Scale Region Recognition Network for Object Counting in Intelligent Transportation System

Xiangyu Guo, Mingliang Gao, Wenzhe Zhai, Qilei Li, *Student Member, IEEE*, and Gwanggil Jeon, *Senior Member, IEEE*

***Abstract*— Self-driving technology and safety monitoring devices in intelligent transportation systems require superb capacity for context awareness. Accurately inferring the counts of crowds and vehicles are the two practical and fundamental tasks in the transportation system. However, the scale variation and background interference in the traffic image hinder the counting performance. To solve the aforementioned problems, a scale region recognition network (SRRNet) is proposed in this paper. It has two key components, termed scale level awareness (SLA) module and object region recognition (ORR) module. The SLA module aims to encode the representations at multiple scales, which are beneficial to address the scale variation. The ORR module is designed to suppress background interference through the visual attention mechanism. Extensive experimental results on four crowd counting datasets and five vehicle counting datasets have demonstrated the superiority of the proposed SRRNet in both counting accuracy and robustness compared with the mainstream competitors. Meanwhile, substantial ablation studies have proved the effectiveness of the proposed SLA and ORS modules.**

***Index Terms*— Intelligent transportation systems, crowd counting, vehicle counting, deep learning.**

I. Introduction

INTELLIGENT transportation system is a substantial component of the smart city, and it has been treasured by academia and industry [1]. Self-driving technology and intelligent monitoring devices, which require extremely high safety and reliability, are the focus of research [2]. Therefore, before intelligent transportation technologies are proliferating in society, they need the superior capacity to perceive surroundings, such as the number, location, and flow of objects. Pedestrians and vehicles are major subjects of interest in the traffic system, and it is a profoundly thought-provoking task to accurately determine the counts of people and vehicles.

Fig. 1. Challenges of scale variation and background interference in crowd and vehicle scenarios.

The solutions to the object counting fall into three categories, *i.e.,* detection-based method [3], regression-based method [4] and density estimation-based method [5]. The detection-based methods intend to deploy a tailored detector to discover an object by marking it with a bounding box. These methods are suited for a relatively sparse scenario. Once the scenario becomes congested, the counting results will get worse dramatically. Therefore, regression-based methods are brought to strengthen the counting performance in dense regions. They are intended to develop a counting model, which is capable to map the representations to the object counts. However, they always ignore spatial information, which is valuable for high-level visual tasks [6], such as crowd poses estimation, crowd anomaly detection and image segmentation. To this end, the density estimation-based method is present, which regresses a high-quality density map and then sums the pixels on the map as the count value. Benefitting from the strong feature extraction capability of convolution neural network (CNN), it has made adequate advances in crowd and vehicle counting tasks [7], [8]. Nevertheless, scale variation and background interference are two major challenges that hinder the further improvement of counting performances. The two challenges in crowd and vehicle counting are depicted in Fig. 1.

Manuscript received 17 December 2022; revised 30 March 2023 and 25 May 2023; accepted 12 July 2023. Date of publication 25 July 2023; date of current version 29 November 2023. This work was supported in part by the National Natural Science Foundation of Shandong Province under Grant ZR2021QD0410. The Associate Editor for this article was Z. Lv. *(Corresponding authors: Mingliang Gao; Gwanggil Jeon.)*

Xiangyu Guo, Mingliang Gao, and Wenzhe Zhai are with the School of Electrical and Electronic Engineering, Shandong University of Technology, Zibo 255000, China (e-mail: xiangyvguo@163.com; mlgao@sdut.edu.cn; wenzhezhai@163.com).

Qilei Li is with the School of Electronic Engineering and Computer Science, Queen Mary University of London, E1 4NS London, U.K. (e-mail: q.li@qmul.ac.uk).

Gwanggil Jeon is with the School of Electrical and Electronic Engineering, Shandong University of Technology, Zibo 255000, China, and also with the Department of Embedded Systems Engineering, Incheon National University, Incheon 22012, South Korea (e-mail: ggjeon@gmail.com).

Digital Object Identifier 10.1109/TITS.2023.3296571

The problem of scale variation is mainly caused by the camera perspective distortion. The left column of Fig. 1 shows the scale variation of head and vehicle in real scenarios (marked in red boxes). Obviously, the object size closer to the camera is larger. To overcome this issue, extracting sufficient multiscale representations is the fundamental and effective solution. Numerous efforts [5], [9], [10], [11], [12] are devoted to this aim. Zhang et al. [9] built a simple but instructive network named MCNN, in which three parallel branches were built to capture features with different scales. Although the counting accuracy is not ideal in the current view, it provides development directions for the following works. Similar to MCNN, Chen et al. [13] proposed a scale pyramid module (SPM) to encode multiscale information. The SPM utilizes four parallel dilated convolution layers to extract the features and then fuse the features by a concatenation. Although the multi-column architecture can extract multi-scale features, it introduces some defects such as a bloated architecture structure, increasing computation, and redundant information [6]. To this end, the single-column structure [8], [10] has drawn more and more attention because it can compress the width of the network while ensuring the counting performance. Li et al. [10] discarded the pooling layers and fully connected layer of VGG-16 and incorporated six dilated convolution layers to enlarge the receptive fields. Cao et al. [14] employed Inception structure as the encoder to extract multiple hierarchy features and leveraged several transposed convolution layers as the decoder to output the density map. Specifically, the number of convolution filters is small (16, 32 and 64) compared with the deep network.

The problem of background interference is another inevitable issue in the domain of object counting in real scenarios. It could distract the network to identify target information, which leads to overestimation or underestimation. The right column of Fig. 1 illustrates the background interference in the crowds and parking lot. It depicts that the heads and vehicles are sheltered by the flags and plants (marked in the blue boxes). Visual attention is an effective way to suppress background disturbance, and it has achieved successful results in counting tasks [7], [15], [16], [17]. Hu et al. [18] proposed the channel attention module named squeeze and excitation (SE) module, which recalibrates the channel weights to highlight the foreground. Liang et al. [16] further analyzed the SE module and improved it from the perspective of denoising. Specifically, it introduces a pair of parameters to linearly transform the feature in terms of scaling and shift. Gao et al. [7] introduced a dual attention module for rich semantic information extraction. Concretely, it consists of a channel attention unit for semantic segmentation and a spatial attention unit for encoding long-range dependencies.

In this paper, we propose a **S**cale **R**egion **R**ecognition **Net**work (SRRNet) to cope with the problems of scale variation and background interference simultaneously. It is composed of four components. First, a backbone is used to extract the low-level features. Subsequently, a scale level awareness (SLA) module is proposed to deal with the scale variation. The SLA module adopts two pre-activate convolution units to avoid invalid features, and an hourglass block to capture multiscale information. Then, an ORR module is designed to suppress the background interference. It recalibrates the channel weights adaptively by introducing additional parameters, which is helpful to discriminate between foreground and background. At last, several transposed convolution filters are leveraged for density map prediction. In a nutshell, the contributions are three-fold:

1) An SRRNet is built in a divide-and-conquer manner to address the scale variation and background interference in inferring the counts of crowds and vehicles in the intelligent transportation system.
2) An SLA module is designed to capture multiscale representations so as to alleviate the scale variation. Meanwhile, an ORR module is proposed to suppress the background interference by introducing the visual attention mechanism.
3) Extensive experiments on four crowd counting datasets and five vehicle counting datasets are conducted to demonstrate the superiority of the SRRNet. Furthermore, sufficient ablation studies are carried out to evaluate the effectiveness of the proposed modules.

