Region-Aware Quantum Network for Crowd Counting

Wenzhe Zhai, Xianglei Xing, Gwanggil Jeon

***Abstract*—Crowd counting has substantial practical applications in various consumer-oriented areas, particularly for safety assessments and marketing strategies. However, considering the complexities of the capturing conditions, the unavoidable background interference possesses the potential to disrupt the effectiveness of established counting methods, and it further poses degraded counting performance. To address this challenge, we propose a Region-Aware Quantum Network (RAQNet) by attentively learning from the crowd region. It consists of four key components, namely the feature extractor, the object region awareness module (ORA), the quantum-driven calibration (QDC) module, and the decoder module. The cascaded ORA modules are engineered for the extraction of local information, which addresses background interference. Additionally, two QDC modules are incorporated to capture global information, which utilizes quantum states to calibrate features. Extensive experimental results conducted on four crowd benchmark datasets and three cross-domain datasets prove that the RAQNet outperforms the state-of-the-art competitors, both subjectively and objectively.**

***Keywords*—Consumption, Crowd counting, Quantum Network, Convolution neural network, Regional attention**

I. Introduction

In smart city systems, crowd counting plays a pivotal role in aiding business operators in understanding traffic patterns [1]–[3]. Businesses can efficiently allocate resources through the analysis of peak hours and popular zones within a store or mall. By identifying areas frequented by customers, stores can strategically position products and displays. This optimization enhances the customer experience by facilitating the easy location of desired items, which results in increased satisfaction and potentially more purchases [4]. Simultaneously, the field of crowd counting is continuously evolving and expanding its applications. It can be applied to high-level vision tasks, such as crowd tracking and 3D human pose estimation [5]–[7].

The crowd counting methods can be categorized into three classes, namely detection-based [8], regression-based [9], and density estimation-based methods [7], [10], [11]. Remarkably, the density estimation-based method leverages the robust feature extraction capabilities of Convolutional Neural Networks (CNNs) to generate a density map. Crowd counting is then accomplished by summing the pixel values on the density map [12], [13]. These methods exhibit superior accuracy and stability compared with the other two types of methods. However, the chaotic background interference often misleads the model to recognize the object areas incorrectly and leads to the overestimation or underestimation of the counting results. Some examples of background interference are illustrated in Fig. 1. The presence of background noise in the crowd image, *e.g.,* billboards and banners, will mislead the model into incorrectly identifying heads. To address this challenge, several attention-based algorithms have been introduced to filter the target region from the background. Sindagi *et al.* [14] proposed a spatial attention module to merge segmentation features and several global attention modules, thereby enhancing information interaction among different channels. Gao *et al.* [15] developed a foreground-background segmentation module to highlight the crowd region. Guo *et al.* [16] introduced a triple-view attention module to adjust weights in both spatial and channel dimensions. Zhai *et al.* [2] introduced the concept of channel-spatial self-attention to bolster the capacity of the model to adapt to crowd dynamics.

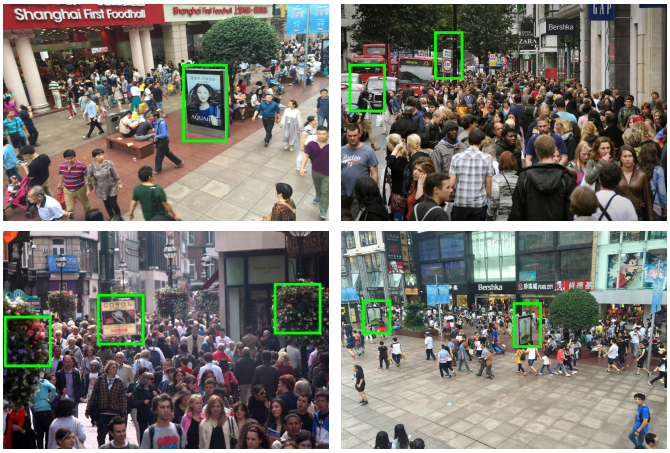

Fig. 1. Challenge of background interference in crowd scenarios.

Despite persistent efforts, performance remains unsatisfactory. The main reason is that the previous methods focus on global features and ignore the local features, which are crucial to distinguish the crowd from the background. In this paper, we proposed a Region-Aware Quantum Network (RAQNet) to learn regionally focused crowds. The proposed RAQNet consists of four parts, *i.e.*, feature extractor, ORA module, QDC module, and decoder. First, a feature extractor is built

The work is supported by the National Natural Science Foundation of China No. 62076078 and the CAAI-Huawei MindSpore Open Fund No. CAAIXSJLJJ-2020-033A. (Corresponding authors: Xianglei Xing)

Wenzhe Zhai and Xianglei Xing are with the College of Intelligent Science and Engineering, Harbin Engineering University, Harbin, 150001, China. (e-mail: wenzhezhai@163.com and xingxl@hrbeu.edu.cn.)

Gwanggil Jeon is with the Department of Embedded Systems Engineering, Incheon National University, Incheon 22012, South Korea (e-mail:ggjeon@gmail.com).

to extract the basic features. Subsequently, the ORA module captures localized information around the crowd region and utilizes the regional attention (RA) unit to mitigate background noise. Furthermore, the QDC module calibrates input features by employing a quantum attention (QA) unit to compute global features via qubit rotation in tandem with the Pauli operator. Six ORA modules and two QDC modules highlight the crucial information within local and global contexts in a mutual-promotion manner to promote counting accuracy. At last, the decoder utilizes multiple transposed convolution to predict the density map. In a nutshell, the contributions of this paper are as follows.

1) We proposed an RAQNet to alleviate the background cluster for dense crowd counting.
2) We built an ORA module to extract the local information and combine it with the RA unit to suppress background interference.
3) We designed a QDC module to capture the global information and integrated the QA unit to calibrate features for increasing generalization performance.
4) Comprehensive experiments on four crowd benchmark datasets and three cross-domain datasets are conducted to verify the superiority and generalization of the RAQNet.

