Scale-Context Perceptive Network for Crowd Counting and Localization in Smart City System

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

***Abstract*—The task of crowd counting and localization is to predict the count and position of people in a crowd, which is a practical and essential subtask in crowd analysis and smart city systems. However, the inherent problems of scale variation and background disturbance restrain their performance. While recent research focus on studying counting and localization independently, a few works are capable of executing both tasks simultaneously. To this end, we propose a scale-context perceptive network (SCPNet) to jointly tackle the crowd counting and localization tasks in a unified framework. Specifically, a scale perceptive (SP) module with a local–global branch schema is designed to capture multiscale information. Meanwhile, a context perceptive (CP) module, by the channel-spatial self-attention mechanism, is derived to suppress the background disturbance. Furthermore, a novel hierarchical scale loss function that combines the Euclidean loss function and structural similarity loss function is designed to prompt the proposed model to fulfill the counting and localization simultaneously. Extensive experiments on challenging crowd data sets prove the superiority of the proposed SCPNet compared with the state-of-the-art competitors in both objective and subjective evaluations.**

***Index Terms*—Convolutional neural network (CNN), crowd counting, crowd localization, self-attention mechanism, smart city.**

I. Introduction

CROWD analysis is an emerging topic in computer vision, and a crucial task in smart city applications, e.g., video monitoring, urban planning, and public security [1], [2]. It has two essential subtasks, namely, counting and localization, that have drawn signification attention in recent years. The objectives are to infer pedestrian numbers and locations, respectively. The approach for crowd counting is constantly refreshed and increasingly more effective. Meanwhile, crowd localization, evolved from crowd counting, is gradually explored and developed. They can be served for high-level vision tasks, e.g., crowd tracking [3] and 3-D human pose estimation [4].

The early detection-based counting methods can be regarded as a solution to crowd counting and localization, simultaneously, as the bounding box can output the location information of pedestrians, and the crowd counts can be obtained by summing the bounding boxes. However, these methods have some limitations, e.g., the scale variation in the dense scene will significantly weaken the counting performance, and the background disturbance will mislead the model to recognize the heads, which results in counting errors. Regression-based methods have been proposed to address the problem of detection-based methods, which directly learn the mapping from an image to count. And the counting performance has been improved satisfactorily. In spite of this, they cannot output the location and size information of the head, and only relying on manual features is tough to generate high-quality density map [1]. Benefiting from the feature extraction capability of convolutional neural network (CNN), numerous researchers have adopted the density estimation methods [5], [6] to accomplish the counting task. These methods attempt to count the crowd by summing the pixel values in an estimated density map [7], [8]. For crowd localization, the mainstream methods are density map-based approaches [9], [10]. The idea of map-based methods is to regress a density map and find the maxima point as the head point. Therefore, generating an accurate ground-truth density map is essential to precisely determine the head location.

The scale variation and the background disturbance are primary issues that degrade the accuracy of crowd counting and localization. The scale variations are attributable to the irregular placement of the cameras, which leads to different distances between crowds and cameras [11]. The background disturbance (e.g., trees, buildings, and vehicles) is similar to the foreground region (e.g., head) which results in an overestimation of the count results in crowded scenarios.

Some works tackle the issue of scale variation by utilizing multicolumn structure [12], or adopting the dilated convolution layer [7]. However, these network models have huge parameters and massive computational complexity. To address these concerns, the scale-aware perception (SP) module is built in this article to extract multiscale features effectively. Specifically, we employ the local–global branches combining

Manuscript received 26 October 2022; revised 14 February 2023; accepted 11 April 2023. Date of publication 19 April 2023; date of current version 24 October 2023. This work was supported in part by the National Natural Science Foundation of China under Grant 61601266, and in part by the National Natural Science Foundation of Shandong Province under Grant ZR2021QD041 and Grant ZR2020MF127. *(Corresponding authors: Mingliang Gao; Gwanggil Jeon.)*

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

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

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/JIOT.2023.3268226

with a hierarchical pyramid structure to capture multiscale features.

Some methods solve the problems of background disturbance by adopting the attention mechanism [13], [14], while they lose the information of potential details, or they misidentify the background area as the pedestrian area leading to overestimation. To this end, we propose the context perceptive (CP) module with channel self-attention and spatial self-attention to preserve the detailed information and suppress background disturbances. Furthermore, a novel hierarchical scale (HS) loss is proposed to promote the global and local conformance between the estimated map and the ground-truth map. Last, the synergy of three losses, i.e., HS loss, MSE loss, and structural similarity (SSIM) loss, facilitate the proposed model to collaboratively complete counting and localization. The main contributions are summered as follows.

1) An SP module is designed to capture multiscale features, including the detailed scale feature and the global scale feature.
2) A CP module consisting of a channel-spatial self-attention mechanism is built to suppress the background disturbance.
3) An HS loss is proposed and integrated with the Euclidean loss function and SSIM loss function to motivate the proposed model to perform the counting and localization simultaneously.
4) Extensive experiments and ablation studies verify the superiority of the proposed method in the tasks of crowd counting and localization.

The following sections are structured as follows. Section II reviews the related work. Section III illustrates the proposed method in detail. The experiment discussion and conclusion are presented in Sections IV and V, respectively.

II. Related Work

Recently, CNN-based methods have widely developed and have been the mainstream in crowd counting, benefiting from the powerful feature representation ability of CNN [9], [15]. In this section, we revisit two types of tasks related to the proposed scale-CP network (SCPNet), i.e., crowd counting and crowd localization.

A. Crowd Counting

1) Solutions to Scale Variation: The scale variation will decrease the quality of estimated density maps, thus decrease the accuracy of crowd counting [11]. The multicolumn structure [12] and dilated convolution layer [7] are utilized to extract multiscale features. Zhang et al. [12] first built a network in a multicolumn structure, which obtains multiscale information in a simple but effective means. On the basis of the aforementioned work, Sam et al. [16] embedded a switching layer into a multicolumn architecture to guide the network to focus on a large-scale region. Li et al. [7] deployed some dilated convolution layers to enlarge the receptive fields, which are helpful to capture information in detail. Cao et al. [17] stacked several convolution blocks to capture multiscale information. The blocks consist of four parallel convolution layers with different kernel sizes to aggregate features with diverse scales. Zhai et al. [15] extracted multiscale features by a discriminative feature extractor and a hierarchical feature aggregator.

2) Solutions to Background Disturbance: The background disturbance will cause the overestimation of the crowd region. The attention mechanism [18], [19], [20], [21] has been widely implemented in object counting by adjusting parameters to emphasize the foreground and weaken the background. Gao et al. [13] introduced a channel attention unit to strengthen the class-specific response, which is beneficial to suppress the background disturbance. Miao et al. [22] designed an attention map generator to obtain attention maps, in which the foreground has a brighter crowd than the background. Zhai et al. [11] adopted channel-spatial attention model to distinguish the crowd region and background. The proposed attention unit generates the refined weights by executing a 1-D convolution operation. Guo et al. [23] built a multispectral channel attention unit by generalizing channel attention to the frequency domain through discrete cosine transformation (DCT) formulation. Lin et al. [24] proposed a multifaceted attention network that combined global attention and local learnable region attention for head location.