The technical roadmap of this paper is organized as follows. Section I introduces the motivation and contribution. Section II reviews the works related to object counting. Section III analyzes the proposed SRRNet in detail. In Section IV, experimental results on object counting datasets are discussed. The conclusion of the paper is provided in Section V.

II. Related Work

In this section, the methods related to the work are revisited in a problem-oriented schema. It revolves around two issues to be solved, *i.e.*, scale variation and background interference.

A. Solutions to Scale Variation

Scale variation is an inherent and thorny challenge in both crowd and vehicle counting. Fundamentally speaking, the solution to the problem is to extract multiscale features from images effectively. Many sophisticated modules or scale fusion mechanisms are built to meet the requirements [7], [9], [10].

Zhang et al. [9] built a three-column network, each column utilizing convolution filters with diverse kernel sizes to capture multiscale features. To boost the counting accuracy of congested crowd scenarios, Li et al. [10] deployed six dilated convolution layers as the backend and proposed the congested scene recognition network (CSRNet). The dilated convolution layers could enlarge the receptive fields to encode more scale information. Cao et al. [14] designed a scale aggregation module to increase the scale diversity of the features. It is composed of four branches, each of which is responsible to acquire features with different scales. The proposed module can enhance the information interaction between columns. Olmschenk [19] designed some dense blocks and map modules to build a multiscale upsampling denseblocks network (MUD-Net). Different from the traditional dense blocks, the proposed denseBlock is equipped with a transposed convolution layer, which can guarantee the input and output have the same resolution. Liu et al. [11] proposed an estimation network to

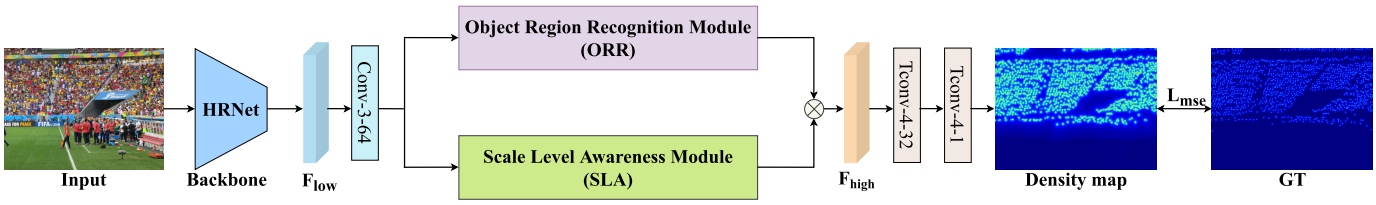

Fig. 2. Architecture of the SRRNet for object counting. The convolution layer is represented as 'Conv-(kernel size)-(output channels)'. 'Tconv' denotes the transposed convolution. '⊗' represents the element-wise multiplication.

improve the generality of different scales. It employs a tailored Xception as an encoder to extract the basic features. Then, several dilated convolutions and transposed convolutions are adopted as an encoder to output the density map. Gao et al. [7] introduced a scale pyramid module (SPM) to address the scale variation and small object feature extraction. The SPM employs four parallel dilated convolution layers with different dilated rates of 2, 4, 6, and 8, respectively. Dai et al. [20] devised a dense dilated convolution block (DDCB) to capture a wide range of scale features. Specifically, the DDCB consists of three dilated convolution layers with the number of 64 convolutions to capture multiscale features, and the dilated rates are set to 1, 2 and 3, respectively. Sam et al. [12] proposed a top-down feature modulator (TFM) to encode multiscale representations. The TFM perceives global context from a large scale and jointly multiple hierarchies feature to identify people. Inspired by the above multiscale feature extraction mechanisms, we put forward an SLA module to capture features at diverse scales.

B. Solutions to Background Interference

The visual attention mechanism aims to highlight the target features, which is helpful to mitigate background interference and focus on the object regions. Many attention mechanisms are proposed and have been widely used in counting tasks [5], [15], [21]. Gao et al. [15] designed a channel attention module (CAM) to boost the class-specific response. It is effective to discriminate the crowd region from the background, which can alleviate the error estimation by misidentifying objects. Zhu et al. [21] set an attention map path to generate a probability map, which could represent head probability at each pixel. Meanwhile, a tailored attention loss was introduced to supervise the attention map generation. Gao et al. [7] proposed an attention module to capture context information. Specifically, a channel attention unit is employed to emphasize the foreground contexts in dense regions and a spatial attention unit is introduced to encode the wide range of dependencies. Liang et al. [16] introduced an attention-based BN to recalibrate the channel weight through a linear transformation.

Apart from the attention mechanism, many semantic segmentation methods are proposed to alleviate background interference. For example, Khan et al. [22] built an area segmentation framework for crowd counting. It consists of a classification module, a semantic scene segmentation (SSS) module, and a density estimation module. Specifically, the SSS module generates a segmentation map to emphasize the foreground regions. Meng et al. [23] introduced a regularized proxy task based on the binary segmentation map. The proposed proxy task can guide the network to generate hard and soft uncertainty maps to suppress background interference. Gao et al. [24] proposed a foreground and background segmentation (FBS) module to discriminate the object region and background. It is built with several stacked convolution blocks, which can produce a segmentation map to highlight the foreground. Liu et al. [25] devised a surrogate task to generate interrelated segmentation maps to estimate the final density map. For each segmentation map, the pixel values higher than the predefined threshold are set to one and otherwise set to zero.

Unlike the aforementioned methods that provide precious experience to suppress background interference, we present an ORR module to address the issue from the perspective of denoising, which is distinct from most methods in this paper.

III. Methodology

A. Architecture Overview

The architecture of the proposed SRRNet is depicted in Fig. 2. First, it adopts a backbone to extract the low-level features F_{low}. Then, an ORR module is designed to highlight the foreground, and an SLA module is proposed to capture the multiscale representations. Finally, two transposed convolution layers are utilized as a decoder to predict the density map.

Specifically, the HRNet [26] is employed as the backbone to generate a feature map 1/4 the size of the input image. Then, a 3×3 convolution layer is leveraged to compress the channels to 64, which can relieve the computation burden. The ORR and SLA modules boost the target features independently in channel and spatial dimensions. The optimized features through the modules are multiplied to produce a high-quality feature map F_{high}. Because the size of F_{high} is only a quarter of the input image, two transposed convolution layers are deployed on the backend to upsample the feature map twice to generate the final prediction.

