IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received December 22, 2021, accepted January 12, 2022, date of publication January 18, 2022, date of current version January 24, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3144075

Congestion-Aware Bayesian Loss for Crowd Counting

JIYEOUP JEONG[1], JONGWON CHOI[2], (Member, IEEE), DAE UNG JO[1], AND JIN YOUNG CHOI[1], (Member, IEEE)

[1]Department of Electrical and Computer Engineering, Automation and Systems Research Institute (ASRI), Seoul National University, Seoul 08826, South Korea
[2]Department of Advanced Imaging, Chung-Ang University, Seoul 06973, South Korea

Corresponding author: Jin Young Choi (jychoi@snu.ac.kr)

This work was supported by the Institute of Information and Communications Technology Planning and Evaluation (IITP) grant funded by Korea government (Ministry of Science and Information and Communications Technology, MSIT) [No. B0101-15-0266, Development of High Performance Visual BigData Discovery Platform for Large-Scale Realtime Data Analysis; No. 2021-0-01343, Artificial Intelligence Graduate School Program (Seoul National University); No. 2021-0-01341, Artificial Intelligence Graduate School Program (Chung-Ang University)].

ABSTRACT Deep learning-based crowd density estimation can greatly improve the accuracy of crowd counting. Though a Bayesian loss method resolves the two problems of the need of a hand-crafted ground truth (GT) density and noisy annotations, counting accurately in high-congested scenes remains a challenging issue. In a crowd scene, people's appearances change according to the scale of each individual (*i.e.*, the person-scale). Also, the lower the sparsity of a local region (*i.e.*, the crowd-sparsity), the more difficult it is to estimate the crowd density. In this paper, we propose a novel congestion-aware Bayesian loss method that considers the person-scale and crowd-sparsity. We estimate the person-scale based on scene geometry, and we then estimate the crowd-sparsity using the estimated person-scale. The estimated person-scale and crowd-sparsity are utilized in the novel congestion-aware Bayesian loss method to improve the supervising representation of the point annotations. We verified the effectiveness of each proposed component through several ablation experiments, and in the various experiments on public datasets, our proposed method achieved state-of-the-art performance.

INDEX TERMS Crowd counting, crowd density estimation, convolution neural network, Bayesian loss.

I. INTRODUCTION

Crowd density estimation can be accomplished with a computer vision-based algorithm to count the number of people in an image, which is one of the challenging tasks for an intelligent surveillance system. Using a crowd density estimation algorithm, we can determine regions of interest where crowds are forming. We can then reduce the computational resources of various algorithms of surveillance system [1]–[3] by concentrating specifically on the detected crowd regions. Furthermore, a crowd density estimation algorithm can also be utilized to count non-human objects, such as cells [4] or vehicles [5].

A crowd density estimation algorithm mainly targets congested scenes, such as the images shown in Fig. 1. In a congested scene, many people are occluded by others.

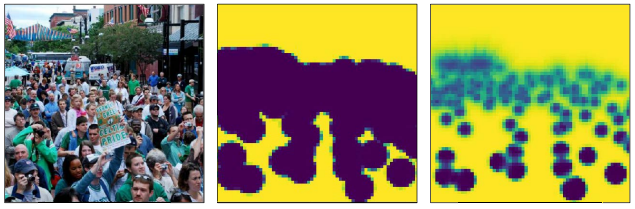
(a) Unevenly distributed crowd (b) Variability in pedestrian scale

FIGURE 1. Examples of target scenes for crowd counting. Crowd counting algorithms mainly target highly congested scenes with (a) an unevenly distributed crowd and (b) a distribution of pedestrians of various scales.

Furthermore, when a crowd is located at a far distance from the camera, each person may only be represented by a few pixels in an image. Due to challenging issues like occlusion and a small occupied region by individuals in a congested scene, it is hard to count the exact number of people in a crowd. Unlike early detection-based methods that counted

The associate editor coordinating the review of this manuscript and approving it for publication was Hiu Yung Wong.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/



(a) Image (b) Bayesian loss method (c) Our method

FIGURE 2. Comparison of background probability map from Bayesian Loss (BL) method and the proposed method. Given a crowd image of (a), the yellow-colored region in (b) and (c) represents the background region indicated by the background probability.

individuals one by one, regression-based density estimation methods can efficiently learn a crowd density map by using only point annotations that mark the location of each person in the image [4], [6], [7].

Regression-based methods have show a large improvement with the advancement of deep learning [8]. Among the deep learning-based methods, the Bayesian loss (BL) method [9] shows impressive performance in training a deep network for crowd density map estimation. Instead of the conventional method that evaluates loss with a desired density value at each pixel, the BL method adopts a novel loss scheme using only the positions of the head point annotations. In contrast to providing the desired density map in conventional methods, the BL method uses a probability that each pixel belongs to a person or background.

In the BL method, the background probability at a pixel is generated by using a fixed distance between the pixel and the nearest head point annotation. However, due to the fixed distance, this background probability model cannot adapt itself to the variation of personal scales and the sparsity of individuals, as shown in Fig. 2. The issue mentioned above results in the limited performance for the various sizes of people from a few pixels to a full face or more, which depends on the scale of person. In addition, the BL method cannot handle the varying degrees of occlusion that arise due to the different sparsity of individuals in a certain region.

In this paper, to solve the issue above, we propose a *congestion-aware Bayesian loss* method in which the estimated scale is used to set up a background probability that is adaptable to personal scale variations. To this end, we have developed schemes to estimate the scale of each person and the sparsity of a local region. These schemes are designed under the assumption that the scale of a person is inversely proportional to the distance the individual and the camera, whereas the sparsity of a region is related to the ratio of the scale and the inter-person distance of the region. Unlike the existing scale-aware schemes [10]–[13], the proposed scale inference method targets the situation where only point annotations are given. Therefore, our method is suitable for single-image crowd density estimation algorithms that provide training images and corresponding point annotations. The estimated sparsity is used to reduce or amplify the loss to adjust for the difficulty in heavily occluded regions. By using the proposed loss, we can learn a diversity of crowd appearances in a weakly supervised manner with only head point annotations instead of density map annotations. Because a diversity of appearances dependent on scale and sparsity are learned in the training phase, estimations of scale and sparsity are not needed at all in the testing phase, and therefore, additional inference costs are not accrued.