B. Crowd Localization

Compared with the crowd counting task, the precise location of pedestrians in an image is also crucial and challenging. Crowd localization is performed to identify the location of each pedestrian in the crowd scenes, which can supply detailed positional information for many real-world applications. Idress et al. [25] located the head by detecting the local maximal points in the density map. Liu et al. [26] proposed a multibranch architecture in which counting and localization tasks promote each other. Cheng et al. [10] introduced a probability map that can regress a localization map by a peak detection strategy for crowd localization. Sam et al. [27] built a fine-grained CNN network, namely, LSC-CNN, which can locate the head at very high resolution. Furthermore, it can represent the size of the located head in a bounding box. Abousamra et al. [28] designed a topological network, termed as TopoCount, to relieve semantic errors in dense crowds and combined with the persistence loss to improve the performance of crowd localization. Liang et al. [29] adopted a transformer to crowd localization. In their work, a set of predicted points are generated by two head paths. Then, a Kaiser–Meyer–Olkin (KMO)-based matcher is subsequently utilized to match the predicted points and ground-truth points.

III. Framework of the Proposed Method

The architecture of the proposed method is illustrated in Fig. 1. In the process of training, the HRNet [30] is adopted as the front-end network to extract basic features. Afterward, the scale perceptive (SP) module and CP module are built to tackle the problems of scale variation and background noise in a challenge-oriented manner. Finally, two transposed convolution layers are adopted to upsample the feature and generate the estimated map. A novel HS loss function is built and

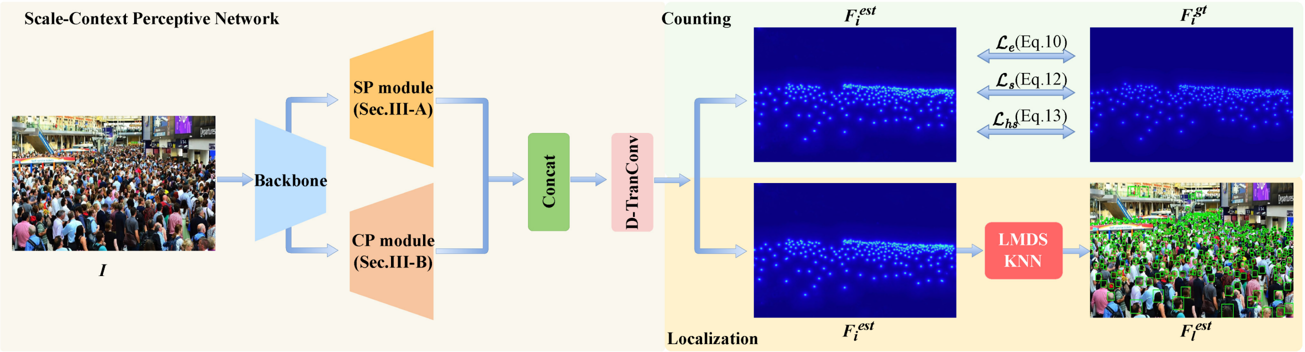


Fig. 1. Architecture of the proposed SCPNet for crowd counting and localization.

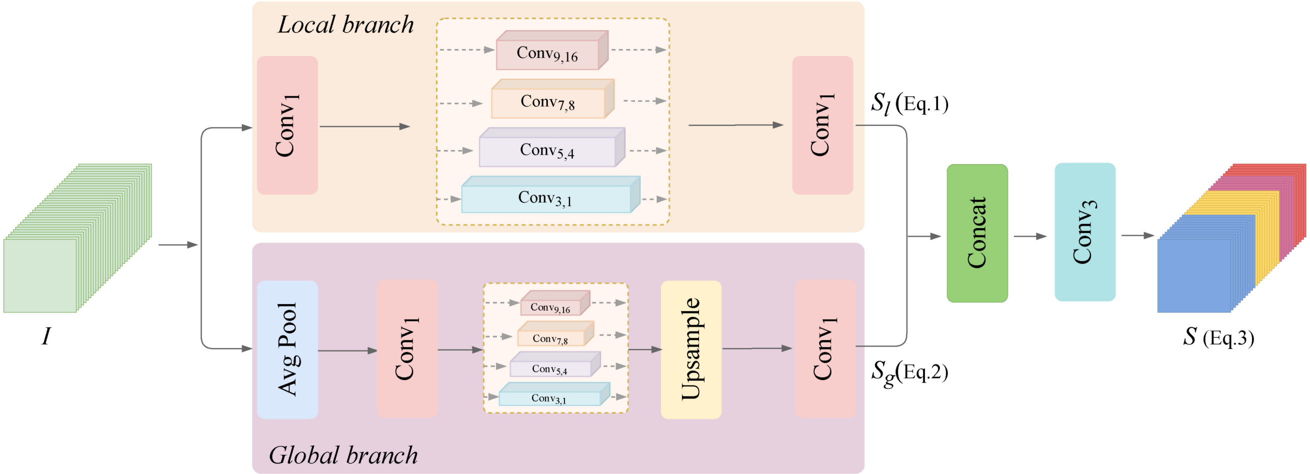


Fig. 2. Framework of the scale SP module.

incorporated with the Euclidean loss function and SSIM loss function to perform the counting and localization in synergy. In the test process, we employ the Local-Maxima-Detection-Strategy [9] and KNN strategy [31] to produce the bounding boxes in the predicted map.

A. Scale Perceptive Module

Scale variation is a pivotal aspect in crowd counting image understanding. To address the challenge, the SP module is proposed to capture multiscale features. The diagram of SP module is depicted in Fig. 2.

The SP module contains two coupled branches, i.e., local branch and global branch. Specifically, the local branch S_l in a hierarchical architecture is responsible to acquire detailed scale information, which is indispensable for the small head regions. First, a 1×1 convolution layer is employed for channel reduction. Afterward, the abundant multiscale features are captured through the hierarchical pyramid structure, which consists of four convolution layers with four kernel sizes, i.e., 9, 7, 5, and 3, respectively. A larger convolutional filter can acquire a wider receptive field with more complexity, while a smaller convolutional filter can capture detailed features. Finally, a 1×1 convolutional filter is used to increase the channels to 512. The local branch S_l is formulated as

$$S_l = \mathrm{Conv}_1\Big\{\mathrm{Conv}_1\Big\langle\mathrm{Cat}\Big[\mathrm{Conv}_{i=3,j=1}^{K,G}(I)\Big]\Big\rangle\Big\}$$
$$K = \{3, 5, 7, 9\}, G = \{1, 4, 8, 16\} \qquad (1)$$

where $I \in \mathcal{R}^{H\times W\times C}$ denotes the input feature map. $\mathrm{Conv}_{i,j}$ represents the convolutional filter of K with group number G. $\mathrm{Cat}(\cdot)$ indicates the concatenation operation.

Unlike the local branch, the global branch aims to capture global features. It resembles the structure of the local branch in a hierarchical pyramid structure to extract the multiscale features. At the front end of the branch, we adopt the adaptive average pooling operation to decrease the input size to 9×9. Subsequently, the hierarchical pyramid structure is utilized to obtain the multiscale features. Finally, the feature map is restored to the same resolution as the input map. The feature of the global branch is formulated as

$$S_g = \mathrm{Up}\Big\{\mathrm{Conv}_1\Big\{\mathrm{Conv}_1\Big\langle\mathrm{Cat}\Big[\mathrm{Conv}_{i=3,j=1}^{K,G}(\mathrm{Avg}(I))\Big\rangle\Big\}\Big\}$$
$$K = \{3, 5, 7, 9\}, G = \{1, 4, 8, 16\} \qquad (2)$$

where $\mathrm{Avg}(\cdot)$ illustrates the adaptive average pooling operation. $\mathrm{Cat}(\cdot)$ represents the concatenation operation and $\mathrm{Up}(\cdot)$ denotes the upsample operation, respectively.