B. Scale Level Awareness Module

Scale variation is an inescapable problem in object counting. Extracting sufficient multiscale information extraction is a fundamental way to address the problem. To this end, an SLA module is introduced to capture features at multiple scales. The architecture and the detailed configuration of the SLA module are shown in Fig. 3 and Table I, respectively.

Given an input $I \in \mathbb{R}^{C\times H\times W}$, two convolution units are first employed to generate an enhanced feature map $M_f \in \mathbb{R}^{C\times H\times W}$. It is helpful to avoid the invalid feature caused by the zero gradients, so as to boost the stable ability of

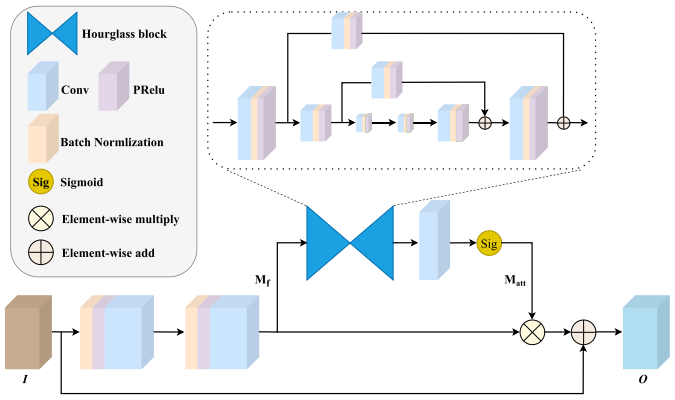

Fig. 3. Architecture of the SLA module.

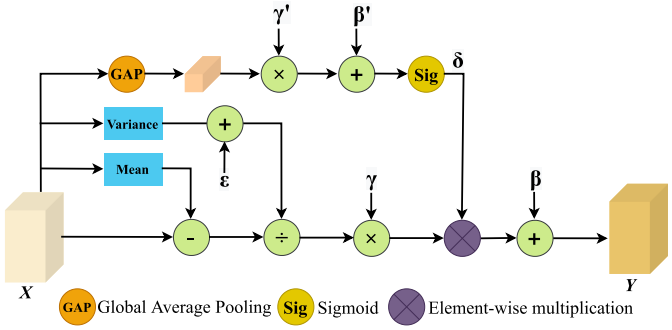

Fig. 4. Architecture of the ORR module.

the network [27]. Specifically, the convolution unit consists of a batch normalization, a PReLU activation function, and a convolution layer. Subsequently, a scale awareness (SA) unit is deployed to generate a spatial attention map $M_f \in \mathbb{R}^{C\times H\times W}$, which contains rich scale information. The SA unit is composed of an hourglass block for extracting multiscale features, a convolution layer for channel reduction, and a Sigmoid function for weight generation. d is the depth of the hourglass block, and it controls the number of different scale levels.

C. Object Region Recognition Module

The attention mechanism has been proven as a viable solution to background interference, which aims to realign the weights of features to better distinguish the foreground and background. According to the work [16], a lot of noise is produced in the process of batch normalization, which is harmful to the network performance. To alleviate the issue, we propose an ORR module to distinguish the object region from the complex background from the perspective of denoising. The architecture of the ORR module is depicted in Fig. 4.

Supposing the feature map be $X \in \mathbb{R}^{B\times C\times H\times W}$, where B, C, H and W denote the batch size, channel, height, and width of the feature, respectively. On the top branch, a global average pooling (GAP) operation is first executed to shrink the input X. The process is formulated as,

$$\mathrm{GAP}(X) = \frac{1}{H\cdot W}\sum_{h=1}^{H}\sum_{w=1}^{W} X. \tag{1}$$

Then, a pair of parameters γ' and β' is introduced to transform linearly from the perspective of scale and shift. They are initially set to 0 and -1, respectively. At last, the Sigmoid function is adopted to generate a set of weight coefficients δ, which are served for adjusting the scaling of each channel. It can be defined as,

$$\delta = \mathrm{Sig}(\gamma' \times \mathrm{GAP}(X) + \beta') \tag{2}$$

To better compensate for the representation, the mean and variance of each channel are calculated. In particular, a constant $\varepsilon = 1e-5$ is added to avoid the variance being zero. Then a couple of learnable parameters γ and β are introduced to restore the representation capacity. Finally, the output Y is generated by:

$$Y = (\frac{X-\mu}{\sigma})\cdot\gamma\cdot\delta+\beta, \tag{3}$$

where μ and σ denote the mean and variance of each channel.

D. Ground Truth Generation

Following the work [28], the focal inverse distance transform (FIDT) map is adopted to generate the ground truth. It can be decoupled into two stages, *i.e.*, distance transform map generation and focal function addition. Assuming that there is annotation at the pixel point (x', y'), the distance transform map is generated by:

$$M_{dt} = \min_{(x',y')\in H}\sqrt{(x-x')^2+(y-y')^2}, \tag{4}$$

where H denotes a set of head labels. Then, the inverse function and focal function are adopted to restrict the distance variations. In a nutshell, the FIDT map is generated by,

$$M_f = \sum_{i=1}^{N}\frac{1}{P_i(x,y)^{\alpha\times P_i(x,y)+\beta}+C}, \tag{5}$$

where α and β are set to 0.02 and 0.75, respectively. N represents the number of head annotations. C is set to 1 to avoid a denominator of zero.

E. Loss Function

The MSE loss is leveraged to train the proposed SRRNet, which optimizes the model by minimizing the Euclidean distance between the predicted and ground truth density maps. It is formulated as follows,

$$l(\theta) = \frac{1}{N}\sum_{i=1}^{N}\left\| X_i^{Est}(\theta) - X_i^{GT}\right\|_2^2, \tag{6}$$

where N represents the batch size, and θ denotes the parameters to be trained. X_i^{GT} and X_i^{Est} denote the ground truth and estimated density map, respectively.

TABLE I
CONFIGURATION OF THE SLA MODULE

Layers	Kernel Size	Input Channels	Output Channels	Activation Function
Convolution unit 1	3×3	64	64	PReLU
Convolution unit 2	3×3	64	64	PReLU
Hourglass block	3×3	64	64	PReLU
Channel reduction layer	1×1	64	1	PReLU

IV. EXPERIMENTAL RESULTS AND ANALYSIS

A. Datasets

Five crowd counting (ShanghaiTech [9], UCF_CC_50 [29], UCF-QNRF [30], JHU-Crowd++ [31] and NWPU-Crowd [32]) and five vehicle counting datasets (CARPK [33], PUCPR+ [33], Large-vehicle [7], Small-vehicle [7] and TRANCOS [34]) are used to assess the performance of the SRRNet.

ShanghaiTech [9] contains 1,198 images, which are divided into two parts, *i.e.*, Part A and Part B based on the density distributions. The images in Part A are randomly taken from the Internet, in which the crowd density is relatively high, while the images in Part B are collected from the street in Shanghai, exhibiting a low crowd density.