Through various experiments, we validate the proposed components including the scale and sparsity estimations of the BL, which contribute to the performance improvement of the proposed method in achieving the state of the art with various benchmark datasets.

Contributions of this paper are summarized as follows:

- We develop schemes to estimate the scale of each person (*i.e.*, person-scale) and the sparsity of a local crowd (*i.e.*, crowd-sparsity) based on the scene geometry.
- Using the estimated person-scale and crowd-sparsity, we propose an extended Bayesian loss method to learn a variety of appearances in a crowd.
- Using the proposed Bayesian loss method, we improve the supervising representation of the point annotations and achieve state-of-the-art performance.

II. RELATED WORKS

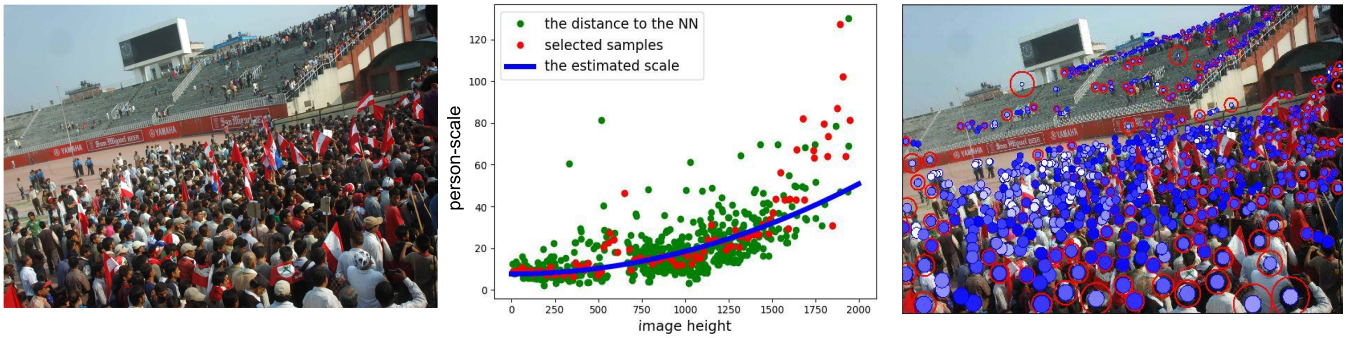
A. DETECTION-BASED AND REGRESSION-BASED CROWD COUNTING

Crowd counting methods can be roughly classified into two groups: detection-based and regression-based methods [3].

Detection-based methods directly identify each target of the count using appropriate pedestrian detectors. When highly congested crowds are formed, however, the appearance of individuals may not be preserved in images, which results in poor algorithmic estimation. To resolve such occlusion issues, some studies use other types of detection targeting, such as faces [14] or the head and shoulders [15]. Despite such efforts, if the surveillance environment is changed, missing or false detection might be caused. Detection-based methods also have unnecessary computational costs that may not require the exact location of people for the sole purpose of counting crowds. Khan *et al.* [10] utilized the results of a conventional head detector to accurately count people by warping the image patch according to the scale of a person. Also, Khan and Basalamah [16] proposed the method using multi-scale fusion module for conventional pedestrian detection [17].

Regression-based methods are more frequently utilized because they perform better than detection-based methods in high-congestion situations or when there is severe occlusion. When regression-based methods were initially proposed, many studies performed mapping of low-level features directly to the size of the crowd [4], [6], [7]. However, these methods that use direct mapping to the count lack spatial information of the crowd so they cannot determine where counting errors occur.

To resolve the limitation of losing spatial information, Lempitsky and Zisserman [18] proposed a method that



(a) Input image (b) Fitting of scale trend (c) Scale estimation result

FIGURE 3. The scale estimation procedure of the proposed method. Given a crowd image (a), the distribution between the nearest neighbor distance and image height of annotations is shown in (b). With the nearest neighbor distances (*green points*), we fit a line (*blue line*) using random sample consensus (RANSAC) by sampling (*red points*) evenly within each section. The circles in (c) express the nearest neighbor distance (*red*) and the estimated scale of the crowd after the fitting process (*blue*). If the estimated scale fits the nearest neighbor distance, it is colored blue; otherwise, it is colored white.

conducts a mapping of local features to an intermediate density map. The density-map regression method has since become the mainstream of crowd counting, enabling the counting of individuals in any region by numerical integration over a density map. Pham *et al.* [19] proposed a non-linear mapping method using the random forest algorithm. Wang and Zou [20] decreased ineffective computational complexity by using subspace learning in the mapping of a density map.

B. CNN-BASED CROWD COUNTING; NETWORK STRUCTURE PERSPECTIVE

A convolutional neural network (CNN) was first applied to estimate the crowd density map by Zhang *et al.* [11]. Motivated by this pioneering work, many studies have been conducted in which a deep network has effectively learned given pairings of an image and its density map. Zhang *et al.* [21] proposed a multi-column network to estimate crowds with varying scales trained in each network column.

There have been several works that use multiple networks to address the multi-scale problem. Sam *et al.* [22] proposed to switch a neural network that classifies image patches according to scale and estimates the crowd density separately. Onoro-Rubio and López-Sastre [5] proposed a hydra-shaped network structure that resizes the image patch to several scales and combines the estimated crowd density. Sindagi and Patel [23] introduced an auxiliary classifier to extract the scale features of an image patch then fused it to a conventional density estimator. Jeong *et al.* [24] constructed multiple network branches according to the sparsity of a local crowds and adopted multi-level refinement network to improve the density estimation accuracy. Hossain *et al.* [13] defined 'scale' as the number of people in a local region and proposed a density estimator using 'scale' as an additional feature. Khan and Basalamah [25] performed small- and large-scaled crowd density estimation successively to improve the accuracy of density estimation.

Several approaches use novel network modules that differ from basic components such as convolution and pooling layers. Li *et al.* [26] used a dilated convolution to effectively extract features of a large field of view. Liu *et al.* [27] used spatial pyramid pooling to adaptively encode the scale as contextual information. Ma *et al.* [12] used human scale quantization at multiple scale levels and trained additional networks to represent scale of a person with a combination of pre-defined scales.

