Sun, Y. Jiang, Y. Ji, W. Fu, S. Yan, Q. Chen, B. Yu, and X. Gan, *Distance measurement system based on binocular stereo vision*, IOP Conf. Ser.: Earth Environm. Sci. **252** (2019), no. 5, DOI 10.1088/1755-1315/252/5/052051.
27. J.-R. Chang, P.-C. Chang, and Y.-S. Chen, *Attention-aware feature aggregation for real-time stereo matching on edge devices*, (Proc. of the Asian Conference on Computer Vision, Kyoto, Japan), 2020, pp. 365–380.
28. P. Wang, P. Chen, Y. Yuan, D. Liu, Z. Huang, X. Hou, and G. Cottrell, *Understanding convolution for semantic segmentation*, (IEEE Winter Conference on Applications of Computer Vision (WACV), Lake Tahoe, NV, USA), 2018, pp. 1451–1460.
29. Y. Li, X. Zhang, and D. Chen, *CSRNet: dilated convolutional neural networks for understanding the highly congested scenes*, (Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake Sity, UT, USA), 2018, pp. 1091–1100.
30. M.-R. Hsieh, Y.-L. Lin, and W. H. Hsu, *Drone-based object counting by spatially regularized regional proposal network*, (Proc. of the IEEE International Conference on Computer Vision (ICCV)), 2017, pp. 4145–4153.
31. G. Gao, Q. Liu, Z. Hu, L. Li, Q. Wen, and Y. Wang, *PSGCNet: a pyramidal scale and global context guided network for dense*

object counting in remote-sensing images, IEEE Trans. Geosci. Remote Sens. **60** (2022), 1–12.
32. E. Kilic and S. Ozturk, *An accurate car counting in aerial images based on convolutional neural networks*, J. Ambient Intell. Humanized Comput. **14** (2021), 1–1268.
33. Y. Zhang, D. Zhou, S. Chen, S. Gao, and Y. Ma, *Single-image crowd counting via multi-column convolutional neural network*, (Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA), 2016, pp. 589–597.
34. J. Yi, Z. Shen, F. Chen, Y. Zhao, S. Xiao, and W. Zhou, *A lightweight multiscale feature fusion network for remote sensing object counting*, IEEE Trans. Geosci. Remote Sens. **61** (2023), 1–13.
35. J. Zhang, J.-J. Qiao, X. Wu, and W. Li, *Vehicle counting network with attention-based mask refinement and spatial-awareness block loss*, (Proc. of the 29th ACM International Conference on Multimedia), 2021, pp. 2889–2898.
36. P. Wang, C. Gao, Y. Wang, H. Li, and Y. Gao, *Mobilecount: an efficient encoder-decoder framework for real-time crowd counting*, Neurocomputing **407** (2020), 292–299.

AUTHOR BIOGRAPHIES

Yu Liu received his BS degree in Computer Science from Nanjing Tech University, Nanjing, China, where he is pursuing his MS degree since September 2021. His research interests include dynamic neural networks and multimodal learning for detecting aerial image objects.

Hang Shen received his PhD degree (with honors) in Computer Science from the Nanjing University of Science and Technology. He worked as a full-time postdoctoral fellow with the Broadband Communications Research Lab, ECE Department, University of Waterloo, Waterloo, ON, Canada, from 2018 to 2019. He is an associate professor with the Department of Computer Science and Technology, Nanjing Tech University, Nanjing, China. His research interests include machine learning for multimedia systems, drone communication networks, cybersecurity, and blockchain. He has published research papers in prestigious international journals and conferences, including IEEE TRANSACTIONS ON MOBILE COMPUTING, IEEE TRANSACTIONS ON BROADCASTING, IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY, JOURNAL OF SYSTEMS ARCHITECTURE, and IEEE ICC. He serves as an associate editor for the *Journal of Information Processing Systems* and IEEE ACCESS and as an academic editor for *Mathematical Problems in Engineering*. He was a Guest Editor for *Peer-to-Peer Networking and Applications* and a TPC member of the 2021 Annual International Conference on Privacy, Security, and Trust. He is a Senior Member of CCF and IEEE, and an executive committee member of the ACM Nanjing Chapter.

Tianjing Wang received her BSc degree in Mathematics from Nanjing Normal University in 2000, MSc degree in Mathematics from Nanjing University in 2002, and PhD degree in Signal and Information Systems from the Nanjing University of Posts and Telecommunications (NUPT) in 2009. From 2011 to 2013, she worked as a full-time postdoctoral fellow with the School of Electronic Science and Engineering, NUPT. She was a visiting scholar with the ECE Department, State University of New York, Stony Book, NY, USA, from 2013 to 2014. She is an associate professor with the Department of Communication Engineering at Nanjing Tech University. Her research interests include deep learning in multimedia systems and connected autonomous vehicles.

Guangwei Bai received his BEng and MEng degrees in Computer Engineering from Xi'an Jiaotong University, Xi'an, China, in 1983 and 1986, respectively, and PhD degree in Computer Science from the University of Hamburg, Hamburg, Germany, in 1999. From 1999 to 2001, he worked as a research scientist at the German National Research Centre for Information Technology. In 2001, he joined the University of Calgary, Calgary, Canada, as a research associate. Since 2005, he has been working at the Nanjing Tech University, Nanjing, China, as a professor in Computer Science. From October to December 2010, he was a visiting professor with the ECE Department, University of Waterloo, Waterloo, ON, Canada. His research interests include architecture and protocol design for future networks, wireless multimedia, quality-of-service provisioning, blockchain, and cybersecurity. He is a member of ACM and a distinguished member of CCF.

How to cite this article: Y. Liu, H. Shen, T. Wang, and G. Bai, *Vehicle counting in drone images: An adaptive method with spatial attention and multiscale receptive fields*, ETRI Journal **47** (2025), 7–19, DOI 10.4218/etrij.2023-0426.