The convolutional filters of 3×3 are utilized to concatenate the feature of local branch and global branch. The final feature map S utilizes upsample operation for restoring to the original image size. It is formulated as

$$S = \mathrm{Up}\big\{\mathrm{Conv}_3\big[\mathrm{Cat}\big(S_l, S_g\big)\big]\big\}. \qquad (3)$$

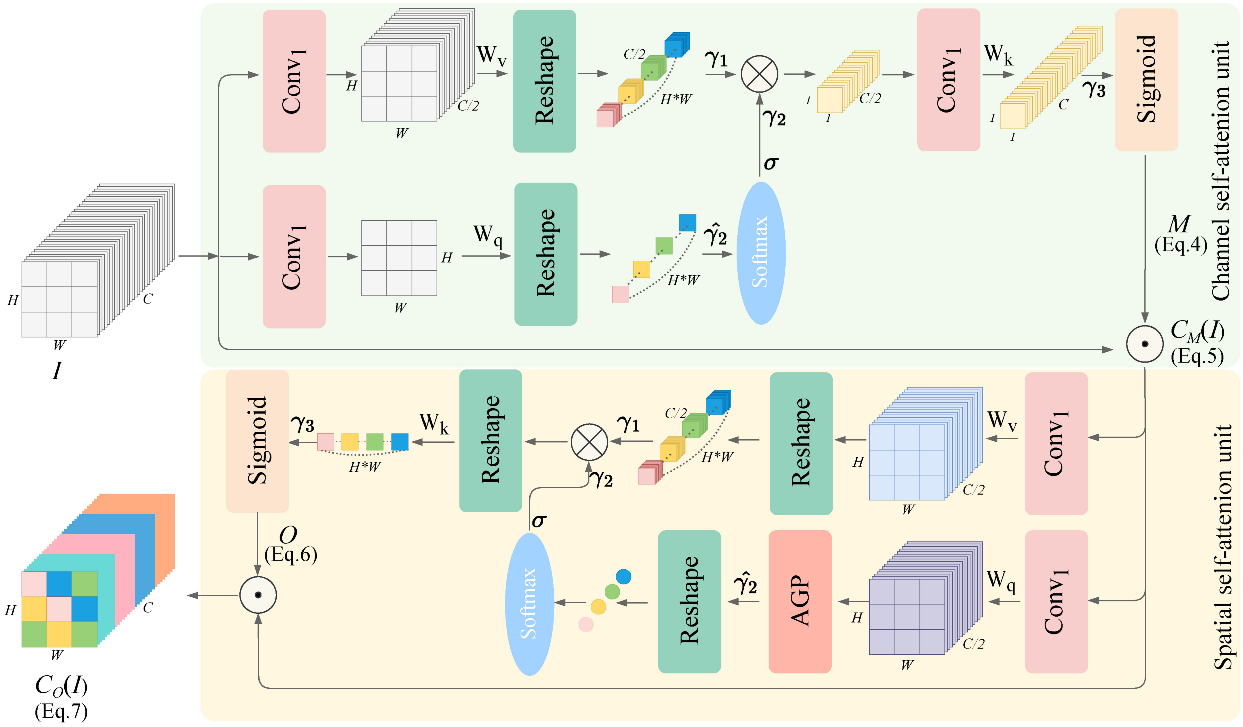


Fig. 3. Framework of the CP module.

B. Context Perceptive Module

Background interference restrains the performance of crowd counting and localization, leading to over- or under-estimation. To solve the problem, the CP module is proposed to identify the head region correctly. Fig. 3 depicts the framework of the CP module.

The CP module consists of two different self-attention units, i.e., a channel self-attention unit and a spatial self-attention unit. The channel self-attention unit is to suppress the background clutter. Given an input $I \in \mathcal{R}^{H\times W\times C}$ generated from the backbone, the channel reduction is first conducted. Specifically, the upper branch channel dimension is reduced to $(C/2)$ and generate $\mathbf{W}_v \in \mathcal{R}^{H\times W\times(C/2)}$. The lower branch is reduced to 1 and generate $\mathbf{W}_q \in \mathcal{R}^{H\times W\times 1}$. Then, $\mathbf{W}_v$ and $\mathbf{W}_q$ are flattened to obtain γ_{c1} and $\hat{\gamma_{c2}}$, respectively. Followed that, the softmax function is employed to reweighting $\hat{\gamma_{c2}}$ and generate γ_{c2}. Next, matrix multiplication is performed between γ_{c1} and γ_{c2}, and layer normalization and the Sigmoid function are used to generate the immediate attention map $m \in \mathcal{R}^{1\times 1\times C}$. The immediate attention map M is defined as

$$M = \varphi\left\{\mathbf{W}_k[\gamma_{c1}\langle \mathbf{W}_v(I)\rangle \otimes \sigma\left[\gamma_{c2}\langle \mathbf{W}_q(I)\rangle\right]]\right\} \tag{4}$$

where φ denotes the Sigmoid function and σ represents the softmax function. The $\otimes$ denotes the matrix dot-product. γ_1 and γ_2 denote the two tensors reshape operators. $\mathbf{W}_k$, $\mathbf{W}_v$, and $\mathbf{W}_q$ represent parameter metrics. The interaction map $C_M(I)(\cdot)$ is obtained by multiplying the input I with the intermediate attention map M, and it is formulated as

$$C_M(I) = M \odot I \tag{5}$$

where $\odot$ denotes the elementwise multiplication.

The spatial self-attention unit is employed to increase the dynamic range of attention, which can enhance the adaptive ability of the model to the crowd region. For the upper branch, it has the same operations as the channel self-attention unit. For the lower branch, the number of channels is first reduced to $C/2$, and then executes an adaptive average operation to decrease the spatial dimension to 1×1 to generate $\hat{\gamma}_{s2}$. Finally, the output attention map $O \in \mathcal{R}^{H\times W\times 1}$ can be obtained through the same operations as the CSA unit. The attention map is formulated as

$$O = \varphi\left\{\gamma_{s1}\left\langle \mathbf{W}_v(M) \otimes \gamma_{s2}\left[\rho\left\langle \mathbf{W}_q(M)\right\rangle\right]\right]\right\rangle\right\} \tag{6}$$

where ρ employs the global pooling operator. The final map C_O interacts with the spatial feature and the channel feature. It is formulated as

$$C_O(M) = O \odot C_M(I). \tag{7}$$

C. Ground-Truth Map

Unlike the traditional Gaussian density map [11], [23], we adopt the focal inverse distance transform (FIDT) map [9] to realize the crowd counting and localization tasks simultaneously instead of Gaussian density map [12] considering that the FIDT can provide the precise location of each head annotation. The FIDT map is generated based on the l_2 transform map as

$$I(x, y) = \min_{(x',y')\in N} \sqrt{(x-x')^2 + (y-y')^2} \tag{8}$$

where $I(x, y)$ reflects the distance between the pixel and corresponding nearest annotation. N denotes a collection of head annotations. Nevertheless, it is hard to regress a density map by (8), as it has a broad distance variation. To overcome this problem, the inverse function is utilized, and a focal function is added to (9) to generate the FIDT map. It is formulated as

$$F(x, y) = \frac{1}{I(x, y)^{(\alpha \times I(x,y) + \beta)} + C} \tag{9}$$

where $F(x, y)$ represents the FIDT map and C is a constant (set to 1) to avoid being divided by 0.