C. CNN-BASED CROWD COUNTING; TRAINING OBJECTIVE PERSPECTIVE

There have been several works that address the limitation of problem setting in crowd density estimation. They tackle (1) the objective function for training and (2) the generation process of the ground truth (GT) density map.

First, several studies pointed out the limitation of the L2 loss between the GT density map and the estimated density map that there is a discrepancy that high-quality representation of density map does not lead to accurate counting. Liu *et al.* [28] utilized the fact that the number of people in a sub-region will always greater than that in the region inside, and proposed what they described as a ranking loss. Shen *et al.* [29] proposed a cross-scale consistency pursuit loss method by using the fact that there is a relationship in which the entire density map is the sum of the partial density maps. Cheng *et al.* [30] designed a spatial awareness loss method to generate a loss when the number of people changes, not when the distribution of people is changed.

In contrast, some studies have tackled the limitations of the definition of the GT density map. Zhao *et al.* [31] utilized auxiliary tasks, such as the estimation of depth, along with the density map to improve the performance. Wan and Chan [32] proposed an adaptive density map generation process that generates learnable density map representations to create sub-optimal density maps. Some works employed segmentation maps [33], the number of people as trainable sources [34], or pedestrian detection results [35].

In particular, Ma *et al.* [9] successfully resolved issues for the training objective and the need for the generation process of a GT density map. They proposed a novel loss (*i.e.*, the Bayesian loss) using the probability of indicating each pixel is included in each point annotation or background.

III. PROPOSED METHOD

In this section, we present the estimation procedures of the scale of a person (*i.e.,* person-scale) and the sparsity of a local region (*i.e.,* crowd-sparsity) and then describe the proposed loss using the estimated person-scale and crowd-sparsity. First, the person-scale estimation procedure is described in Sec. III-A. The method for crowd-sparsity estimation is then described in Sec. III-B. With the estimated person-scale and crowd-sparsity, the proposed loss is described in Sec. III-C.

A. PERSON-SCALE ESTIMATION

To estimate a person-scale in an image, we use the following two scene characteristics. First, the person-scale is represented as inversely proportional to the distance from the person to the camera. We assume a typical surveillance situation where only one ground plane exists, such as a scene without additional layers. In that situation, every person at the same image height is assumed to have the same scale. Also, the person-scale is proportional to the image height that is generally defined in ascending order from the top to the bottom of the image. That is, as shown in Fig. 3(a), people in the bottom region of an image are represented in a large scale, and *vice versa*. Second, in a congested scene, where people are distributed evenly, the person-scale is represented by the nearest neighbor distance of each person..

Under the assumptions described above, we can estimate the person-scale $s(h)$ at the image height h using the inter-person distance as

$$s(h) \approx \frac{1}{P}\sum_{i=1}^{P} |p_i - p_{\mathcal{N}(i)}|, \tag{1}$$

where P is the number of head points at the image height h, p_i is the head position of the i-th person, and $p_{\mathcal{N}(i)}$ is the head position of the nearest neighbor of the i-th person.

However, in some cases, if we directly estimate the person-scale using Eq. (1), the scale estimation results can be noisy because outliers can exist with sparsely distributed people. To resolve the outlier issue, we use a regression of the height-scale relationship, as depicted in Fig. 3. From Fig. 3(b), it can be observed that our assumption on the relation between person-scale and image height is valid. To fit the relationship between person-scale and the image height, we (1) follow the aforementioned scene geometry and (2) consider unevenly distributed crowds as outliers. Hence, we conduct a second-order linear RANSAC (random sample consensus) operation without fitting a constant of the first-order variable, in other words, find a and b in $ax^2 + b$ such that most of the points of x are satisfied. The fitted curve in Fig. 3(b) for estimating the person-scale is obtained by the RANSAC regressor, which models the observed data with little influence of outliers. We utilize the estimated person-scale in designing the congestion-aware Bayesian loss method in Sec. III-C.

B. CROWD-SPARSITY ESTIMATION

We can utilize the estimated crowd-sparsity to improve the learning capability of the crowd density estimation network. When learning the crowd density map, in a densely crowded region, it is difficult to distinguish the crowd from the background clutter. Thus, regions of low crowd-sparsity will significantly affect the overall counting performance. Motivated by the hard-negative mining in object detection algorithms [17], we reduce the influence of loss on annotations in sparsely crowded regions, and amplify the influence of loss on annotations in densely crowded regions.

To estimate crowd-sparsity, we utilize the estimated person-scale in Eq. (1). If a person has a greater distance to his/her nearest neighbor than the estimated scale, we can assume that the crowd in a local region around the person is sparsely distributed, and *vice versa*. The crowd-sparsity around a person is then defined by the ratio of the nearest neighbor distance of the person to the estimated person-scale, in other words,

$$S_n = s(h_n)/s'(h_n), \tag{2}$$

where h_n is the image height of the n-th person, $s(h_n)$ is the distance to his/her nearest neighbor given by $|p_n - p_{\mathcal{N}(n)}|$, and $s'(h_n)$ is the estimated scale for the person. In the region under the fitted curve in Fig. 3(b), the people are highly occluded, so S_n is less than one. In contrast, the people in the region above the curve are sparsely distributed, so S_n becomes larger than one. Hence, using the crowd-sparsity S_n, we can reduce or amplify the influence of the annotations depending on the crowd-sparsity of a local region. In Sec. III-C, we describe the derivation of the proposed loss including the estimated person-scale and crowd-sparsity.