UCF_CC_50 [29] consists of 50 high-resolution gray images captured in diverse densities and scenarios.

UCF-QNRF [30] is characterized by a broad variety of backgrounds, diversity of perspectives, and lighting variations. It comprises 1,535 high-quality images, some of which have very high resolutions. Therefore, it is necessary to resize the images during training to avoid out-of-memory.

JHU-Crowd++ [31] is an unconstrained dataset, which is composed of 4,250 images. Specifically, it has three characteristics, namely sufficient training samples, different weather conditions, and a series of distractor images.

CARPK [33] is collected by drones from four diverse parking lots. It has 1,448 images with 89,777 labelled cars.

PUCPR+ [33] includes 125 images captured from the parking lot under different weather conditions, *e.g.*, rainy, cloudy and sunny.

Large-vehicle [7] comprises 172 remote sensing images with an average resolution of 1552×1573. The marked object is the large vehicle in the image.

Small-vehicle [7] is also a remote sensing counting dataset. It has 280 high-resolution images with 148,838 small vehicles. Compared with the Large-vehicle dataset, it has a larger scale variation.

TRANCOS [34] consists of 1,244 images taken from congested traffic environments. Each image is provided with a mask.

B. Implementation Details

All experiments are implemented under the Pytorch toolbox. For the datasets (CARPK, PUCPR+, Large-vehicle and Small-vehicle) annotated by the boundary boxes, we take the center of gravity of them as the central location to generate the ground truth. To augment the training data, the images are arbitrarily cropped and mirror-flipped. Specifically, for the small resolution datasets (ShanghaiTech and TRANCOS), the crop size is set to 256 × 256, and 512 × 512 for other datasets. The training batch size is set to 16. Adam optimizer is utilized to optimize the proposed model. The learning rate is set to 1e-4, and the weight decay is set to 5e-4. To avoid out-of-memory, we resize the images to make sure the longer side is less than 2048.

C. Evaluation Protocols

The Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) are widely leveraged to assess the counting accuracy and robustness [5], [9], [35]. They are defined as,

$$MAE = \frac{1}{T}\sum_{i=1}^{T}|E_i - G_i|, \tag{7}$$

$$RMSE = \sqrt{\frac{1}{T}\sum_{i=1}^{T}|E_i - G_i|^2}, \tag{8}$$

where T denotes the number of test images, E_i and G_i represent the estimated value and ground truth of the i-th image, respectively.

In particular, for the TRANCOS dataset, we adopt the specific Grid Average Mean Error (GAME) [34], which can reflect the counting accuracy both in global and local regions. It is formulated as,

$$GAME(L) = \frac{1}{N}\sum_{n=1}^{N}\left(\sum_{l=1}^{4^L}|E_i - G_i|\right), \tag{9}$$

where 4^L represents the non-overlapping subregions of the image. In addition, the $GAME$ is equal to MAE when L is set as 0.

D. Comparison on Crowd Counting

The comparison results of the proposed SRRNet and other crowd counting models are reported in Table II. On the Part A dataset, the SRRNet scores 60.8 and 103.0 in MAE and RMSE, both outperforming all the competitors. Compared with the second-best TEDNet [36], it reduces the MAE and RMSE by 5.3% and 5.6%, respectively. Furthermore, compared with a multimodel fusion method LSC-CNN [12], the SRRNet improves the MAE and RMSE by 8.4% and 12.0%, respectively. On the relatively sparse Part B dataset, the SRRNet achieves the best MAE of 7.4. Meanwhile, the RMSE is 13.6 which is still competitive. The objective results

TABLE II
OBJECTIVE COMPARISON RESULTS ON CROWD COUNTING. THE BEST RESULTS ARE HIGHLIGHTED IN **BOLD**

Method	Part A		Part B		UCF_CC_50		UCF-QNRF		JHU++	
	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
MCNN [9]	110.2	173.2	26.4	41.3	377.6	509.1	277.0	426.0	188.9	483.4
Switch-CNN [38]	90.4	135.0	21.6	33.4	318.1	439.2	-	-	-	-
A-CCNN [39]	85.4	124.6	19.2	31.5	-	-	367.3	-	171.2	453.1
SANet [14]	67.0	104.5	8.4	13.6	258.4	334.9	-	-	91.1	320.4
CSRNet [10]	68.2	115.0	10.6	16.0	266.1	397.5	-	-	85.9	309.2
SFCN [37]	64.8	107.5	7.6	13.0	214.2	318.2	102.0	171.4	77.5	297.6
RAZ [40]	65.1	106.7	8.4	14.1	-	-	116.0	195.0	-	-
TEDNet [36]	64.2	109.1	8.2	12.8	249.4	354.5	113.0	118.0	75.0	299.9
LSC-CNN [12]	66.4	117.0	8.1	12.7	-	-	120.5	218.2	112.7	454.4
MUD-iKNN [19]	68.0	117.7	13.4	21.4	237.7	305.7	104.0	172.0	-	-
DUBNet [41]	64.6	106.8	7.7	**12.5**	243.8	329.3	105.6	180.5	133.5	416.5
SUA-Fully [23]	66.9	125.6	12.3	17.9	-	-	119.2	213.3	-	-
PCCNet [24]	73.5	124.0	11.0	19.0	240.0	315.5	148.7	247.3	-	-
CG-DRCN [31]	64.0	**98.4**	8.5	14.4	-	-	112.2	176.3	71.0	278.6
SRRNet (Ours)	**60.8**	103.0	**7.4**	13.6	**172.9**	**256.3**	**89.5**	**162.9**	**62.4**	**254.6**

on Part A and B are depicted in Fig. 5. It proves that SRRNet can adapt to congested and sparse crowd scenarios.

On the extremely dense UCF_CC_50 dataset, the proposed SRRNet scores 172.9 and 256.3 in MAE and RMSE, which perform best among the competitors. Compared with the second-best SFCN [37], the SRRNet significantly improves by 19.3% and 19.5% in MAE and RMSE. The results indicate that the proposed SRRNet is capable of counting crowds in high-density scenarios.

On the UCF-QNRF dataset, which is characterized by scale variation, the proposed SRRNet also exhibits the best results compared with other models. Compared with MUD-ikNN [19] which is specific to address the scale variation, the proposed SRRNet achieves an improvement of 13.9% and 16.2% in MAE and RMSE, respectively.It proves that the SRRNet is beneficial to address the scale variation.

On the JHU++ dataset which has various weather scenarios, the SRRNet scores 62.4 and 254.6 in MAE and RMSE which are the best results. Compared with CG-DRCN [31], the SRRNet exhibits a 5.6% and 8.6% reduction in MAE and RMSE, respectively. Notably, CG-DRCN is also built to remedy background interference. The results verify that the SRRNet is helpful to suppress the side effect of background noise. The subjective results on the datasets are shown in Fig. 5. It can be observed that the estimated density maps and counting numbers are very close to the ground truth.