II. Related Work

A. CNN-based crowd counting

Crowd counting is a dedicated field focused on the precise estimation of the number of individuals [17]. Zhang *et al.* [18] developed a lightweight model with three columns to facilitate multi-scale feature extraction. Sindagi *et al.* [14] proposed a hierarchical attention-based network termed HA-CCN. This network comprises a spatial attention module designed for the fusion of segmentation features and a sequence of global attention modules aimed at enhancing information interaction across the channel dimension. Gao *et al.* [15] developed a fore-background segmentation module to generate a segmentation map. This map serves to highlight the object region while simultaneously mitigating the influence of background clutter. Sam *et al.* [19] presented a locate-size-count network termed LSC-CNN to obtain multiscale information and incorporated a Grid Winner-Take-All (GWTA) loss function to mitigate the problem of local minima. Guo *et al.* [20] integrated a triple-view attention module to address background disturbances and incorporated a pyramid feature aggregation module to extract multiscale features. Kilic *et al.* [21] introduced a heatmap learner CNN (HLCNN) for car counting tasks. The HLCNN employs a customized VGG architecture as its backbone, with the final three fully connected layers substituted with three convolutional adaptation layers.

B. Quantum machine learning

Quantum machine learning is an emerging discipline that leverages the principles of quantum mechanics to enhance traditional machine learning algorithms. By harnessing qubits and quantum operations, quantum machine learning demonstrates the capability to perform complex computations exponentially faster than classical computers. This empowers the exploration of new algorithms, optimization methods, and data analysis techniques.

Recent research has witnessed the development of quantum counterparts to classical models, exemplified by the Quantum Regressor [22] and Quantum Variational Autoencoder [23]. The primary objective is to exploit the distinctive characteristics of spatial transformations within quantum circuits when incorporated into the realm of machine learning. The introduction of the Quantum network aims to capitalize on its beneficial spatial feature transformations and its capability for modeling intricate functional relationships. Nguyen *et al.* [24] introduced a quantum neural network (QNN) that exhibits resilience to both noise and decoherence. The introduction of the QNN module presents solutions to challenges related to population counting. Besides, the quantum network is applied in the crowd counting domain. It can exploit the quantum superposition and entanglement to achieve higher-dimensional and non-linear feature representation, which can improve the accuracy and robustness of crowd counting in complex scenarios. The Quantum Image Feature Extraction with Dense Distribution-Aware Learning (QE-DAL) [25] framework was proposed by Hu *et al.* to address challenges arising in densely populated crowd scenes. The multi-scale module of QE-DAL acquires four distinct feature maps by employing convolutional blocks of different scales, which are subsequently integrated to create a density map. In addition, Hu *et al.* [26] employed a refined distance compensating with a quantum scale-aware learning framework (RDC-SAL) to tackle crowd counting and localization tasks. The RDC-SAL incorporates a dedicated quantum layer into its front-end feature extraction module. Therefore, this model enables the efficient extraction of densely populated crowd features of varying sizes.

III. Method

A. Overview

The architecture of the proposed RAQNet is illustrated in Fig 2. It consists of four components, *i.e.*, feature extractor, object region awareness (ORA) module, quantum-driven calibration (QDC) module, and decoder. For the input image I, the feature extractor adopts VGG-19 to extract the low-level feature and generate a feature map. Afterward, an ORA module extracts the local information around the crowd region and utilizes the regional attention (RA) unit to suppress background noise. In addition, a QDC module is integrated with the quantum attention (QA) unit, which utilizes the qubit rotation in combination with Pauli operators to calibrate the feature. Then, six ORA modules and two QDC modules are stacked to highlight the crucial information within local and global contexts. The final decoder is applied to increase the resolution of the enhanced feature map and make predictions for the density map.

B. Object Region Awareness Module

In the crowd counting domain, the task of quantifying individuals within intricate background scenes is a common

This article has been accepted for publication in IEEE Transactions on Consumer Electronics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCE.2024.3378166

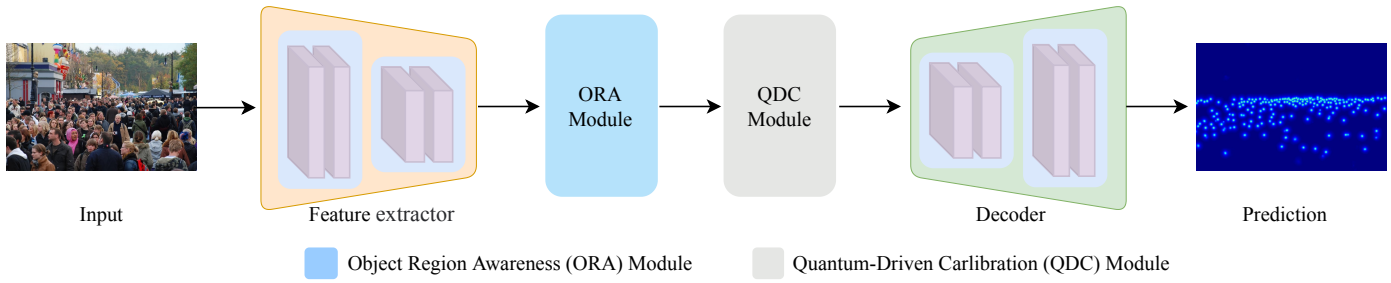

Fig. 2. Architecture of the proposed RAQNet for crowd counting.

challenge. To tackle this problem, we utilize an ORA module to focus on the foreground elements while disregarding the complexities of the background. Specifically, the ORA module incorporates a regional attention (RA) unit for foreground-related feature learning. Meanwhile, the residual connection mechanism is employed to ensure the preservation of the original feature information. Subsequently, the "add & norm" operation is employed to implement residual connections and layer normalization within the ORA module. The feed-forward (FF) unit performs non-linear channel dimension mapping by employing a 1×1 convolutional layer, thereby enhancing the effectiveness of feature representation. To extract positional features between different features, we build the positional encoding generator (PEG) between two consecutive ORA modules. Meanwhile, the proposed module employs a 3×3 convolution operation to extract spatial features. It is essential to emphasize the application of deep separable convolution within the PEG framework, where individual convolution operations are executed independently for each channel.

The RA unit aims to guide the focus of the proposed network toward the pertinent human subjects, rather than intricate and irrelevant backgrounds. Additionally, it facilitates the model in gaining a comprehensive understanding of the correlations and density distribution among distinct regions. The architecture of the RA unit is shown in Fig. 3. Given an input feature map, a series of 1×1 convolution operations are employed to generate query (Q), key (K), and value (V) matrices. Simultaneously, the dimension of the channel is halved to reduce the computational demands on the model and enhance the computational efficiency. Then, an ensemble of eight attention heads is employed to acquire diverse attention weights. It enables the model to focus more effectively on the regions of interest. The regional attention is formulated as,

$$R(Q, K, V) = S\left(\frac{Q \cdot K^T}{\sqrt{d_k}}\right) \cdot V, \tag{1}$$

where $S(\cdot)$ represents a softmax activation function. d_k denotes the dimension associated with each attention head and it is set to 32.