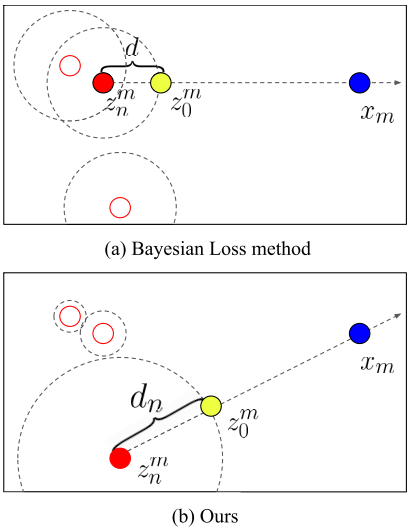
C. CONGESTION-AWARE BAYESIAN LOSS

Let x_m $(m = 1, 2, \ldots, M)$ be a random variable that denotes the spatial location, where M is the number of pixels. Given N number of people, z_n $(n = 1, 2, \ldots, N)$ is a head point annotation. The label for z_n is defined by a random variable y_n. Assuming that the likelihood of the head point annotation follows a Gaussian distribution, the likelihood probability of x_m given label y_n is given by

$$p(x_m|y_n) = \mathcal{N}(z_n, \sigma^2\mathbb{I}_{2\times 2}), \tag{3}$$

where σ is a parameter that controls a region that is affected by each head point annotation and $\mathbb{I}_{2\times 2}$ denotes an identity matrix. In addition, given background label y_0, the likelihood probability of x_m is set to a Gaussian kernel with a centroid z_0^m as

$$p(x_m|y_0) = \mathcal{N}(z_0^m, \sigma^2\mathbb{I}_{2\times 2}). \tag{4}$$



(a) Bayesian Loss method

(b) Ours

FIGURE 4. Dummy background annotation settings. For a pixel x_m of a density map, (a) the Bayesian loss method adopts a dummy background annotation, z_0^m, at a distance d pixels from z_n^m, which is the nearest neighbor head annotation of x_m. In contrast, (b) the proposed method adopts an adaptable distance d_n depending on the person-scale instead of the fixed d.

In this paper, in contrast to the original Bayesian loss, we propose an adjustable centroid z_0^m depending on the person-scale. As shown in Fig. 4, in the original work [9], the centroid z_0^m is located at a distance d pixels from the nearest head annotation. In our method, the adjusted centroid z_0^m is located at a distance d_n pixels from the nearest head annotation, where d_n depends on the person-scale $s'(h_n)$. To this end, we define d_n by

$$d_n = d_0 \cdot s_I \cdot exp(\frac{s'(h_n) - s_0}{s_0}), \tag{5}$$

where s_I is the shorter side length of the image, d_0 (*e.g.*, 0.15) is a fractional scale of s_I, and s_0 is the average person-scale of the dataset. When the person-scale $s'(h_n)$ becomes larger than the average scale s_0, we set d_n to grow exponentially. The adjusted centroid is then obtained by

$$z_0^m = z_n^m + d_n \frac{x_m - z_n^m}{\|x_m - z_n^m\|_2}, \tag{6}$$

where z_n^m denotes the nearest annotation point of x_m.

Using the likelihoods, given the spatial position x_m, the posterior probability of each head point annotation or background is given by

$$p(y_n|x_m) = \frac{p(y_n)p(x_m|y_n)}{\sum_{n'=0}^{N}[p(y_{n'})p(x_m|y_{n'})]}, \tag{7}$$

where $p(y_n) = \frac{1}{N+1}$ denotes the prior probability with label index $n = 0, 1, 2, \ldots, N$, including the background.

If the posterior probability of each head point annotation in Eq. (7) is expressed as a map, it represents the contributed region of each head annotation. Similarly, the posterior probability for the background annotation can represent the background region, as illustrated in Fig. 2(c). It can be seen that the proposed method more accurately represents the background according to the person-scale of annotations than the original work in Fig. 2(b).

If an estimated crowd density at location x_m is denoted as $D^{est}(x_m)$, the Bayesian loss is derived as follows. Let c_n^m be a count at x_m contributed by y_n, and c_n is a count of n-th annotation. Following [9], the expectation of c_n is derived as

$$\begin{aligned} E[c_n] &= E[\sum_{m=1}^{M} c_n^m] = \sum_{m=1}^{M} E[c_n^m] \\ &= \sum_{m=1}^{M} p(y_n|x_m)D^{est}(x_m). \end{aligned} \tag{8}$$

The count value of each annotation c_n should be one and that of background c_0 should be zero. Using the crowd-sparsity for each annotation in Eq. (2), the proposed congestion-aware Bayesian loss (CBL) is proposed by

$$\begin{aligned} \mathcal{L}_{CBL} &= \sum_{n=1}^{N} \frac{1}{S_n} |1 - E[c_n]| + |E[c_0]| \\ &= \sum_{n=1}^{N} \frac{1}{S_n} \left| 1 - \sum_{m=1}^{M} p(y_n|x_m)D^{est}(x_m) \right| \\ &\quad + \sum_{m=1}^{M} p(y_0|x_m)D^{est}(x_m), \end{aligned} \tag{9}$$

where S_n reduces or amplifies the influence of the annotations depending on the crowd-sparsity of a local region. At inference time, we can obtain the number of people without the posterior label probability $p(y_n|x_m)$ as follows:

$$\begin{aligned} C &= \sum_{n=1}^{N} E[c_n] = \sum_{n=1}^{N}\sum_{m=1}^{M} p(y_n|x_m)D^{est}(x_m) \\ &= \sum_{m=1}^{M}\sum_{n=1}^{N} p(y_n|x_m)D^{est}(x_m) \\ &= \sum_{m=1}^{M} D^{est}(x_m), \end{aligned} \tag{10}$$

where C denotes the number of people in the entire image.

IV. EXPERIMENTS

In this section, we describe the evaluation of the effectiveness of the proposed components and illustrate that our method was able to achieve the state-of-the-art on various benchmark datasets.

A. DATASETS

As summarized in Table 2, we have evaluated the proposed method on four challenging crowd counting datasets: UCF_QNRF, ShanghaiTech Part A, ShanghaiTech Part B, and UCF_CC_50.