D. Loss Function

1) Euclidean Loss: First, the Euclidean loss is adopted to train the proposed network for crowd counting. The Euclidean loss function measures the estimation difference at pixel level between the estimated map and ground truth. It is formulated as

$$\mathcal{L}_e = \frac{1}{N} \sum_{i=1}^{N} \left\| F_i^{\text{est}} - F_i^{gt} \right\|_2^2 \tag{10}$$

where N denotes the number of images. F_i^{est} and F_i^{gt} represent the estimated maps and ground truth for each of the diverse samples.

2) Structural Similarity Loss: To supplement the blur effect and local structure information, some approaches employ SSIM loss to enhance the quality of the prediction map [17], [32]. It is formulated as

$$\text{SSIM}\left(F_i^{\text{est}}, F_i^{gt}\right) = \frac{\left(2\mu_{F_i^{\text{est}}}\mu_{F_i^{gt}} + \lambda_1\right)\left(2\sigma_{F_i^{\text{est}}F_i^{gt}} + \lambda_2\right)}{\left(\mu^2_{F_i^{\text{est}}} + \mu^2_{F_i^{gt}} + \lambda_1\right)\left(\sigma^2_{F_i^{\text{est}}} + \sigma^2_{F_i^{gt}} + \lambda_2\right)} \tag{11}$$

$$\mathcal{L}_s\left(F_i^{\text{est}}, F_i^{gt}\right) = 1 - \text{SSIM}\left(F_i^{\text{est}}, F_i^{gt}\right) \tag{12}$$

where μ and σ represent the mean and variance of local structure information. λ_1 and λ_2 are set to 1e-4 and 9e-4, respectively.

3) Hierarchical Scale Loss: To promote global and local conformance between the estimated map and the ground-truth map, a novel HS loss is illustrated as follows:

$$\mathcal{L}_{hs} = \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{S} \frac{1}{k_j^2} \left\| P_j\left(F_i^{\text{est}}\right) - P_j\left(F_i^{gt}\right) \right\|_1 + \eta \left\| F_i^{\text{est}} - F_i^{gt} \right\|_1 \tag{13}$$

where S stands for HS level. $P_j(\cdot)$ represents average pooling operation. k_j controls the average pooling output size (i.e., 1×1, 2×2, and 4×4). η is a balance weight.

The estimated map and ground-truth map are processed into three levels with various average pooling output kernel sizes. Depending on the context of the density level, the estimated map is mandated to be concordant with the ground truth at various scales. The global context feature level is derived from the output size 1×1 and the local context feature level of image patches is captured by 2×2 and 4×4. For point localization, we apply the mean absolute error (MAE) loss to improve the robustness of localization points.

The final loss is obtained by summing the three bundle losses, and it is formulated as follows:

$$\mathcal{L} = \mathcal{L}_e + \omega_1 \mathcal{L}_s + \omega_2 \mathcal{L}_{hs} \tag{14}$$

where $\{\omega_1, \omega_2\}$ represents the weighting parameters of the various loss functions and is set to $\{0.01, 0.1\}$.

IV. Experiments

A. Setup of Experiments

To augment the training samples, the training data sets are cropped randomly and flipped horizontally. The crop size is set to 256×256 for the ShanghaiTech data set and 512×512 for other data sets. The Adam optimization [33] is applied to optimize the network, in which the learning rate is initialized as 10^{-4} and the weight decay is set to 5×10^{-4}, respectively. The training batch size is set to 16. The configuration is equipped with Intel Core i7-9700K CPU@3.60 GHz and implemented in PyTorch framework [9] with NVIDIA GeForce GTX 3090Ti GPU.

1) Crowd Counting Evaluation Protocols: For the counting task, the MAE and root mean square error (RMSE) are employed to provide an indication of the accuracy and robustness of the prediction for crowd counting. They are formulated as

$$\text{MAE} = \frac{1}{M} \sum_{m=1}^{M} \left| C_m^{\text{est}} - C_m^{gt} \right| \tag{15}$$

$$\text{RMSE} = \sqrt{\frac{1}{M} \sum_{m=1}^{M} \left(C_m^{est} - C_m^{gt} \right)^2} \tag{16}$$

where M represents the number of images, C_m^{est} represents the estimated count in the mth test image, and C_m^{gt} is the ground truth in the mth test image.

2) Crowd Localization Evaluation Protocols: For the localization task, the Precision, Recall, and F1-measure (F_{1-m}) are measured to reflect the location information of the predicted localization. These three metrics can be formulated as

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \tag{17}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{18}$$

$$F_{1-m} = \frac{2\text{TP}}{2\text{TP} + \text{FN} + \text{FP}} \tag{19}$$

where TP, FP, TN, and FN are the true positive, false positive, true negative, and false negative, respectively.

B. Data Sets

To verify the effectiveness of the proposed SCPNet, comparative experiments are performed on four crowd counting data sets, i.e., ShanghaiTech [12], UCF_CC_50 [34], UCF-QNRF [25], and JHU-Crowd++ [35]. The details of the benchmark data sets are summarized in Table I.

ShanghaiTech [12] consists of two subsets, i.e., Part A and Part B. The former was collected from the Internet, which includes 300 images for training and 182 images for testing. The latter was obtained on the Shanghai shopping streets,

TABLE I
DETAILS OF THE BENCHMARK DATA SETS

Dataset	Num.Images	Avg.Size	Train/Test	Avg.Count	Total.Count
ShanghaiTech Part A [12]	482	589×868	300/ 182	501	241,677
ShanghaiTech Part B [12]	716	768×1024	400/ 316	123	88,488
UCF_CC_50 [34]	50	2101×2888	50/ 50	1,279	63,974
UCF-QNRF [25]	1,535	2013×2902	1,201/ 334	815	1,251,642
JHU-Crowd++ [35]	4,250	1450×900	3,888/ 1,062	262	1,114,785