C. Quantum-Driven Carlibration Module

While the ORA module effectively mitigates background noise, it inevitably sacrifices positional accuracy and captures only a limited amount of global perceptual information. To

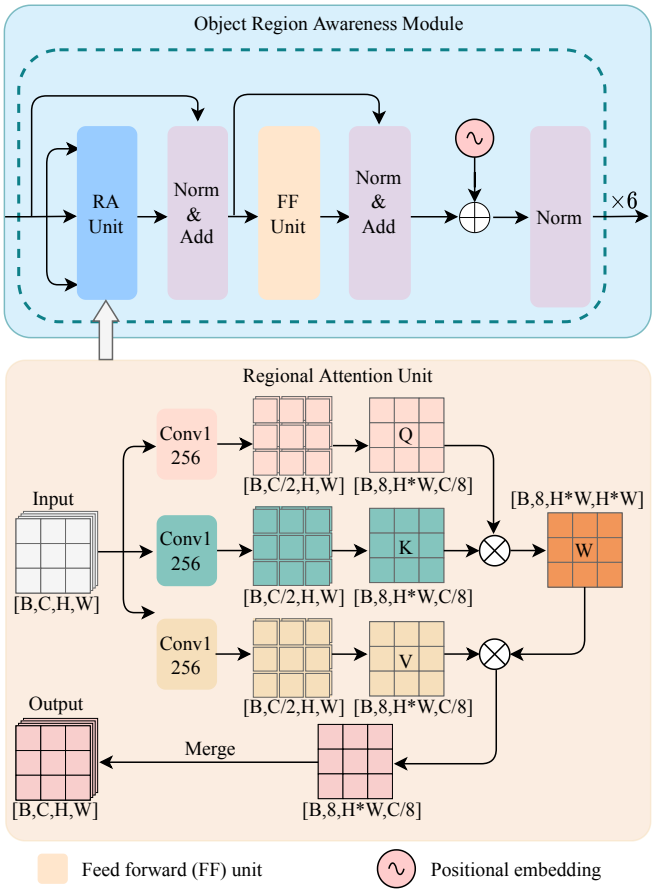

Fig. 3. Architecture of the ORA module.

enhance the refinement of information extracted from the ORA module, we use the quantum convolutional neural network to transform feature maps into quantum states. The diagram of the QDC module is illustrated in Fig. 4. It exhibits structural similarities with the ORA module which contains the residual structure, the FF unit, and the positional code generator. The major difference is the embedding of the QA unit in the QDC module. The QA unit comprises three key components, namely the quantum encoder, quantum circuits, and the quantum decoder. Firstly, the input feature map f_x is encoded through

This article has been accepted for publication in IEEE Transactions on Consumer Electronics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCE.2024.3378166

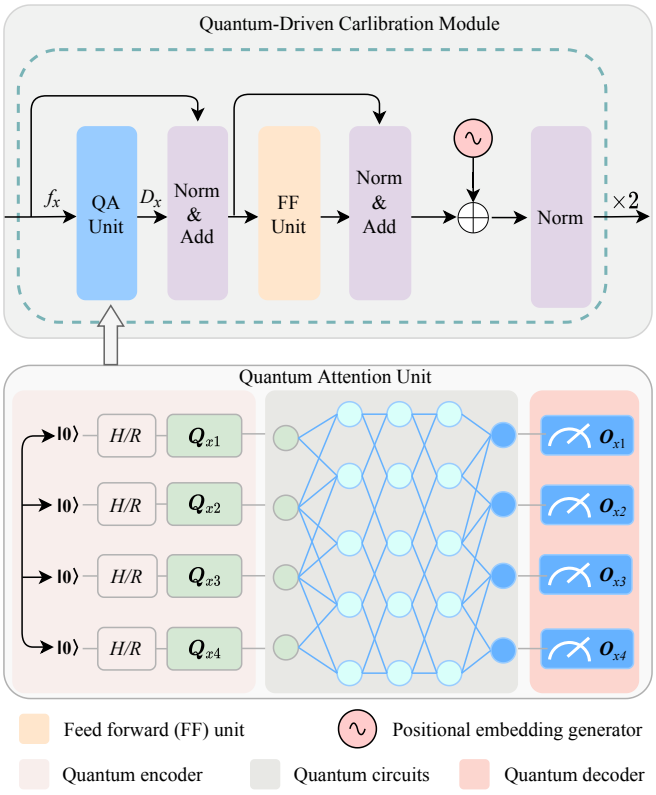

Fig. 4. Architecture of the QDC module.

the input channels within the encoder layer. The Hadamard H gate is applied to each quantum bit (qubit), which transforms each qubit from the ground state to a superposition state and places them in a superposition of $|0\rangle$ and $|1\rangle$. Subsequently, the R gate is applied to each qubit. The R gate is a rotation gate, with its rotation angle being determined by the respective values in the input data [24]. This step rotates each qubit in the superposition state. It is formulated as follows,

$$\boldsymbol{Q}_x = \cos(f_x)\,|0\rangle + \sin(f_x)\,|1\rangle, \tag{2}$$

where f_x denotes the input feature map. $\boldsymbol{Q}_x$ represents the initial quantum state. $\{\cos(\cdot), \sin(\cdot)\}$ represents the qubit rotation gate.

The quantum circuit is triggered by the initial state $\boldsymbol{Q}_x$. Subsequently, the quantum filter can calculate quantum states utilizing the sliding window. Furthermore, random parameters are selected from a uniform distribution and utilized to calculate the expected output value $\boldsymbol{O}_x$. It is facilitated using the Pauli operators [27], which are denoted as follows,

$$c(\theta) = e^{-ih\theta}, \tag{3}$$

$$\boldsymbol{O}_x = c(\theta) \cdot |F(\boldsymbol{Q}_x \cdot W_{m,n})\rangle\,, \tag{4}$$

where $\boldsymbol{O}_x$ represents output quantum stats. $c(\theta)$ denotes the parametric quantum circuit. "$-i$" represents the imaginary unit. h represents the Hamiltonian operator, which describes the energy of the system [28]. θ is a vector of parameters. $F(\cdot)$ employs the quantum filter, which uses the parameterized sliding window $W_{m,n}$ with dimensions $m \times n$.