- **UCF_QNRF [36]** is the latest and largest crowd counting dataset, which includes 1,535 images crawled from

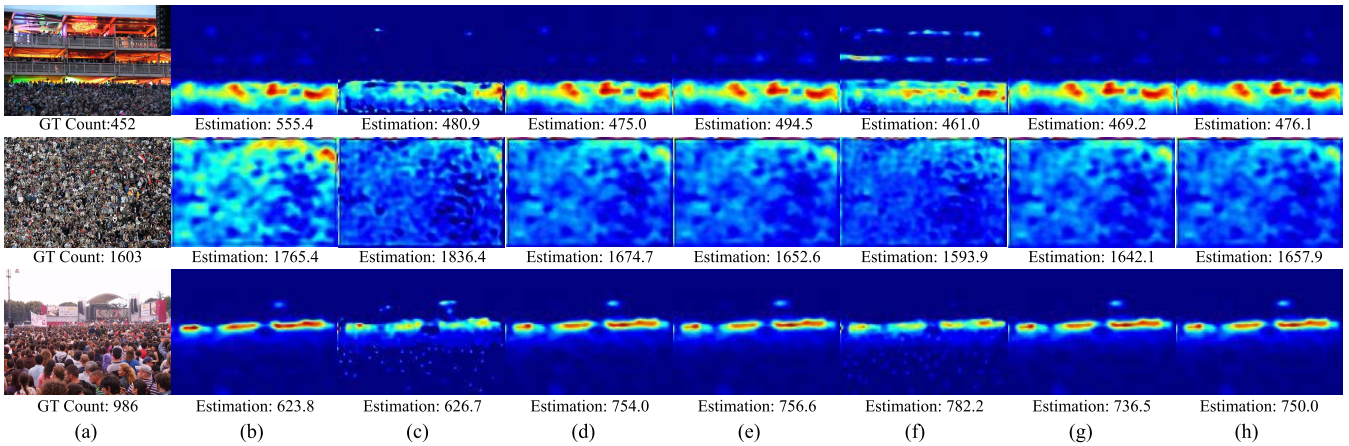


FIGURE 5. Qualitative results of the ablation study: (a) Input image; (b) Estimated density maps from the baseline; (c) Estimated density maps after the scale estimation; (d–h) Estimated density maps after sparsity estimation, varying the fraction of the shorter side of the input image, (d) $d_0 = 1.5$, (e) $d_0 = 2$, (f) $d_0 = 2.25$, (g) $d_0 = 2.5$, and (h) $d_0 = 3$.

TABLE 1. Network structure. The configuration of the convolution layer is expressed as [kernel size]-[number of channels].

Layer	Feature Extraction	Layer	Regression
1-1,2	conv3-64		Bilinear Interpolation
	max pool	1	conv3-256
2-1,2	conv3-128	2	conv3-128
	max pool	3	conv3-1
3-1,2,3,4	conv3-256		
	max pool		
4-1,2,3,4	conv3-512		
	max pool		
5-1,2,3,4	conv3-512		

TABLE 2. Datasets for the experiments.

Dataset	Images	Mean Resolution	Annotations
UCF_QNRF	1,535	2338×1607	1,007,316
ShanghaiTech Part A	482	873×599	162,413
ShanghaiTech Part B	716	1024×768	49,151
UCF_CC_50	50	2101×2888	63,974

Flickr with 1.01 million point annotations. It is a challenging dataset because it has a wide range of counts, image resolutions, light conditions, and viewpoints. The training set has 1,201 images and the remaining 334 images are used for testing.

- **ShanghaiTech [21]** contains 1,198 images with a total of 330,165 people and is divided into two parts: Part A containing 482 images of congested scenes (300 images for training and 182 images for testing), and Part B containing 716 images of sparse scenes (400 images for training and 316 for testing).
- **UCF_CC_50 [36]** contains only 50 gray-scale images which are considered to be challenging due to the high crowd density in the images. Its count value varies from 94 to 4,543. Due to its small quantity, experiments are conducted by 5-fold cross validation followed by the original literature [36].

B. IMPLEMENTATION DETAILS

The proposed network consists of a VGG19 CNN model as described in Table 1. We trained the network in an end-to-end fashion. The first 19 convolutional layers were initialized with a pre-trained VGG19. For the data augmentation processes, we performed random flipping and the cropping of the given images with a size of 512×512 for the UCF_QNRF, ShanghaiTech Part A and UCF_CC_50 datasets and 256×256 for the ShanghaiTech Part B dataset. The parameters were updated by an Adam (adaptive moment estimation) optimizer. All the experiments were performed on an NVIDIA 1080Ti GPU.

C. EVALUATION METRICS

To evaluate our proposed method, we used both the mean absolute error (MAE) and mean squared error (MSE) as evaluation metrics:

$$MAE = \frac{1}{L}\sum_{i=1}^{L}\left|C_i - C_i'\right|, \tag{11}$$

$$MSE = \sqrt{\frac{1}{L}\sum_{i=1}^{L}\left|C_i - C_i'\right|^2}, \tag{12}$$

where L is the number of test images, C_i is the number of people in the i-th image, and C_i' is the estimated number of C_i. The number of people in the image is obtained by the integration of the crowd density over all the image regions as

$$C_i = \sum_{m=1}^{M} D^{GT}(x_m). \tag{13}$$

TABLE 3. Experimental results for the ablation study.

	Base	Scale	Sparsity while varying d_0				
			1.5	2	2.25	2.5	3
MAE	68.7	73.6	68.5	68.3	**61.8**	65.7	66.4
MSE	114.7	115.5	105.3	104.0	**101.7**	103.1	102.8

Similar to the approach used in Eq. (13), the estimated count is obtained as follows:

$$C_i' = \sum_{m=1}^{M} D^{est}(x_m). \tag{14}$$

D. ABLATION STUDY

In this section, we describe the conduct of several experiments to verify the extent to which each proposed component contributed to performance improvement. The ablation experiments were performed on the ShanghaiTech Part A dataset because it could represent well the effectiveness of the proposed method due to its diversity of person-scale and crowd-sparsity within a relatively small quantity of images. According to the configuration of the proposed method, the following three cases were tested:

- **Base** was conducted the same way of Bayesian loss [9]. d_0 in Eq. (6) was set to 0.15.
- **Scale** had the same setting as **Base**, including the proposed person-scale estimation process.
- **Sparsity** trains the network with the proposed loss, including the proposed crowd-sparsity estimation in Eq. (9).

The proposed method has only one hyper-parameter, d_0, which is a guideline for estimating the person-scale. If d_0 varies, the represented scale also varies as the proposed definition. Therefore, we also conducted a comparison experiments varying d_0 after adopting the **Sparsity** setting from the ablation study, which was named **Sparsity-**d_0.