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-Crowd++	
	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
MCNN [12]	110.2	173.2	26.4	41.3	377.6	509.1	277.0	426.0	188.9	483.4
A-CCNN [36]	85.4	124.6	19.2	31.5	-	-	367.3	-	-	-
PCCNet [37]	73.5	124.0	11.0	19.0	240.0	315.5	148.7	247.3	-	-
CSRNet [7]	68.2	115.0	10.6	16.0	266.1	397.5	-	-	85.9	309.2
MUD-iKNN [31]	68.0	117.7	13.4	21.4	-	-	104.0	172.0	-	-
LSC-CNN [27]	66.4	117.0	8.1	12.7	225.6	302.7	120.5	218.2	112.7	454.4
SANet [17]	67.0	104.5	8.4	13.6	258.4	334.9	-	-	91.1	320.4
SUA-Fully [38]	66.9	125.6	12.3	17.9	-	-	119.2	213.3	-	-
RAZ [26]	65.1	106.7	8.4	14.1	-	-	116.0	195.0	-	-
SFCN [39]	64.8	107.5	7.6	13.0	214.2	318.2	102.0	171.4	77.5	297.6
DUBNet [40]	64.6	106.8	7.7	12.5	243.8	329.3	105.6	180.5	-	-
CG-DRCN [35]	64.0	98.4	8.5	14.4	-	-	112.2	176.3	71.0	278.6
KDMG [41]	63.8	99.2	7.8	12.7	-	-	105.6	180.5	69.7	268.3
HA-CCN [8]	62.9	94.9	8.1	13.4	256.2	348.4	118.1	180.4	-	-
CAN [42]	62.3	100.0	7.8	**12.2**	212.2	301.3	107.0	183.0	100.1	314.0
MBTTBF [43]	60.2	**94.1**	8.0	15.5	233.1	300.9	97.5	165.2	81.8	299.1
Ours	**57.3**	102.1	**7.5**	13.8	**132.0**	**295.0**	**93.7**	**164.3**	**66.2**	**251.0**

including 400 images in the training set and 316 images in the testing set.

UCF_CC_50 [34] comprises 50 annotated images of 63 974 individuals, which were collected from the Internet with various views and resolutions. The number of crowd counts in the images ranges from 94 to 4543 with an average of 1280 people per image. We utilize the fivefold cross-validation principle [34] to evaluate the performance of the proposed method.

UCF-QNRF [25] includes 1201 training images and 334 test images. It comprises scenes with diverse viewpoints, lighting variations, and varying densities.

JHU-Crowd++ [35] contains 4372 high-diversity images, which are divided into 2272 training images, 500 validation images, and 1600 test images. The number of crowd counts in the images ranges from 0 to 25 791.

C. Experimental Results and Analysis

1) Experiments on Crowd Counting: Objective comparison results on crowd counting are reported in Table II. Intuitively, the proposed SCPNet achieves convincing results on all the data sets.

Specifically, on the ShanghaiTech Part A data set, it ranks first with an MAE of 57.3, indicating its superior counting accuracy. Meanwhile, it scores 102.1 in RMSE, which still makes it highly competitive. Especially, compared with the SANet [17] which also utilizes the scale aggregation mechanism, the SCPNet archives 14.4% and 2.3% improvement in MAE and RMSE. On the ShanghaiTech Part B data set, it still ranks first with the lowest MAE, and has a 2.1% reduction in RMSE compared to RAZ [26], which addresses crowd counting and localization tasks simultaneously.

Besides, on the extremely dense UCF_CC_50 data set, the SCPNet surpasses all the competitors, which demonstrates that it can handle congested scenarios. Compared with the second-best CAN [42], the SCPNet improves the MAE and RMSE by 37.8% and 2.1%, respectively.

On the UCF-QNRF data set with large-scale variation, the SCPNet scores 94.9 and 165.2 in MSE and RMSE, both ranking the first place. Compared with the HA-CCN [8] which is especially for addressing the scale variation, it improves the MAE and RMSE by 19.6% and 8.9%, respectively. The results verify the effectiveness of addressing that it is helpful to capture multiscale variation.

On the JHU-Crowd++ data set with serious background noise, the proposed method takes first place in both MAE and RMSE, respectively. Compared with CG-DRCN [35] which introduces a confidence map to alleviate the background clutter, the SCPNet achieves an improvement of 6.8% and 9.9% in MAE and RMSE.

Some subject results on crowd counting are illuminated in Fig. 4. It demonstrates that the estimated map and counts are approximate to the ground truth. The proposed method can effectively capture the scale variation of pedestrians and suppress the background disturbance, benefiting from the SP module and CP module.

2) Experiments on Crowd Localization: The objective comparison results on crowd localization against the competitors are shown in Table III.

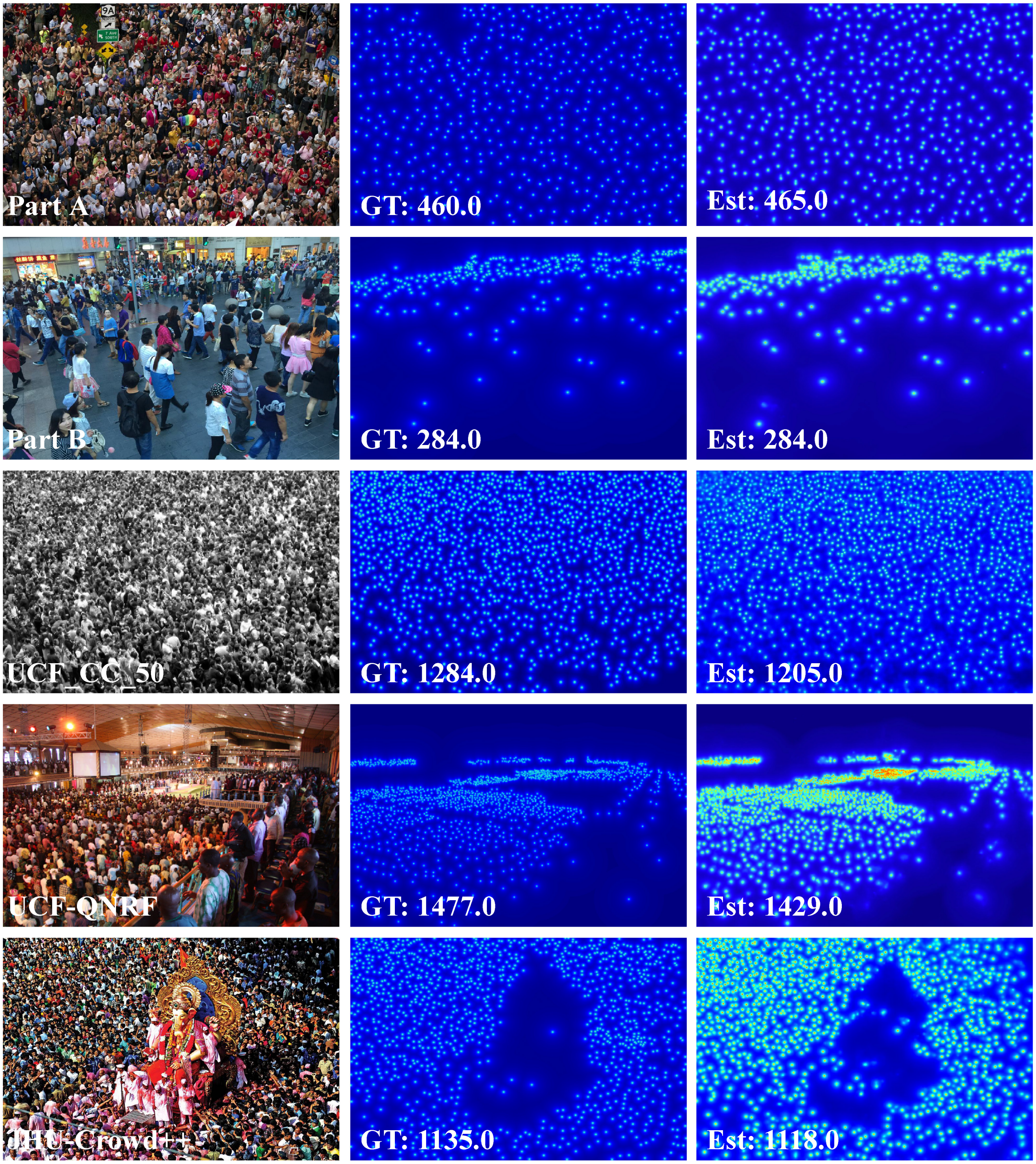


Fig. 4. Subjective results on crowd counting.