The quantum decoder primarily serves the purpose of aligning the output of the quantum circuit layer with the expected output of a typical convolution layer. In this research, the sigmoid operator is employed to compute the final decoded state. It is formulated as follows,

$$D_x = \mathrm{Sig}(\boldsymbol{O}_x), \tag{5}$$

where $\mathrm{Sig}(\cdot)$ represents the sigmoid function.

D. Loss Function

The MSE loss is deployed to optimize the model during the training stage. It is represented as follows,

$$L_{\mathrm{MSE}} = \frac{1}{N}\sum_{i=1}^{N}\left\|F_i^{est} - F_i^{gt}\right\|_2^2, \tag{6}$$

where N denotes the overall headcount. F_i^{est} and F_i^{gt} represent the estimated and the ground-truth count of the i-th image. $\|\cdot\|_2^2$ represents Euclidean norm squared.

IV. Experiments and Analysis

A. Implementation Details

The training samples are subjected to random cropping, generating dimensions of 256×256 for the ShanghaiTech and Ship datasets, and 512×512 for other datasets. This disparity in dimensions is attributed to the smaller image size of the ShanghaiTech dataset. Network optimization is facilitated by the Adam optimizer, initialized with a learning rate of 10^{-4} and a weight decay rate of 5×10^{-4}. The batch size is set to 16 during the training process. The experimental framework is founded on the MindSpore platform, with computational support provided by the NVIDIA 3090 GPU.

B. Datasets

ShanghaiTech [18] is segmented into two categories: SHA and SHB. The SHA dataset includes randomly sourced images from the internet with a wide range of resolutions. The SHB dataset comprises images obtained from surveillance devices installed along a street in Shanghai, and they are characterized by a uniform resolution. SHA displays higher crowd density compared to SHB.

UCF_CC_50 [29] is a densely crowded dataset featuring a wide array of scenarios, including crowd parades, sporting events, and religious assemblies. The images within the dataset are in greyscale, and certain samples contain only head information, which makes accurate counting considerably challenging.

UCF-QNRF [30] is one of the most challenging datasets, distinguished by its extensive variation in scale and diverse perspectives. The dataset comprises high-quality images that impose a substantial computational load on GPUs, which necessitate resizing during the training process.

JHU-Crowd++ [31] is a relatively new dataset that encompasses various challenges, such as weather variations and scale variations. Additionally, it contains a considerably large number of samples, making it a challenging dataset.

CARPK [32] is specifically designed for car counting, and the images are captured from various parking lots in a bird's-eye view using drones.

PUCPR+ [32] is another vehicle counting dataset, which contains fewer samples compared to CARPK. It includes diverse weather conditions, adding an extra layer of complexity to the counting task.

RSOC [33] is subdivided into four distinct categories, namely, buildings, large vehicles, small vehicles, and ships, taking into account various object attributes. **Building sub-dataset** is composed of fixed-size images with relatively low resolution, obtained from Google Earth. These images exhibit a high density of architectural structures within the dataset. **Ship sub-dataset** comprises a collection of high-resolution images that primarily depict small watercraft. These images exhibit a range of characteristics, including diverse orientations, non-uniform spatial distribution within the dataset, and substantial variations in terms of their overall scale.

C. Evaluation Metrics

The Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) are adopted to measure the counting performance. They are defined as,

$$\text{MAE} = \frac{1}{N}\left|\hat{c}_i - c_i\right|, \tag{7}$$

$$\text{RMSE} = \sqrt{\frac{1}{N}\left|\hat{c}_i - c_i\right|^2}, \tag{8}$$

where N denotes the number of test images. $\hat{c}_i$ and c_i represent the estimated value and ground truth of the i-th image, respectively. A lower value of the two criteria represents better counting accuracy and robustness of the model.

D. Comparison with State-of-the-art Methods

To validate the effectiveness of the proposed method, we conducted comparative experiments with 12 state-of-the-art (SOTA) methods on four distinct datasets. The objective comparison results are shown in Table I. In the densely populated SHA dataset, the proposed RAQNet attained scores of 59.0 and 101.2 on MAE and RMSE, respectively. Due to the scene diversity of SHA, the proposed model cannot adapt well to all scenarios leading to a relatively high RMSE. Compared with the second-ranked TEDNet [10], the proposed network demonstrates a decrease of 8.1% in the MAE and a 7.2% in the RMSE. Compared with the fusion approach that involves multiple models LSC-CNN [19], the proposed RAQNet leads to an improvement in both MAE and RMSE, with increments of 11.1% and 13.5%, respectively. On the relatively sparse Part B dataset, the RAQNet attains an MAE of 9.0. Meanwhile, the value of RMSE is 15.4. Similar to the results on SHA, the MSE and RMSE are not optimal on SHB. The reason lies that the SHB is captured by surveillance cameras on the street, and the irregular placement results in perspective distortions.

When evaluated on the densely populated UCF_CC_50 dataset, the proposed RAQNet demonstrates remarkable competitiveness, and it achieves MAE and RMSE scores of 177.1 and 247.6, respectively. In comparison to the second-best PCC-Net [15], the RAQNet shows a remarkable 26.2% improvement in MAE and a corresponding 21.5% improvement in RMSE. The results indicate that the proposed method performs well in densely populated scenarios.

Compared to other SOTA methods, the proposed RAQNet exhibits substantial competitiveness on the UCF-QNRF dataset. In contrast to the TEDNet [10], the proposed RAQNet exhibits remarkable performance enhancements. The RAQNet delivers substantial improvements of 5.7% and 1.0% in MAE and RMSE, respectively. Compared with the DFNet [34] which leverages the hierarchical features to address the background noise, the RAQNet archives 51.2% and 47.9% improvement in MAE and RMSE. The objective results demonstrate the efficacy of the proposed network in addressing the challenge of background noise.

In the JHU++ dataset, which encompasses crowd-counting data collected in diverse scenarios, the proposed RAQNet exhibits strong performance. It scores 66.7 and 196.3 in MAE and RMSE, respectively. In comparison to the CG-DRCN [31], which is built for eliminating background interference, the proposed RAQNet demonstrates an improvement of 6.1% in MAE and 29.5% in RMSE.