The qualitative results of the ablation study are depicted in Fig. 5, and the quantitative results are summarized in Table 3. Among the testing cases, the best performance was achieved when d_0 was set to 2.25 while considering both the person-scale and crowd-sparsity. The following analysis was derived from the ablation study.

- ***Scale*** *improved the representation of individual's locations but slightly lost counting accuracy when compared with* ***Base***. As shown in Figs. 5(a) and (c), estimated density map of a large-scaled person in the front is represented by a point-shape in **Scale** (c). The point-shaped result means that the density estimation network accurately estimated the location of a person; however, it resulted in a slight loss of some of the counting performance. A strict restriction of point annotation could lead to an inaccurate estimation of the density map around a person. As depicted in Fig. 2(c), performing person-scale estimation concentrates more on the location of the head point annotation than the original work. It could falsely learn the density in more tightly crowded regions containing noisy annotations.
- ***Sparsity-1.5*** *started to improve performance compared to* ***Base***. When d_0 in Eq. (6) was set to be 10 times larger (*i.e.*, $d_0 = 1.5$) than the original work, the performance became similar. In other words, when d_0 was set to 1.5, the training started to consider the diversity of the person-scale without losing the counting performance. As shown in the first row of Figs. 5(c) and (d), false positives were reduced at the top of the estimated map in **Sparsity-1.5** (d), compared to **Scale** (c). Also, **Sparsity-1.5** successfully estimated the density at the bottom of the first row of (d), which was incorrectly estimated as zero in (c) by **Scale**.
- ***Sparsity-2.25*** *showed the best performance.* As depicted in the first row of Fig. 5, a small-scaled individual at the bottom of the image was hard to represent in the density map, except for **Sparsity-2.25**. We can observe the effect of the proposed method through the third row of Fig. 5(f) in which the density in the background region was successfully estimated to be zero; in the other cases, false positives were shown in the background region.
- ***Sparsity*** *settings except for* $d_0 = 2.25$ *had similar density map representations.* We can see that every setting except for **Sparsity-2.25** failed to learn the hard cases mentioned above, such as missing small-scaled people and the falsely estimated densities in the background region. From the results, we confirm that only one parameter setting ($d_0 = 2.25$) improved both counting accuracy and representation capability.

In the remaining experiments, d_0 was set to 2.25 for all the datasets according to the results from the ablation study.

E. COMPARISONS WITH STATE OF THE ART

For the four datasets (UCF_QNRF, ShanghaiTech Part A, ShanghaiTech Part B and UCF_CC_50), we performed extensive comparison experiments with 16 state-of-the-art algorithms, including the early deep-learning models (CCNN [11] and MCNN [21]), models with novel network structures (CMTL [34], SCNN [22], CP-CNN [23], ACSCP [29], DCNet [37], IG-CNN [38], IC-CNN [2], CL-CNN [40], DA-Net [40], ISANet [41], and SDSP [25]), models with network layers specialized in the crowd density estimation (SANet [39], SAAN [13], CSRNet [26] and CAN [27]), a model based on detection scheme [16], and BL method [9]. As summarized in Table 4, the proposed method CBL exhibited the best performance on the MAE metric and also showed a competitive result on the MSE metric. A noticeable improvement was found in the UCF_QNRF, ShanghaiTech Part A, and UCF_CC_50 datasets, in which at least thousands of people were depicted in images. In contrast, it was limited in finding a performance improvement in the ShanghaiTech Part B dataset, which consisted of hundreds of people in relatively simple surveillance environments with few occlusions.

• **UCF_QNRF**: Fig. 6 illustrates the qualitative results for UCF_QNRF dataset. In the first column, false positives in

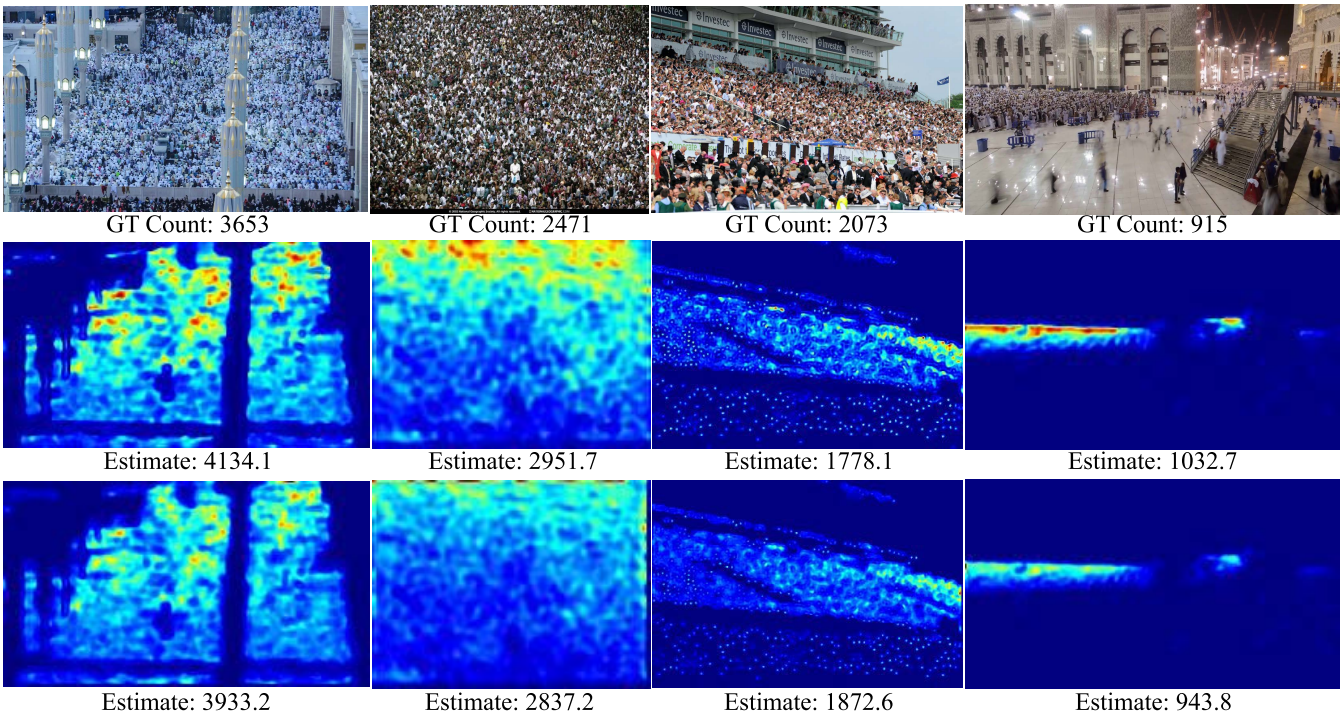


FIGURE 6. Qualitative results with the UCF_QNRF. *Top row*: Input image; *Middle row*: Bayesian loss [9]; *Bottom row*: Our method.