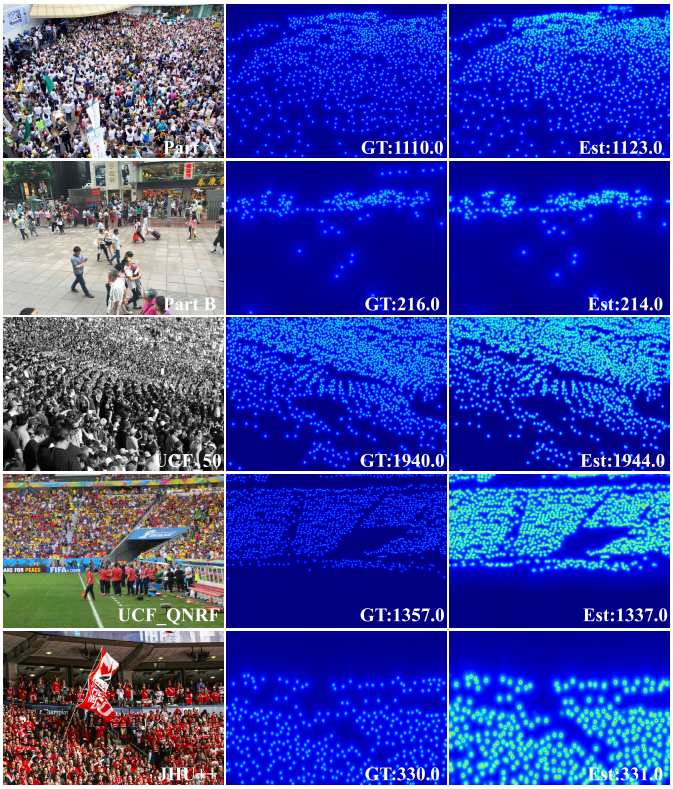

Fig. 5. Subjective results on crowd counting datasets. 'GT' and 'Est' indicate the ground truth and estimated value.

E. Comparison on Vehicle Counting

The comparison results of the vehicle counting methods are reported in Table III, IV and V.

As shown in Table III, the proposed SRRNet outperforms the mainstream methods on the CARPK and PUCPR+ datasets. Among them, the first six rows are detection-based methods, and they are inferior to the proposed SRRNet. The One-Look Regression [42] is a regression-based method, which is also unsatisfactory. Compared with the other density estimation-based methods, *i.e.*, MCNN [9], SFANet [10] and BL [8], the SRRNet exhibits the best performance in counting accuracy and robust. Specifically, compared with the second-best method BL [8] on the CARPK dataset, the SRRNet improves the MAE and RMSR by 11.3% and 3.5%, respectively. Furthermore, on the PUCPR+ dataset, the SRR-

TABLE III

OBJECTIVE EXPERIMENTAL RESULTS ON THE CARPK AND PUCPR+ DATASETS. THE BEST RESULTS ARE HIGHLIGHTED IN **BOLD**

Methods	CARPK		PUCPR+	
	MAE	RMSE	MAE	RMSE
YOLO [43]	102.89	110.02	156.72	200.54
FRCN [44]	103.48	110.64	156.76	200.59
LEP [3]	51.83	-	15.17	-
LPN [33]	23.80	36.79	22.76	34.46
SSD [45]	37.33	42.32	119.24	132.22
RetinaNet [46]	16.62	22.30	24.58	33.12
One-Look Regression [42]	59.46	66.84	21.88	36.73
MCNN [9]	39.10	43.30	21.86	29.53
CSRNet [10]	11.48	13.32	8.65	10.24
BL [8]	9.58	11.38	6.54	8.13
SRRNet (Ours)	**8.50**	**10.98**	**2.04**	**2.79**

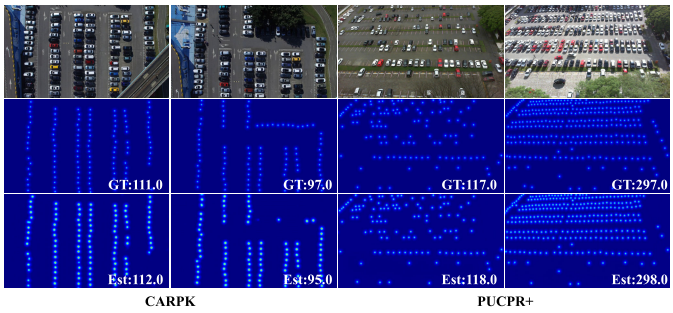

Fig. 6. Subjective results on CARPK and PUCPR+ datasets.

Net has a 68.8% and 65.7% significant margin with BL [8] in MAE and RMSE. It is worth noting that the PUCPR+ dataset has three different weather conditions (sunny, rainy and cloudy), the results prove that the SRRNet can deal with complex scenes with various challenging weather conditions. Some subjective results on CARPK and PUCPR+ datasets are shown in Fig. 6. It can be observed that the estimated value is very close to the ground truth.

As shown in Table IV, on the Large-vehicle and Small-vehicle datasets, the SRRNet still performs best in MAE. The ASPDN [7] which is also specific to solving the problems of scale variation and background disturbances performs best in RMSE in the Large-vehicle dataset. Compared with ASPDN, the SRRNet achieves the second-best results, but the difference is only 0.6%. On the Small-vehicle dataset, it can be seen that the SRRNet improves significantly in MAE and RMSE. Specifically, compared with the SPN [47], the SRRNet improves the MAE and RMSE by 73.0% and 66.5%, respectively. The results also demonstrate that the SRRNet is suitable for small object counting, since the vehicles in the remote sensing images present tiny sizes. The subjective results illustrated in Fig. 7 further prove the effectiveness of the proposed SRRNet in density estimation and counting.

The objective comparison results on the TRANCOS datasets are shown in Table V. One can see that the SRRNet performs best across all the other methods in GAME(0), GAME(1), GAME(2) and GAME(3). It indicates that the SRRNet can not only have a satisfactory global counting performance,

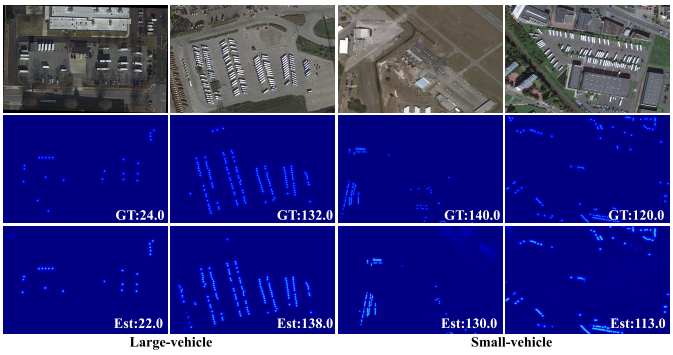

Fig. 7. Subjective results on Large-vehicle and Small-vehicle datasets.