Table III shows that the SCPNet ranks first in precision and F-measure, and the suboptimal score in recall on Part A. Compared with the TinyFaces [44] which perform best in Recall, the SCPNet improves the Precision and F1-measure by 82.0% and 35.1%, respectively. For Part B, compared with the second-best method, TopoCount [28], the proposed method achieves the best localization performance and improves the Precision, Recall, and F1-measure by 1.1%, 2.7%, and 2.2%, respectively.

Meanwhile, the proposed method scores first place in Precision, Recall, and F-measure on the UCF-QNRF. Specifically, compared with LSC-CNN [27] which adopts the multicolumn architecture to detect the dense crowds, the proposed SCPNet improves the Precision by 12.4%, Recall by 16.1%, and F1-measure by 15.8%, respectively.

Some visualization results of crowd localization are shown in Fig. 5. It proves that the SCPNet can learn the semantic scale of each head to contain head region information.

TABLE III
OBJECTIVE COMPARISON RESULTS ON CROWD LOCALIZATION. THE BEST RESULTS ARE HIGHLIGHTED IN BOLD

Method	Part A			Part B			UCF-QNRF		
	Precision(%)	Recall(%)	$\mathrm{F}_{1-\mathrm{m}}$(%)	Precision(%)	Recall(%)	$\mathrm{F}_{1-\mathrm{m}}$(%)	Precision(%)	Recall(%)	$\mathrm{F}_{1-\mathrm{m}}$(%)
MCNN [12]	-	-	-	-	-	-	59.9%	63.5%	61.6%
TinyFaces [44]	43.1%	**85.5%**	57.3%	64.7%	79.0%	71.1%	36.3%	77.3%	49.4%
LCFCN [45]	75.1%	45.1%	56.3%	-	-	-	77.9%	52.4%	62.7%
LSC-CNN [27]	63.9%	61.0%	62.4%	71.7%	70.6%	71.2%	76.6%	73.5%	74.0%
TopoCount [28]	74.6%	72.7%	73.6%	82.3%	81.8%	82.0%	81.8%	79.0%	80.3%
Ribera *et al.* [46]	67.7%	44.8%	53.9%	-	-	-	75.5%	49.8%	60.1%
Ours	**78.4%**	76.4%	**77.4%**	**83.8%**	**83.9%**	**83.8%**	**86.1%**	**85.3%**	**85.7%**

TABLE IV
ABLATION STUDIES ON THE CRITICAL MODULES IN THE PROPOSED SCPNET

Methods	MAE	RMSE
baseline	64.5	119.0
baseline+SP	64.2	109.4
baseline+CP	63.8	120.1
baseline+SP_CP	112.4	204.2
baseline+CP_SP	62.3	112.1
baseline+SP||CP	**61.8**	**109.7**

TABLE V
ABLATION STUDIES ON THE LOSS FUNCTIONS IN THE PROPOSED SCPNET

Loss			MAE	RMSE
$\mathcal{L}_e$	$\mathcal{L}_s$	$\mathcal{L}_{hs}$		
√			61.8	109.7
√	√		60.1	105.6
√		√	57.7	108.6
√	√	√	**57.3**	**102.1**

Meanwhile, the proposed method obtains accurate estimations and generates precise localization maps in various crowd scenes.

D. Ablation Studies

To validate the effectiveness of the critical constituents and the proposed loss function, two ablation studies are conducted in Sections IV-D1 and IV-D2.

1) Ablation Studies on Critical Modules: In Table IV, the ablation experiments are conducted to validate the critical constituents and combination types on the ShanghaiTech data set.

The counterparts are illustrated as follows.

1) "baseline" is the model without any constituent parts. The scores of MAE and RMSE are 64.5 and 119.0, respectively.
2) "baseline+SP" refers to the baseline model with the single SP. It improves the MAE and MSE by 0.4% and 8.0% which proves the effectiveness of the SP module to suppress scale variation issues.
3) "baseline+CP" denotes the baseline model with a single CP module, which is beneficial to suppress the background to improve the network robustness.
4) "baseline+SP_CP" represents the cascade model in which the SP module and CP module are connected sequentially. This connection is not conducive to counting.
5) "baseline+CP_SP" is the cascade model in which CP and SP modules are connected sequentially. It shows that this compound mode can facilitate the performance in both MAE and RMSE.
6) "baseline+SP||CP" indicates a parallel model in which the SP module and CP module are parallelly connected. It shows that the performance outperforms the other combination models.

2) Ablation Studies on Loss Functions: To validate the impact of the loss functions, ablation studies are performed on the ShanghaiTech data set. The results are depicted in Table V.

The compounds of different loss functions are illustrated as follows.

1) "$\mathcal{L}_e$" means the only Euclidean loss function. It scores 61.8 and 109.7 in MAE and RMSE, respectively.
2) "$\mathcal{L}_e$+$\mathcal{L}_s$" refers to the Euclidean loss function integrated with the SSIM loss function. One can see that adding $\mathcal{L}_s$" can improve the performance in MAE and RMSE simultaneously.
3) "$\mathcal{L}_e$+$\mathcal{L}_{hs}$" denotes the Euclidean loss function with the HS loss function. It shows that the $\mathcal{L}_{hs}$ loss proves the MAE improves the MAE by 6.8%, and also the RMSE is improved.
4) "$\mathcal{L}_e$+$\mathcal{L}_s$+$\mathcal{L}_{hs}$" represents the final loss function. It achieves the best scores of 57.3 and 102.1 in MAE and RMSE, respectively. It proves that the final loss function enables the network to learn the local and global interaction of crowds at different scales.

V. CONCLUSION

In this article, the SCPNet is proposed to address the problems of scale variation and background disturbance in crowd counting and crowd localization. It mainly consists of two modules, i.e., the SP module and CP module. The SP module is built with a local branch and a global branch to capture multiscale features. The CP module adopts self-attention in both channel and spatial dimensions to suppress the interference of background disturbance. Meanwhile, a novel HS loss is proposed and integrated with the Euclidean loss and SSIM loss to promote the SCPNet to accomplish the tasks of crowd counting and localization synergistically. Extensive ablation studies and experiments prove the validity of the proposed SP module, CP module, and loss function,