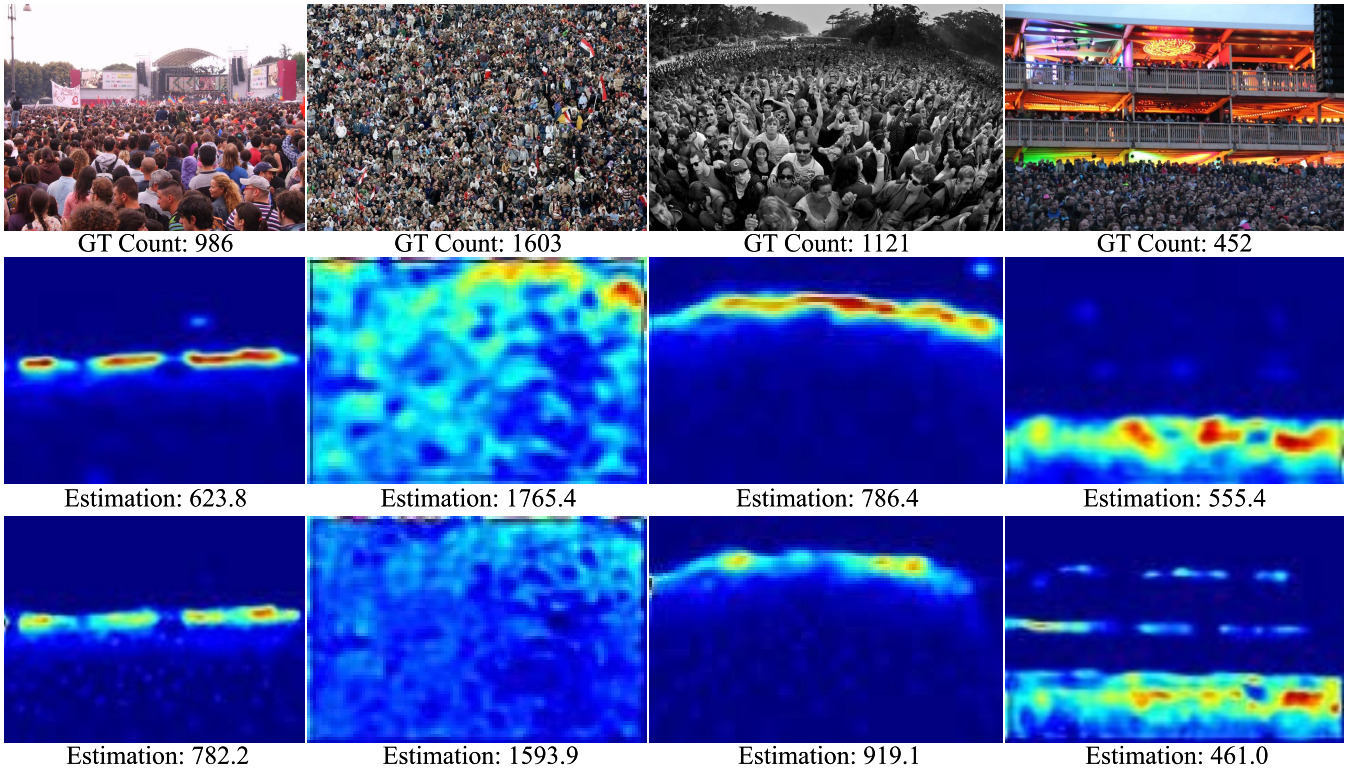


FIGURE 7. Qualitative results with ShanghaiTech Part A. *Top row*: Input image; *Middle row*: Bayesian loss [9]; *Bottom row*: Our method.