TABLE IV

OBJECTIVE EXPERIMENTAL RESULTS ON THE LARGE-VEHICLE AND SMALL-VEHICLE DATASETS. THE BEST RESULTS ARE HIGHLIGHTED IN **BOLD**

Methods	Large-vehicle		Small-vehicle	
	MAE	RMSE	MAE	RMSE
MCNN [9]	36.56	55.55	488.65	1317.44
CMTL [48]	61.02	78.25	490.53	1321.11
SANet [14]	62.78	79.65	497.22	1276.66
CSRNet [10]	34.10	46.42	443.72	1252.22
SCAR [15]	62.78	79.64	497.22	1276.65
SPN [13]	36.21	50.65	455.16	1252.92
CAN [49]	34.56	49.63	457.36	1260.39
SFCN [37]	33.93	49.74	440.70	1248.27
SFANet [21]	29.04	47.01	435.29	1284.15
ASPDN [7]	18.76	**31.06**	433.23	1238.61
SRRNet (Ours)	**18.25**	31.24	**122.79**	**419.65**

TABLE V

OBJECTIVE EXPERIMENTAL RESULTS ON THE TRANCOS DATASETS. THE BEST RESULTS ARE HIGHLIGHTED IN **BOLD**

Methods	GAME(0)	GAME(1)	GAME(2)	GAME(3)
Method in [50]	17.77	20.14	23.65	25.99
Method in [4]	13.76	16.72	20.72	24.36
Method in [51]	10.98	13.75	16.69	19.32
PSDDN [52]	4.79	5.43	6.68	8.40
LSC-CNN [12]	4.60	5.40	6.90	8.30
SRRNet (Ours)	**3.89**	**4.83**	**5.58**	**7.95**

but also an ideal local counting performance. To be specific, the first three rows are the detection-based methods and the performance is generally unsatisfactory. Compared with the LSC-CNN [12], the SRRNet improves by 15.4% in GAME(0). Some subjective results on TRANCOS dataset are shown in Fig. 8.

F. Ablation Studies

1) Ablation Studies on the Proposed Modules: To assess the effectiveness of the proposed SLA and ORR modules, a set of ablation studies is carried out. The experimental results are listed in Table VI.

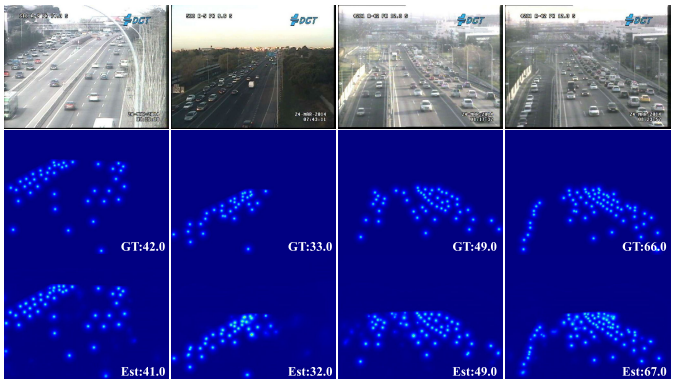

Fig. 8. Subjective results on TRANCOS datasets.

TABLE VI

OBJECTIVE ABLATION STUDIES ON THE PROPOSED SLA AND ORR MODULES. THE BEST RESULTS ARE HIGHLIGHTED IN **BOLD**

Baseline	SLA	ORR	MAE	RMSE
✓			64.5	119.0
✓	✓		62.1	111.5
✓		✓	63.7	114.5
✓	✓	✓	**60.8**	**103.0**

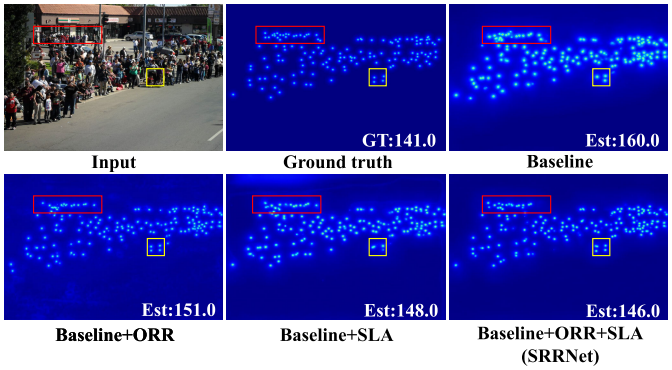

Fig. 9. Subjective results of different models on a representative sample from ShanghaiA datasets.

The 'Baseline' denotes the model with the basic backbone and backend. It obtains the worst MAE and RMSE with 64.5 and 119.0, respectively. When the SLA module is added to the baseline, the MAE and RMSE improve by 3.7% and 6.3%, respectively. Meanwhile, when the ORR module is incorporated into the baseline, it achieves an improvement by 1.2% and 3.9% reduction in MAE and RMSE, respectively. Finally, when the SLA and ORR modules are introduced to the baseline simultaneously, it achieves the best MAE and RMSE with 60.8 and 103, respectively. Fig. 9 provides the subjective results of different configurations on a representative sample. The given image depicts a street scenario with scale variation. Comparative results prove that both the ORR module and SLA module are helpful to boost the counting performance, as shown in highlighted in the red and yellow boxes. The final SRRNet equipped with these two modules achieves the best performance.

TABLE VII

OBJECTIVE ABLATION STUDIES ON THE DIFFERENT DEPTHS. THE BEST RESULTS ARE HIGHLIGHTED IN **BOLD**

Baseline	MAE	RMSE
SRRNet (d=1)	63.3	112.9
SRRNet (d=2)	**60.8**	**103.0**
SRRNet (d=3)	62.4	105.7
SRRNet (d=4)	62.5	109.1

2) Ablation Studies on the Depth d: To further explore the effect of the depth d of the hourglass block in the SLA module, we execute an ablation experiment on Shanghai Part A dataset. The results are reported in Table VII.

It proves that all the hourglass blocks of different depths are useful to improve the counting performance. Specifically, the SRRNet with $d = 1$ obtains a relatively lower improvement compared with the other three models, because the hourglass block cannot extract multiscale information. Meanwhile, when $d = 3$ and $d = 4$, the counting performance achieves certain improvements. On the contrary, the SRRNet equipped with $d = 2$ achieves the best results with the best MAE and RMSE.

V. CONCLUSION

In this work, we propose an SRRNet to address the issues of scale variation and background interference for crowd and vehicle counting in the intelligent transportation system. The proposed SRRNet consists of a backbone for low-level feature extraction, an SLA module to capture the multiscale information, an ORR module to suppress the background interference, and a decoder for density map prediction. Experimental results on four public crowd counting datasets and five vehicle counting datasets have demonstrated the superiority of the SRRNet in terms of subjective and objective evaluation. Ablation studies are carried out to prove the effectiveness of the components.

DECLARATIONS

Conflict of interest The authors declare that they have no conflict of interest.

REFERENCES

[1] L. Claussmann, M. Revilloud, D. Gruyer, and S. Glaser, "A review of motion planning for highway autonomous driving," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 5, pp. 1826–1848, May 2020, doi: 10.1109/TITS.2019.2913998.