The subjective results from the datasets are presented in Fig. 5. It shows that the estimated density maps and count numbers closely align with the ground truth. These results underscore the capability of the proposed method to accurately perform crowd counting even in complex backgrounds.

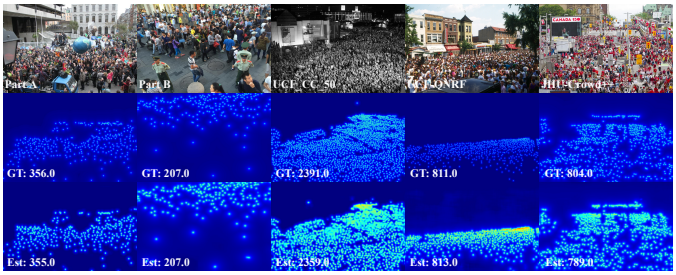

Fig. 5. Subjective results on cross-domain datasets.

E. Cross-domain Analysis

To further validate the generalization ability of the proposed method, we perform cross-domain analysis on two vehicle counting datasets and a remote sensing dataset, *i.e.*, CARPK, PUCPR+ [32], RSOC [33]. Table II exhibits competitive results between the proposed RAQNet and the state-of-the-art vehicle counting methods. The quantitive results prove that the RAQNet performs better than other competitors. Specifically, on the CARPK dataset, the proposed method scores 5.38 in MAE and 7.83 in RMSE. Compared to the second-ranked

TABLE I. 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 [18]	110.2	173.2	26.4	41.3	377.6	509.1	277.0	426.0	188.9	483.4
Switch-CNN [35]	90.4	135.0	21.6	33.4	318.1	439.2	-	-	-	-
A-CCNN [36]	85.4	124.6	19.2	31.5	-	-	367.3	-	171.2	453.1
SANet [37]	67.0	104.5	8.4	13.6	258.4	334.9	-	-	91.1	320.4
CSRNet [38]	68.2	115.0	10.6	16.0	266.1	397.5	-	-	85.9	309.2
RAZ [39]	65.1	106.7	8.4	14.1	-	-	116.0	195.0	-	-
TEDNet [10]	64.2	109.1	8.2	12.8	249.4	354.5	113.0	188.0	75.0	299.9
LSC-CNN [19]	66.4	117.0	**8.1**	**12.7**	-	-	120.5	218.2	112.7	454.4
SUA-Fully [40]	66.9	125.6	12.3	17.9	-	-	119.2	213.3	-	-
DFNet [34]	77.6	129.7	14.1	21.1	402.3	434.1	218.2	357.4	-	-
PCCNet [15]	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
RAQNet (Ours)	**59.0**	101.2	9.0	15.4	**177.1**	**247.6**	**106.5**	186.1	**66.7**	**196.3**

BL [41], the proposed RAQNet demonstrates a notable improvement of 43.8% in MAE and 31.1% in MAE and RMSE, respectively. Unlike the CARPK dataset, the PUCPR+ dataset exhibits a significant challenge due to the inconsistency in weather conditions, which poses a formidable obstacle for object counting. In comparison to these counterparts, the RAQNet under consideration exhibits a notable level of performance on the dataset. The objective results illustrate that the RAQNet achieves an MAE of 1.71 and an RMSE of 2.54, which performs best among the competitors.

The objective results of the proposed method on Building and Ship datasets are depicted in Table III. On the Building dataset, the RAQNet obtains the first place, and it achieves an MAE of 7.08 and an RMSE of 10.58. Compared with the suboptimal method TASNet [16], it shows a 7.2% reduction in MAE and a 6.0% reduction in RMSE. On the Ship dataset, the proposed method performs best compared with the competitors. It scores 79.98 and 188.76 on the metrics for MAE and RMSE, respectively.

Some visualization results on the CARPK, PUCPR+, and RSOC cross-domain datasets are illustrated in Fig. 6. The subjective results highlight the effectiveness of the proposed approach in accurately conducting vehicle counting. Besides, some visual comparison of RAQNet with other models (MCNN [18], SCAR [42], ASPDN [33]) on the ship dataset is illustrated in Fig. 7. It proves that the proposed method can accurately predict the number of identities with precise locating of objects.

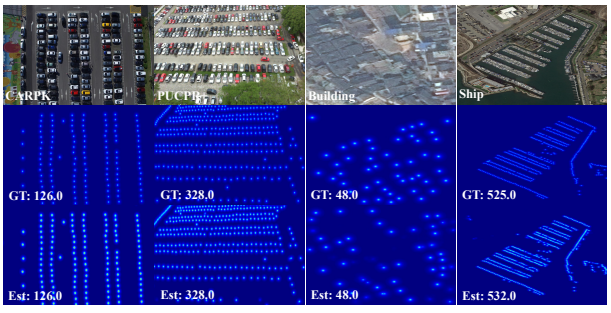

Fig. 6. Subjective results on cross-domain datasets.

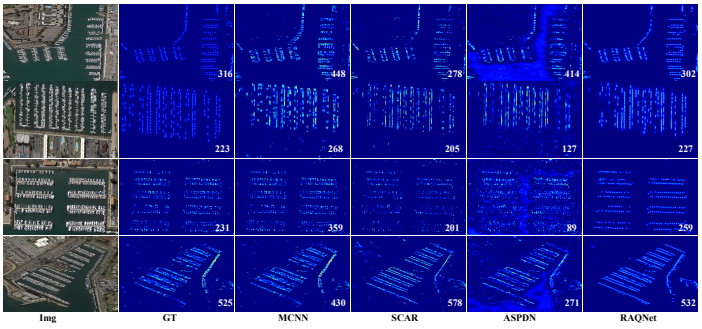

Fig. 7. Visual analysis with State-Of-The-Art (SOTA) approaches. The proposed RAQNet model adeptly determines the object locations and generates a precise density map. Compared with other SOTA models, RAQNet outperforms in counting objects across diverse scenarios.