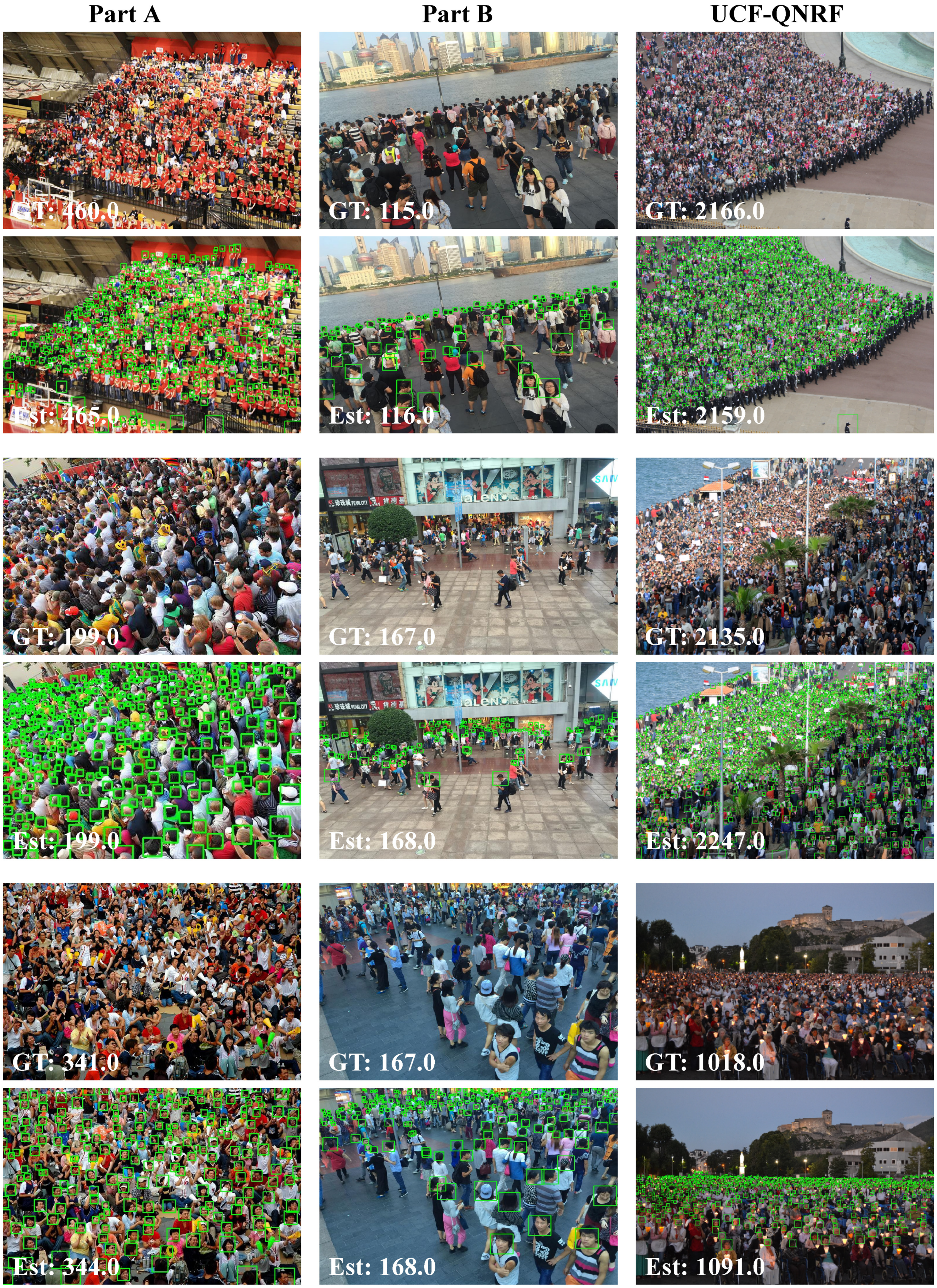


Fig. 5. Subjective results on crowd localization.

and verify the effectiveness of the SCPNet in crowd counting and localization. In the future, more efforts are expected to extend the proposed method to a wide range of video crowd scenes [47], [48]. Meanwhile, drones-based crowd counting [49], [50] is also a direction worthy of further study.

DECLARATIONS

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

REFERENCES

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

[2] Z. Fan, H. Zhang, Z. Zhang, G. Lu, Y. Zhang, and Y. Wang, "A survey of crowd counting and density estimation based on convolutional neural network," *Neurocomputing*, vol. 472, pp. 224–251, Feb. 2022. [Online]. Available: https://doi.org/10.1016/j.neucom.2021.02.103

[3] R. Sundararaman, C. Braga, É. Marchand, and J. Pettré, "Tracking pedestrian heads in dense crowd," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2021, pp. 3864–3874. [Online]. Available: https://doi.org/10.1109/CVPR46437.2021.00386

[4] C. Zheng, S. Zhu, M. Mendieta, T. Yang, C. Chen, and Z. Ding, "3D human pose estimation with spatial and temporal transformers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2021, pp. 11636–11645. [Online]. Available: https://doi.org/10.1109/ICCV48922.2021.01145

[5] A. de Santana Correia and E. Colombini, "Attention, please! A survey of neural attention models in deep learning," 2021, *arXiv:2103.16775*.

[6] Y. Hu et al., "NAS-count: Counting-by-density with neural architecture search," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020, pp. 747–766. [Online]. Available: https://doi.org/10.1007/978-3-030-58542-6_45

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

[8] V. A. Sindagi and V. M. Patel, "HA-CCN: Hierarchical attention-based crowd counting network," *IEEE Trans. Image Process.*, vol. 29, pp. 323–335, 2020. [Online]. Available: https://doi.org/10.1109/TIP.2019.2928634

[9] 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.

[10] J. Cheng, H. Xiong, Z. Cao, and H. Lu, "Decoupled two-stage crowd counting and beyond," *IEEE Trans. Image Process.*, vol. 30, pp. 2862–2875, 2021 [Online]. Available: https://doi.org/10.1109/TIP.2021.3055631

[11] W. Zhai et al., "DA2Net: A dual attention-aware network for robust crowd counting," *Multimedia Syst.*, to be published. [Online]. Available: https://doi.org/10.1007/s00530-021-00877-4

[12] 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)*, 2016, pp. 589–597. [Online]. Available: https://doi.org/10.1109/CVPR.2016.70

[13] 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. [Online]. Available: https://doi.org/10.1016/j.neucom.2019.08.018

[14] N. Liu, Y. Long, C. Zou, Q. Niu, L. Pan, and H. Wu, "ADCrowdNet: An attention-injective deformable convolutional network for crowd understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019, pp. 3220–3229. [Online]. Available: https://doi.org/10.1109/CVPR.2019.00334

[15] W. Zhai et al., "An attentive hierarchy ConvNet for crowd counting in smart city," *Clust. Comput.*, vol. 26, pp. 1099–1111, Sep. 2022. [Online]. Available: https://doi.org/10.1007/s10586-022-03749-2

[16] 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)*, 2017, pp. 4031–4039. [Online]. Available: https://doi.org/10.1109/CVPR.2017.429

[17] 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. [Online]. Available: https://doi.org/10.1007/978-3-030-01228-1_45

[18] M.-H. Guo et al., "Attention mechanisms in computer vision: A survey," 2022, *arXiv:2111.07624*.