[2] J. Guo, U. Kurup, and M. Shah, "Is it safe to drive? An overview of factors, metrics, and datasets for driveability assessment in autonomous driving," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 8, pp. 3135–3151, Aug. 2020, doi: 10.1109/TITS.2019.2926042.

[3] T. Stahl, S. L. Pintea, and J. C. van Gemert, "Divide and count: Generic object counting by image divisions," *IEEE Trans. Image Process.*, vol. 28, no. 2, pp. 1035–1044, Feb. 2019.

[4] V. Lempitsky and A. Zisserman, "Learning to count objects in images," in *Proc. Annu. Conf. Neural Inf. Process. Syst. (NeurIPS)*, 2010, pp. 1324–1332.

[5] X. Guo, M. Gao, W. Zhai, J. Shang, and Q. Li, "Spatial-frequency attention network for crowd counting," *Big Data*, vol. 10, no. 5, pp. 453–465, Oct. 2022, doi: 10.1089/big.2022.0039.

[6] G. Gao, J. Gao, Q. Liu, Q. Wang, and Y. Wang, "CNN-based density estimation and crowd counting: A survey," 2020, *arXiv:2003.12783*.

[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.*, vol. 59, no. 5, pp. 3642–3655, May 2021, doi: 10.1109/TGRS.2020.3020555.

[8] Z. Ma, X. Wei, X. Hong, and Y. Gong, "Bayesian loss for crowd count estimation with point supervision," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6141–6150.

[9] Y. Zhang, D. Zhou, S. Chen, S. Gao, and Y. Ma, "Single-image crowd counting via multi-column convolutional neural network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 589–597, doi: 10.1109/CVPR.2016.70.

[10] Y. Li, X. Zhang, and D. Chen, "CSRNet: Dilated convolutional neural networks for understanding the highly congested scenes," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1091–1100, doi: 10.1109/CVPR.2018.00120.

[11] L. Liu et al., "DENet: A universal network for counting crowd with varying densities and scales," *IEEE Trans. Multimedia*, vol. 23, pp. 1060–1068, 2021, doi: 10.1109/TMM.2020.2992979.

[12] D. B. Sam, S. V. Peri, M. N. Sundararaman, A. Kamath, and R. V. Babu, "Locate, size, and count: Accurately resolving people in dense crowds via detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 8, pp. 2739–2751, Aug. 2021, doi: 10.1109/TPAMI.2020.2974830.

[13] X. Chen, Y. Bin, N. Sang, and C. Gao, "Scale pyramid network for crowd counting," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2019, pp. 1941–1950.

[14] X. Cao, Z. Wang, Y. Zhao, and F. Su, "Scale aggregation network for accurate and efficient crowd counting," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 734–750, doi: 10.1007/978-3-030-01228-1_45.

[15] J. Gao, Q. Wang, and Y. Yuan, "SCAR: Spatial-/channel-wise attention regression networks for crowd counting," *Neurocomputing*, vol. 363, pp. 1–8, Oct. 2019, doi: 10.1016/j.neucom.2019.08.018.

[16] S. Liang, Z. Huang, M. Liang, and H. Yang, "Instance enhancement batch normalization: An adaptive regulator of batch noise," in *Proc. 34th AAAI Conf. Artif. Intell.*, 2019, pp. 1–9, doi: 10.1609/AAAI.V34I04.5917.

[17] X. Guo, M. Anisetti, M. Gao, and G. Jeon, "Object counting in remote sensing via triple attention and scale-aware network," *Remote Sens.*, vol. 14, no. 24, p. 6363, Dec. 2022. [Online]. Available: https://www.mdpi.com/2072-4292/14/24/6363

[18] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141, doi: 10.1109/CVPR.2018.00745.

[19] G. Olmschenk, H. Tang, and Z. Zhu, "Improving dense crowd counting convolutional neural networks using inverse k-nearest neighbor maps and multiscale upsampling," in *Proc. 15th Int. Joint Conf. Comput. Vis., Imag. Comput. Graph. Theory Appl.*, 2020, pp. 1–11, doi: 10.5220/0009156201850195.

[20] F. Dai, H. Liu, Y. Ma, X. Zhang, and Q. Zhao, "Dense scale network for crowd counting," in *Proc. Int. Conf. Multimedia Retr.*, Aug. 2021, pp. 64–72, doi: 10.1145/3460426.3463628.

[21] L. Zhu, Z. Zhao, C. Lu, Y. Lin, Y. Peng, and T. Yao, "Dual path multi-scale fusion networks with attention for crowd counting," 2019, *arXiv:1902.01115*.

[22] K. Khan et al., "Crowd counting using end-to-end semantic image segmentation," *Electronics*, vol. 10, no. 11, p. 1293, May 2021.

[23] Y. Meng et al., "Spatial uncertainty-aware semi-supervised crowd counting," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 15529–15539.

[24] J. Gao, Q. Wang, and X. Li, "PCC Net: Perspective crowd counting via spatial convolutional network," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 10, pp. 3486–3498, Oct. 2020, doi: 10.1109/TCSVT.2019.2919139.

[25] Y. Liu, L. Liu, P. Wang, P. Zhang, and Y. Lei, "Semi-supervised crowd counting via self-training on surrogate tasks," 2020, *arXiv:2007.03207*.

[26] J. Wang et al., "Deep high-resolution representation learning for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 10, pp. 3349–3364, Oct. 2021, doi: 10.1109/TPAMI.2020.2983686.

[27] C. Dong, C. C. Loy, and X. Tang, "Accelerating the super-resolution convolutional neural network," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 391–407, doi: 10.1007/978-3-319-46475-6_25.

[28] D. Liang, W. Xu, Y. Zhu, and Y. Zhou, "Focal inverse distance transform maps for crowd localization," *IEEE Trans. Multimedia*, early access, Sep. 2, 2022, doi: 10.1109/TMM.2022.3203870.

[29] H. Idrees, I. Saleemi, C. Seibert, and M. Shah, "Multi-source multi-scale counting in extremely dense crowd images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 2547–2554, doi: 10.1109/CVPR.2013.329.

[30] H. Idrees et al., "Composition loss for counting, density map estimation and localization in dense crowds," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 544–559, doi: 10.1007/978-3-030-01216-8_33.

[31] V. A. Sindagi, R. Yasarla, and V. M. Patel, "JHU-CROWD++: Large-scale crowd counting dataset and a benchmark method," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 5, pp. 2594–2609, Nov. 2022.

[32] Q. Wang, J. Gao, W. Lin, and X. Li, "NWPU-crowd: A large-scale benchmark for crowd counting and localization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 6, pp. 2141–2149, Jun. 2021, doi: 10.1109/TPAMI.2020.3013269.

[33] M.-R. Hsieh, Y.-L. Lin, and W. H. Hsu, "Drone-based object counting by spatially regularized regional proposal network," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4165–4173.