TABLE II. COMPARATIVE 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 [45]	51.83	-	15.17	-
LPN [32]	23.80	36.79	22.76	34.46
RetinaNet [46]	16.62	22.30	24.58	33.12
One-Look Regression [47]	59.46	66.84	21.88	36.73
MCNN [18]	39.10	43.30	21.86	29.53
CSRNet [38]	11.48	13.32	8.65	10.24
BL [41]	9.58	11.38	6.54	8.13
PSGCNet [48]	8.15	10.46	5.24	7.36
TASNet [16]	7.16	10.23	5.16	6.76
RAQNet (Ours)	**5.38**	**7.83**	**1.71**	**2.54**

TABLE III. COMPARATIVE RESULTS ON THE RSOC DATASET. THE BEST RESULTS ARE HIGHLIGHTED IN **BOLD**.

Methods	Building		Ship	
	MAE	RMSE	MAE	RMSE
MCNN [18]	13.65	16.56	263.91	412.30
SANet [37]	29.01	32.96	302.37	436.91
CSRNet [38]	8.00	11.78	240.01	394.81
SCAR [42]	26.90	31.35	302.37	436.92
SPN [49]	7.74	11.48	241.43	392.88
CAN [50]	9.12	13.38	282.69	423.44
SFCN [51]	8.94	12.87	240.16	394.81
TASNet [16]	7.63	11.25	191.82	278.17
SFANet [52]	8.18	11.75	201.61	332.87
ASPDN [33]	7.54	10.66	193.83	318.95
RAQNet (Ours)	**7.08**	**10.58**	**79.98**	**188.76**

F. Efficiency comparison

To verify the efficiency of the proposed RAQNet, efficiency comparisons are carried out to measure the complexity of the calculations, *e.g.*, GFLOPs, inferring time and frame per second (FPS). The input size is set to 576×768. The comparison results are illustrated in Table IV. Comparative results prove that the proposed method RAQNet can achieve the best values of 250.8, 35.2, and 28.3 in GFLOPs, inferring time and FPS. Specifically, the RAQNet reduces the GFLOPs, inferring time and FPS by 2.1%, 2.5%, and 2.2% compared with the ASPDN [33] which employs a feature pyramid module and deformable convolution module to suppress clutter backgrounds.

TABLE IV. COMPARISON RESULTS OF THE RAQNET AND OTHER METHODS IN CALCULATIONS AND PARAMETERS.

Methods	GFLOPs	Inferring time (ms)	FPS
SASNet [53]	393.2	45.6	21.9
ASPDN [33]	256.2	36.1	27.7
SFCN [51]	274.1	39.9	25.1
RAQNet (Ours)	**250.8**	**35.2**	**28.3**

G. Ablation Studies

1) Ablation studies on the proposed ORA module: To assess the effectiveness of the proposed ORA module, we conducted ablation experiments on the combination of the ORA module. Table V shows the objective results of the ablation study of the ORA module on the SHA dataset. The notation "Number of ORA" indicates the introduction of ORA modules into the baseline, where "n" represents the number of ORA modules. According to Table V, the six ORA modules structure demonstrates an enhancement in system performance with the increasing number of incorporated ORA modules. When equipped with six ORA modules, the model achieves an MAE score of 60.0 and an RMSE score of 103.1. Compared with the baseline, the proposed model improves MSE and RMSE by 8.95% and 13.3%, respectively, when using six ORA modules. Along with increasing the number of ORA modules to eight, the performance of the proposed approach experiences a degradation, and the model scores 65.4 in MAE and 108.2 in RMSE. Moreover, the proposed network is based on the transformer architecture, which suffers from overfitting problems when the training data is insufficient. Therefore, we opt to integrate six ORA modules into the baseline to enhance system performance.

TABLE V. OBJECTIVE EXPERIMENTAL RESULTS OF THE ABLATION STUDY ON THE ORA MODULE. THE BEST RESULTS ARE HIGHLIGHTED IN **BOLD**.

Number of ORA	MAE	RMSE
0	65.9	114.3
2	62.4	110.8
4	61.5	107.2
6	**60.0**	**103.1**
8	65.4	108.2

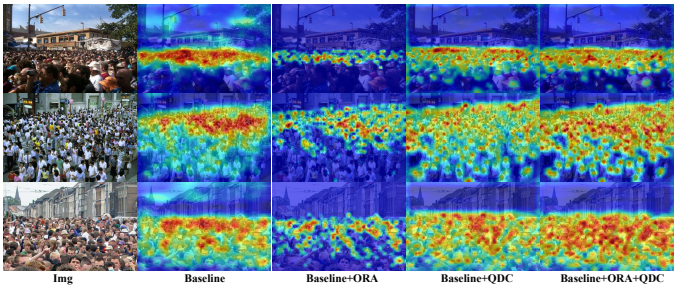

Fig. 8. The qualitative comparison of the baseline with different components.

2) Ablation studies on the proposed QDC module: To further evaluate the effectiveness of the QDC module, we conducted ablation experiments on the number of QDC modules within the context of six ORA modules. Table VI presents the ablation results of the QDC module quantities on the SHA dataset. In Table VI, the term "Number of ORA" signifies the integration of QDC modules into the baseline within the context of the six ORA modules. When the QDC module is integrated into the model, it is observed that the proposed model demonstrates a 1.1% and 0.3% improvement in the MSE and RMSE, respectively. This demonstrates that the independent integration of the QDC module can enhance the performance of the method. The results show that the ORA and QDC modules adeptly acquire both local and global information within the images, thereby resulting in improved model performance. Particularly, when using two QDC modules, it leads to a 4.1% increase in MAE and a 7.1% increase in RMSE compared to the baseline model.

TABLE VI. OBJECTIVE EXPERIMENTAL RESULTS OF THE ABLATION STUDIES ON THE PROPOSED QDC MODULE. (THE BEST RESULTS ARE HIGHLIGHTED IN **BOLD**.)

Number of QDC	MAE	RMSE
0	61.5	108.9
1	60.8	108.6
2	**59.0**	**101.2**
3	65.4	115.1

3) Visualization of the ORA module and QDC module: The visualization results of the baseline with different components are shown in Fig. 8. It shows that both the ORA module and QDC module are helpful in boosting the counting performance. The proposed ORA module can effectively extract local information to suppress the background noise. The QDC module can capture more global information to calibrate input features. Moreover, the problem of background clutter is alleviated by adding the two modules to the baseline, but the compound mode of "Baseline+ORA+QDC" (RAQNet) achieves the best results.