[19] 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 Netw. Appl.*, to be published. [Online]. Available: https://doi.org/10.1007/s11036-023-02090-1

[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, no. 24, p. 6363, 2022. [Online]. Available: https://www.mdpi.com/2072-4292/14/24/6363

[21] W. Zhai, M. Gao, M. Anisetti, Q. Li, S. Jeon, and J. Pan, "Group-split attention network for crowd counting," *J. Electron. Imag.*, vol. 31, no. 4, 2022, Art. no. 41214. [Online]. Available: https://doi.org/10.1117/1.JEI.31.4.041214

[22] Y. Miao, Z. Lin, G. Ding, and J. Han, "Shallow feature based dense attention network for crowd counting," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, vol. 34, 2020, pp. 11765–11772. [Online]. Available: https://doi.org/10.1609/AAAI.V34I07.6848

[23] 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, 2022. [Online]. Available: https://doi.org/10.1089/big.2022.0039

[24] H. Lin, Z. Ma, R. Ji, Y. Wang, and X. Hong, "Boosting crowd counting via multifaceted attention," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022, pp. 19628–19637. [Online]. Available: https://doi.org/10.1109/CVPR52688.2022.01901

[25] 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. 532–546. [Online]. Available: https://doi.org/10.1007/978-3-030-01216-8_33

[26] C. Liu, X. Weng, and Y. Mu, "Recurrent attentive zooming for joint crowd counting and precise localization," in *Proc. CVPR*, 2019, pp. 1217–1226. [Online]. Available: https://doi.org/10.1109/CVPR.2019.00131

[27] 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. [Online]. Available: https://doi.org/10.1109/tpami.2020.2974830

[28] S. Abousamra, M. Hoai, D. Samaras, and C. Chen, "Localization in the crowd with topological constraints," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, 2021, pp. 872–881. [Online]. Available: https://doi.org/10.1609/aaai.v35i2.16170

[29] D. Liang, W. Xu, and X. Bai, "An end-to-end transformer model for crowd localization," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2022, pp. 1–17.

[30] 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. [Online]. Available: https://doi.org/10.1109/TPAMI.2020.2983686

[31] 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.*, vol. 5, 2020, pp. 1–9. [Online]. Available: https://doi.org/10.5220/0009156201850195

[32] L. Liu, Z. Qiu, G. Li, S. Liu, W. Ouyang, and L. Lin, "Crowd counting with deep structured scale integration network," in *Proc. Int. Conf. Comput. Vis. (ICCV)*, 2019, pp. 1774–1783. [Online]. Available: https://doi.org/10.1109/ICCV.2019.00186

[33] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2015, pp. 1–15.

[34] 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. (CVPR)*, 2013, pp. 2547–2554. [Online]. Available: https://doi.org/10.1109/CVPR.2013.329

[35] 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, May 2022.

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

[37] 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. [Online]. Available: https://doi.org/10.1109/TCSVT.2019.2919139

[38] Y. Meng et al., "Spatial uncertainty-aware semi-supervised crowd counting," in *Proc. Int. Conf. Comput. Vis. (ICCV)*, 2021, pp. 15549–15559. [Online]. Available: https://doi.org/10.1109/ICCV48922.2021.01526

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

[40] 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. [Online]. Available: https://doi.org/10.1609/AAAI.V34I07.6852

[41] J. Wan, Q. Wang, and A. B. Chan, "Kernel-based density map generation for dense object counting," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 3, pp. 1357–1370, Mar. 2022. [Online]. Available: https://doi.org/10.1109/TPAMI.2020.3022878

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

[43] V. A. Sindagi and V. M. Patel, "Multi-level bottom-top and top-bottom feature fusion for crowd counting," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2019, pp. 1002–1012. [Online]. Available: https://doi.org/10.1109/ICCV.2019.00109

[44] P. Hu and D. Ramanan, "Finding tiny faces," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 951–959. [Online]. Available: https://doi.org/10.1109/CVPR.2017.166

[45] I. H. Laradji, N. Rostamzadeh, P. O. Pinheiro, D. Vazquez, and M. Schmidt, "Where are the blobs: Counting by localization with point supervision," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 547–562. [Online]. Available: https://doi.org/10.1007/978-3-030-01216-8_34

[46] J. Ribera, D. Guera, Y. Chen, and E. J. Delp, "Locating objects without bounding boxes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019, pp. 6479–6489. [Online]. Available: https://doi.org/10.1109/CVPR.2019.00664

[47] Z. Wu, X. Zhang, G. Tian, Y. Wang, and Q. Huang, "Spatial-temporal graph network for video crowd counting," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 1, pp. 228–241, Jan. 2023. [Online]. Available: https://doi.org/10.1109/TCSVT.2022.3187194

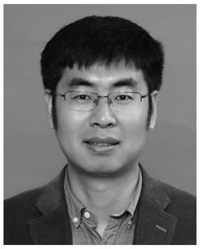
[48] L. Dong, H. Zhang, J. Ma, X. Xu, Y. Yang, and Q. M. J. Wu, "CLRNet: A cross locality relation network for crowd counting in videos," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Oct. 10, 2022, doi: 10.1109/TNNLS.2022.3209918.

[49] A. S. Chakravarthy, S. Sinha, P. Narang, M. Mandal, V. Chamola, and F. R. Yu, "DroneSegNet: Robust aerial semantic segmentation for UAV-based IoT applications," *IEEE Trans. Veh. Technol.*, vol. 71, no. 4, pp. 4277–4286, Apr. 2022. [Online]. Available: https://doi.org/10.1109/TVT.2022.3144358

[50] L. Wen et al., "Detection, tracking, and counting meets drones in crowds: A benchmark," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2021, pp. 7808–7817. [Online]. Available: https://doi.org/10.1109/CVPR46437.2021.00772

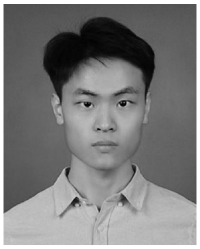
Wenzhe Zhai is currently pursuing the M.S. degree with the School of Electrical and Electronic Engineering, Shandong University of Technology, Zibo, China.

His research interests include smart city system, information fusion, crowd analysis, and deep learning.



Mingliang Gao received the Ph.D. degree in communication and information systems from Sichuan University, Chengdu, China, in 2013.

He is currently an Associate Professor with Shandong University of Technology, Zibo, China. He was a Visiting Lecturer with the University of British Columbia, Vancouver, BC, Canada, from 2018 to 2019. He has been the Principal Investigator for a variety of research funding, including the National Natural Science Foundation, the China Postdoctoral Foundation, and the National Key Research Development Project. He has published over 150 journal/conference papers in IEEE, Springer, Elsevier, and Wiley. His research interests include computer vision, machine learning, and intelligent optimal control.



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.

Qilei Li (Student Member, IEEE) received the M.S. degree from Sichuan University, Chengdu, China, in 2020. He is currently pursuing the Ph.D. degree in computer science with Queen Mary University of London, London, U.K., supervised by Prof. Shaogang (Sean) Gong.

His research interests include computer vision and deep learning, particularly focusing on person ReID, and video/image enhancement.

Mr. Li 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 Postdoctoral 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, École Normale Supérieure Paris-Saclay (ENS-Cachan), Cachan, France. From 2019 to 2020, he was a Prestigious Visiting Professor with Dipartimento di Informatica, Università degli Studi di Milano Statale, Milan, Italy. He is currently a Full Professor with Incheon National University, Incheon, South Korea. He was a Visiting Professor with Sichuan University, Chengdu, China; Universitat Pompeu Fabra, Barcelona, Spain; Xinjiang University, Ürümqi, China; King Mongkut's Institute of Technology Ladkrabang, Bangkok, Thailand; and the University of Burgundy, Dijon, France.

Dr. Jeon 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. He is an Associate Editor of *Sustainable Cities and Society*, IEEE ACCESS, *Real-Time Image Processing*, *Journal of System Architecture*, and *Remote Sensing MDPI*.