[34] R. Guerrero-Gómez-Olmedo, B. Torre-Jiménez, R. J. López-Sastre, S. Maldonado-Bascón, and D. Oñoro-Rubio, "Extremely overlapping vehicle counting," in *Proc. Iberian Conf. Pattern Recognit. Image Anal.*, 2015, pp. 423–431, doi: 10.1007/978-3-319-19390-8_48.

[35] W. Zhai et al., "DA^{2}Net: A dual attention-aware network for robust crowd counting," *Multimedia Syst.*, Jan. 2022, doi: 10.1007/s00530-021-00877-4.

[36] X. Jiang et al., "Crowd counting and density estimation by trellis encoder–decoder networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 6126–6135, doi: 10.1109/CVPR.2019.00629.

[37] Q. Wang, J. Gao, W. Lin, and Y. Yuan, "Learning from synthetic data for crowd counting in the wild," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 8190–8199, doi: 10.1109/CVPR.2019.00839.

[38] D. B. Sam, S. Surya, and R. V. Babu, "Switching convolutional neural network for crowd counting," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4031–4039, doi: 10.1109/CVPR.2017.429.

[39] S. Amirgholipour, X. He, W. Jia, D. Wang, and M. Zeibots, "A-CCNN: Adaptive CCNN for density estimation and crowd counting," in *Proc. 25th IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2018, pp. 948–952, doi: 10.1109/ICIP.2018.8451399.

[40] C. Liu, X. Weng, and Y. Mu, "Recurrent attentive zooming for joint crowd counting and precise localization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1217–1226, doi: 10.1109/CVPR.2019.00131.

[41] M. H. Oh, P. Olsen, and K. N. Ramamurthy, "Crowd counting with decomposed uncertainty," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, 2020, pp. 11799–11806, doi: 10.1609/AAAI.V34I07.6852.

[42] T. N. Mundhenk, G. Konjevod, W. A. Sakla, and K. Boakye, "A large contextual dataset for classification, detection and counting of cars with deep learning," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 785–800, doi: 10.1007/978-3-319-46487-9_48.

[43] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 779–788, doi: 10.1109/CVPR.2016.91.

[44] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017, doi: 10.1109/TPAMI.2016.2577031.

[45] W. Liu et al., "SSD: Single shot MultiBox detector," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2016, pp. 21–37, doi: 10.1007/978-3-319-46448-0_2.

[46] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 2, pp. 318–327, Feb. 2020.

[47] X. Zeng, Y. Wu, S. Hu, R. Wang, and Y. Ye, "DSPNet: Deep scale purifier network for dense crowd counting," *Exp. Syst. Appl.*, vol. 141, Mar. 2020, Art. no. 112977.

[48] V. A. Sindagi and V. M. Patel, "CNN-based cascaded multi-task learning of high-level prior and density estimation for crowd counting," in *Proc. 14th IEEE Int. Conf. Adv. Video Signal Based Surveill. (AVSS)*, Aug. 2017, pp. 1–6, doi: 10.1109/AVSS.2017.8078491.

[49] W. Liu, M. Salzmann, and P. Fua, "Context-aware crowd counting," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5094–5103, doi: 10.1109/CVPR.2019.00524.

[50] L. Fiaschi, U. Koethe, R. Nair, and F. A. Hamprecht, "Learning to count with regression forest and structured labels," in *Proc. 21st Int. Conf. Pattern Recognit. (ICPR)*, Nov. 2012, pp. 2685–2688.

[51] D. Oñoro-Rubio and R. López-Sastre, "Towards perspective-free object counting with deep learning," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2016, pp. 615–629, doi: 10.1007/978-3-319-46478-7_38.

[52] Y. Liu, M. Shi, Q. Zhao, and X. Wang, "Point in, box out: Beyond counting persons in crowds," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 6462–6471, doi: 10.1109/CVPR.2019.00663.

Xiangyu Guo is currently pursuing the M.S. degree with the School of Electrical and Electronic Engineering, Shandong University of Technology, Zibo, China. His research interests include smart city systems, computer vision, and deep learning.

Mingliang Gao received the Ph.D. degree in communication and information systems from Sichuan University. He is currently an Associate Professor and the Vice Dean of the Shandong University of Technology. He was a Visiting Lecturer with The University of British Columbia (2018–2019). He has been a Principal Investigator for a variety of research funding, including the National Natural Science Foundation, China Post-Doctoral Foundation, and National Key Research Development Project. He has published over 170 journals/conference papers in IEEE, Springer, Elsevier, and Wiley. His research interests include computer vision, machine learning, and intelligent optimal control.

Wenzhe Zhai is currently pursuing the M.S. degree with the School of Electrical and Electronic Engineering, Shandong University of Technology, Zibo, China. His research interests include smart city systems, information fusion, crowd analysis, and deep learning.

Qilei Li (Student Member, IEEE) received the M.S. degree from Sichuan University in 2020. He is currently pursuing the Ph.D. degree in computer science with the Queen Mary University of London under the supervision of Prof. Shaogang (Sean) Gong. His research interests include computer vision and deep learning, particularly focusing on person ReID and video/image enhancement. He serves as a reviewer for *Information Fusion*, IEEE TRANSACTIONS ON INSTRUMENTATION AND MEASUREMENT, IEEE ACCESS, *Concurrency and Computation: Practice and Experience*, and *Multimedia System*.

Gwanggil Jeon (Senior Member, IEEE) received the B.S., M.S., and Ph.D. (summa cum laude) degrees from the Department of Electronics and Computer Engineering, Hanyang University, Seoul, South Korea, in 2003, 2005, and 2008, respectively. From September 2009 to August 2011, he was with the School of Information Technology and Engineering, University of Ottawa, Ottawa, ON, Canada, as a Post-Doctoral Fellow. From September 2011 to February 2012, he was with the Graduate School of Science and Technology, Niigata University, Niigata, Japan, as an Assistant Professor. From December 2014 to February 2015 and June 2015 to July 2015, he was a Visiting Scholar with Centre de Mathématiques et Leurs Applications (CMLA), École Normale Supérieure Paris-Saclay (ENS-Cachan), France. From 2019 to 2020, he was a Prestigious Visiting Professor with Dipartimento di Informatica, Universitá degli Studi di Milano Statale, Italy. He is currently a Full Professor with Incheon National University, Incheon, South Korea. He was a Visiting Professor with Sichuan University, China; Universitat Pompeu Fabra, Barcelona, Spain; Xinjiang University, China; King Mongkut's Institute of Technology Ladkrabang, Bangkok, Thailand; and University of Burgundy, Dijon, France. He is an Associate Editor of *Sustainable Cities and Society*, IEEE ACCESS, *Real-Time Image Processing*, *Journal of System Architecture*, and *Remote Sensing* (MDPI). He was a recipient of the IEEE Chester Sall Award in 2007, the ETRI Journal Paper Award in 2008, and Industry-Academic Merit Award by the Ministry of SMEs and Startups of Korea Minister in 2020.