4) Ablation studies on backbone networks: In addition to investigating the efficacy of the proposed modules, we conduct ablation studies on the backbone networks. Three networks, namely HRNet [54], VGG-16 [55], and VGG-19, serve as the adopted backbones. Comparative results are presented

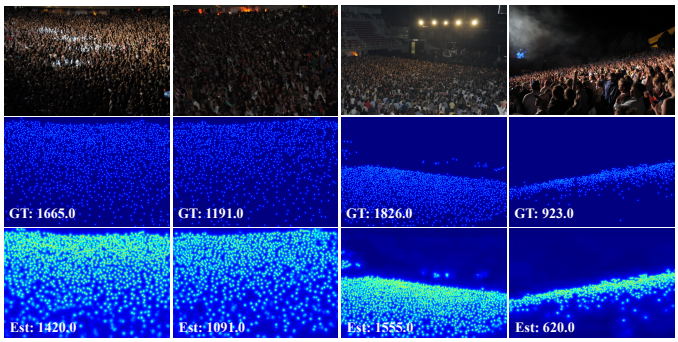

Fig. 9. The failure cases. The first row, the second row, and the third row depict the exemplars, the ground truth, and the estimated results, respectively.

in Table VII. It demonstrates that utilizing VGG-19 as the backbone network yields the most favorable performance. In fact, VGG is a commonly used backbone for feature extraction in numerous counting tasks [33], [38] owing to its robust generalization capability.

TABLE VII. COMPARATIVE RESULTS OF DIFFERENT BACKBONES ON THE SHIP DATASET. THE BEST PERFORMANCES ARE HIGHLIGHTED IN **BOLD**.

Methods	MAE	RMSE
RAQNet(HRNet)	65.2	111.2
RAQNet(VGG-16)	64.8	107.4
RAQNet(VGG-19)	**59.0**	**101.2**

H. Failure cases

While the RAQNet exhibits superior performance in dense crowd scenarios, some unsatisfactory outcomes are observed in challenging scenes, as illustrated in Fig. 9. It shows that low-light conditions cause the estimated density maps to include unwanted background noise. This is because the features of the head regions and the background regions are similar in dim-light environments. In future work, we will further investigate to explore more robust crowd feature map extraction methods for low-light environments.

V. CONCLUSION

In this paper, we proposed an RAQNet to address the problem of background interference for dense crowd counting. The RAQNet consists of a feature extractor, an ORA module, a QDC module, and a decoder. The feature extractor can abstract the low-level feature. The ORA module is combined with the RA unit to suppress the background noise. Meanwhile, the QDC module is devised based on the QA unit, and it can calibrate features through qubit rotation combined with the Pauli operator. The decoder is utilized to generate the prediction map. Experiments are carried out on four crowd counting datasets and three cross-domain datasets, and comparative results verify the superior counting performance of the proposed RAQNet.

REFERENCES

[1] D. Yang, X. Gao, L. Kong, Y. Pang, and B. Zhou, "An event-driven convolutional neural architecture for non-intrusive load monitoring of residential appliance," *IEEE Transactions on Consumer Electronics*, vol. 66, no. 2, pp. 173–182, 2020.

[2] W. Zhai, M. Gao, X. Guo, and Q. Li, "Scale-context perceptive network for crowd counting and localization in smart city system," *IEEE Internet of Things Journal*, vol. 10, no. 21, pp. 18 930–18 940, 2023.

[3] W. Zhai, M. Gao, M. Anisetti, Q. Li, S. Jeon, and J. Pan, "Group-split attention network for crowd counting," *Journal of Electronic Imaging*, vol. 31, no. 4, p. 041214, 2022.

[4] L. Dong, H. Zhang, D. Zhou, J. Shi, and J. Ma, "Cctwins: A weakly-supervised transformer-based crowd counting method with adaptive scene consistency attention," *IEEE Transactions on Consumer Electronics*, 2023.

[5] W. Zhai, M. Gao, A. Souri, Q. Li, X. Guo, J. Shang, and G. Zou, "An attentive hierarchy convnet for crowd counting in smart city," *Cluster Computing*, pp. 1–13, 2022.

[6] X. Guo, M. Gao, W. Zhai, Q. Li, K. H. Kim, and G. Jeon, "Dense attention fusion network for object counting in iot system," *Mobile Networks and Applications*, 2023.

[7] W. Zhai, M. Gao, Q. Li, G. Jeon, and M. Anisetti, "Fpanet: feature pyramid attention network for crowd counting," *Applied Intelligence*, pp. 1–18, 2023.

[8] I. S. Topkaya, H. Erdogan, and F. M. Porikli, "Counting people by clustering person detector outputs," in *2014 11th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, 2014, pp. 313–318.

[9] V. Lempitsky and A. Zisserman, "Learning to count objects in images," in *Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2010, pp. 1324–1332.

[10] X. Jiang, Z. Xiao, B. Zhang, X. Zhen, X. Cao, D. S. Doermann, and L. Shao, "Crowd counting and density estimation by trellis encoder-decoder networks," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6126–6135, 2019.

[11] W. Zhai, Q. Li, Y. Zhou, X. Li, J. Pan, G. Zou, and M. Gao, "Da2net: a dual attention-aware network for robust crowd counting," *Multimedia Systems*, 2022.

[12] S. Basalamah, S. D. Khan, and H. Ullah, "Scale driven convolutional neural network model for people counting and localization in crowd scenes," *IEEE Access*, vol. 7, pp. 71 576–71 584, 2019.

[13] S. D. Khan and S. Basalamah, "Scale and density invariant head detection deep model for crowd counting in pedestrian crowds," *The Visual Computer*, vol. 37, no. 8, pp. 2127–2137, 2021.

[14] V. A. Sindagi and V. M. Patel, "Ha-ccn: Hierarchical attention-based crowd counting network," *IEEE Transactions on Image Processing*, vol. 29, pp. 323–335, 2020.

[15] J. Gao, Q. Wang, and X. Li, "Pcc net: Perspective crowd counting via spatial convolutional network," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, pp. 3486–3498, 2020.

[16] X. Guo, M. Anisetti, M. Gao, and G. Jeon, "Object counting in remote sensing via triple attention and scale-aware network," *Remote Sensing*, vol. 14, no. 24, 2022.

[17] L. Dong, H. Zhang, K. Yang, D. Zhou, J. Shi, and J. Ma, "Crowd counting by using top-k relations: a mixed ground-truth cnn framework," *IEEE Transactions on Consumer Electronics*, vol. 68, no. 3, pp. 307–316, 2022.

[18] Y. Zhang, D. Zhou, S. Chen, S. Gao, and Y. Ma, "Single-image crowd counting via multi-column convolutional neural network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 589–597.

[19] 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 Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, pp. 2739–2751, 2021.

[20] 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, p. 6363, 2022.

[21] E. Kiliç and S. Ozturk, "An accurate car counting in aerial images based on convolutional neural networks," *Journal of Ambient Intelligence and Humanized Computing*, 2021.

[22] J. Shi, S. Chen, Y. Lu, Y. Feng, R. Shi, Y. Yang, and J. Li, "An approach to cryptography based on continuous-variable quantum neural network," *Scientific reports*, vol. 10, no. 1, p. 2107, 2020.

[23] N. Gao, M. Wilson, T. Vandal, W. Vinci, R. Nemani, and E. Rieffel, "High-dimensional similarity search with quantum-assisted variational autoencoder," in *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, 2020, pp. 956–964.

[24] N. H. Nguyen, E. C. Behrman, and J. E. Steck, "Quantum learning with noise and decoherence: a robust quantum neural network," *Quantum Machine Intelligence*, vol. 2, pp. 1–15, 2020.

[25] R. Hu, Z. Tang, and R. Yang, "Qe-dal: A quantum image feature extraction with dense distribution-aware learning framework for object counting and localization," *Applied Soft Computing*, vol. 138, p. 110149, 2023.

[26] R. Hu, Z.-R. Tang, E. Q. Wu, Q. Mo, R. Yang, and J. Li, "Rdc-sal: Refine distance compensating with quantum scale-aware learning for crowd counting and localization," *Applied Intelligence*, vol. 52, no. 12, pp. 14 336–14 348, 2022.

[27] R. Sarkar and E. van den Berg, "On sets of maximally commuting and anticommuting pauli operators," *Research in the Mathematical Sciences*, vol. 8, no. 1, p. 14, 2021.

[28] G. L. Light, "An introductory note to quantum modeling in finance: quantum finance," *European Journal of Applied Physics*, vol. 5, no. 5, pp. 16–20, 2023.

[29] H. Idrees, I. Saleemi, C. Seibert, and M. Shah, "Multi-source multi-scale counting in extremely dense crowd images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013, pp. 2547–2554.

[30] H. Idrees, M. Tayyab, K. Athrey, D. Zhang, S. Al-Maadeed, N. Rajpoot, and M. Shah, "Composition loss for counting, density map estimation and localization in dense crowds," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 532–546.

[31] V. A. Sindagi, R. Yasarla, and V. M. Patel, "Jhu-crowd++: Large-scale crowd counting dataset and a benchmark method," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 5, pp. 2594–2609, 2022.

[32] M.-R. Hsieh, Y.-L. Lin, and W. H. Hsu, "Drone-based object counting by spatially regularized regional proposal network," in *Proceedings of the International Conference on Computer Vision (ICCV)*, 2017, pp. 4165–4173.

[33] 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 Transactions on Geoscience and Remote Sensing*, vol. 59, pp. 3642–3655, 2021.

[34] S. D. Khan, Y. Salih, B. Zafar, and A. Noorwali, "A deep-fusion network for crowd counting in high-density crowded scenes," *International Journal of Computational Intelligence Systems*, vol. 14, no. 1, p. 168, 2021.

[35] D. B. Sam, S. Surya, and R. V. Babu, "Switching convolutional neural network for crowd counting," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 4031–4039.

[36] S. A. Kasmani, X. He, W. Jia, D. Wang, and M. Zeibots, "A-ccnn: Adaptive ccnn for density estimation and crowd counting," in *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, 2018, pp. 948–952.

[37] X. Cao, Z. Wang, Y. Zhao, and F. Su, "Scale aggregation network for accurate and efficient crowd counting," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 734–750.

[38] Y. Li, X. Zhang, and D. Chen, "Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 1091–1100.

[39] C. Liu, X. Weng, and Y. Mu, "Recurrent attentive zooming for joint crowd counting and precise localization," *CVPR*, pp. 1217–1226, 2019.

[40] Y. Meng, H. Zhang, Y. Zhao, X. Yang, X. Qian, X. Huang, and Y. Zheng, "Spatial uncertainty-aware semi-supervised crowd counting," in *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021, pp. 15 549–15 559.

[41] Z. Ma, X. Wei, X. Hong, and Y. Gong, "Bayesian loss for crowd count estimation with point supervision," *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6141–6150, 2019.

[42] J. Gao, Q. Wang, and Y. Yuan, "Scar: Spatial-/channel-wise attention regression networks for crowd counting," *Neurocomputing*, vol. 363, pp. 1–8, 2019.

[43] J. Redmon, S. K. Divvala, R. B. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 779–788.

[44] S. Ren, K. He, R. B. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, pp. 1137–1149, 2015.

[45] T. Stahl, S. L. Pintea, and J. C. V. Gemert, "Divide and count: Generic object counting by image divisions," *IEEE Transactions on Image Processing*, vol. 28, pp. 1035–1044, 2019.

[46] T.-Y. Lin, P. Goyal, R. B. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, pp. 318–327, 2020.

[47] 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 *ECCV*, 2016.

[48] 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 Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–12, 2022.

[49] X. Chen, Y. Bin, N. Sang, and C. Gao, "Scale pyramid network for crowd counting," in *Proceedings of the IEEE Workshop on Applications of Computer Vision (WACV)*, 2019, pp. 1941–1950.

[50] W. Liu, M. Salzmann, and P. Fua, "Context-aware crowd counting," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 5094–5103.

[51] Q. Wang, J. Gao, W. Lin, and Y. Yuan, "Learning from synthetic data for crowd counting in the wild," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 8190–8199.

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

[53] Q. Song, C. Wang, Y. Wang, Y. Tai, C. Wang, J. Li, J. Wu, and J. Ma, "To choose or to fuse? scale selection for crowd counting," in *AAAI*, 2021, pp. 2576–2583.

[54] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang *et al.*, "Deep high-resolution representation learning for visual recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 10, pp. 3349–3364, 2020.

[55] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *International Conference on Learning Representations (ICLR)*, 2015, pp. 1–14.