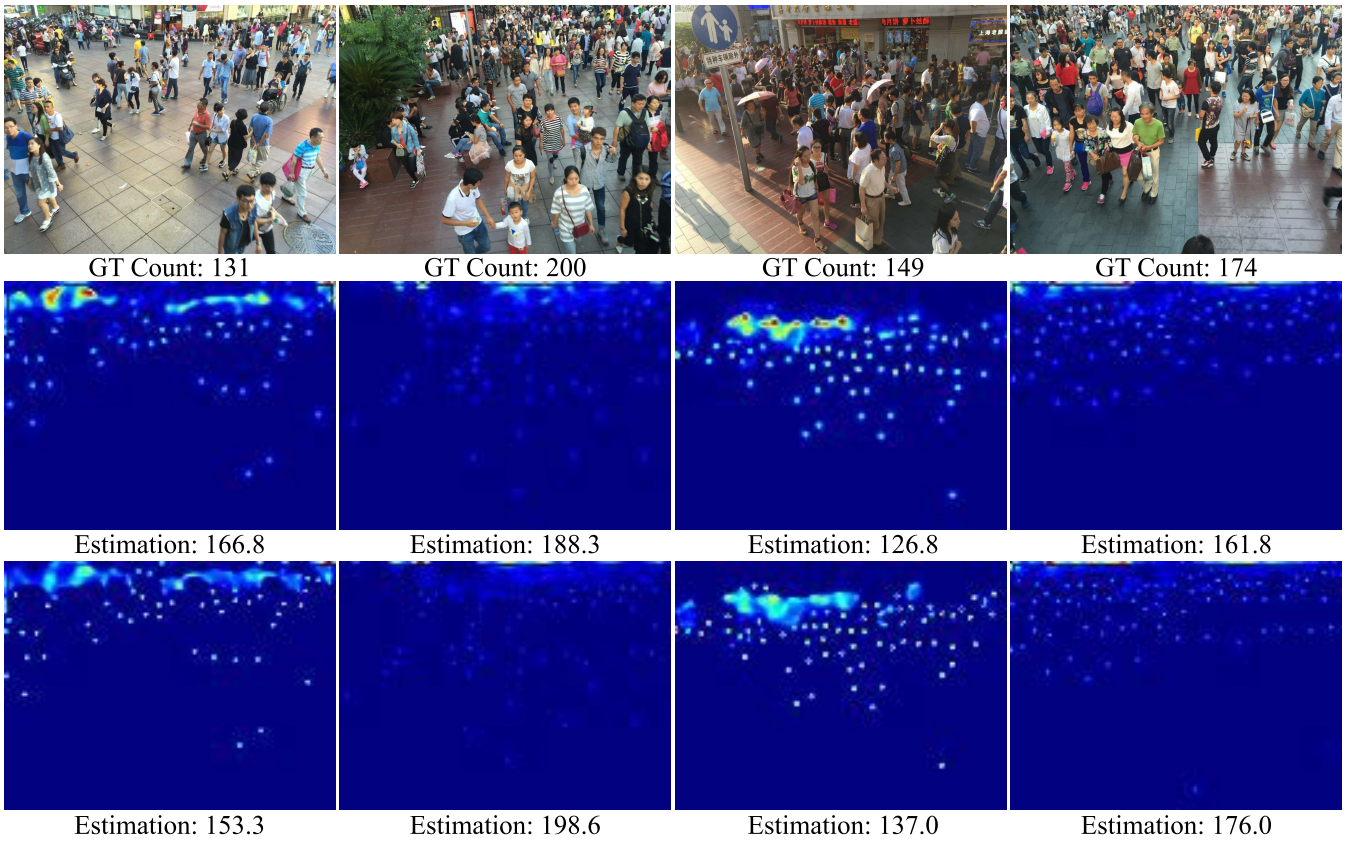


FIGURE 8. Qualitative results with ShanghaiTech Part B. *Top row*: Input image; *Middle row*: Bayesian loss [9]; *Bottom row*: Our method.

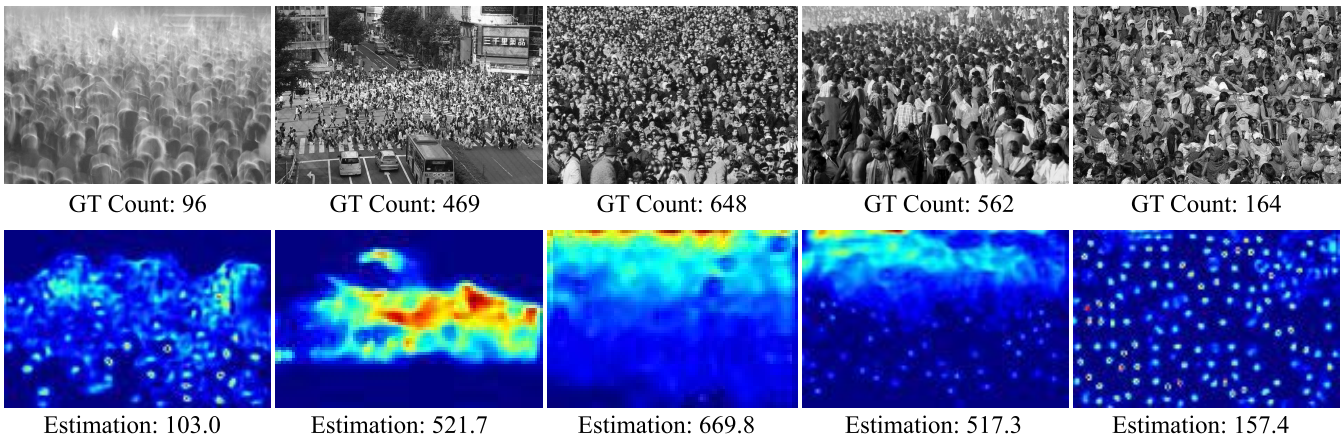


FIGURE 9. Qualitative results with UCF_CC_50. *Top row*: Input image; *Bottom row*: Our method.

the background were removed more in our method compared to the BL method. It was because the foreground and the background were well separated by the proposed person-scale estimation. In the second and the fourth column, the BL method provided an overestimation in a congested region, while such errors were reduced in the proposed method. It is inferred that the propose method provided more accurate learning in the congested region to improve the counting performance. In the third column, however, the localization performance became degraded, resulting in a reduced resolution of crowd density in the grandstand region. It is because people that are densely crowded and severely occluded made the representation in the density map worse.

- **ShanghaiTech Part A**: Fig. 7 depicts the qualitative results for the ShanghaiTech Part A dataset. In the first and third column, the representation of the density map was

TABLE 4. Experimental results for comparison with state-of-the-arts.

Method	UCF_QNRF		SHT Part A		SHT Part B		UCF_CC_50	
	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE
CCNN [11]	-	-	181.8	277.7	32.0	49.8	467.0	498.5
MCNN [21]	277	426	110.2	173.2	26.4	41.3	377.6	509.1
CMTL [34]	252	514	101.3	152.4	20.0	31.1	322.8	341.4
SCNN [22]	228	445	90.4	135.0	21.6	33.4	318.1	439.2
CP-CNN [23]	-	-	73.6	106.4	20.1	30.1	295.8	320.9
ACSCP [29]	-	-	75.7	102.7	17.2	27.4	291.0	404.6
DCNet [37]	-	-	73.5	112.3	18.7	26.0	288.4	404.7
IG-CNN [2]	-	-	72.5	118.2	13.6	21.1	291.4	349.4
CSRNet [26]	98.2	157.2	68.2	115.0	10.6	16.0	266.1	397.5
IC-CNN [38]	-	-	68.5	116.2	10.7	16.0	260.9	365.5
SANet [39]	-	-	67.0	104.5	8.4	13.6	258.4	334.9
SAAN [13]	-	-	-	-	16.9	28.4	271.6	391.0
CL-CNN [40]	132	191	-	-	-	-	-	-
DA-Net [41]	-	-	71.6	104.9	15.0	21.9	290.8	326.5
ISANET [42]	-	-	75.8	124.9	11.0	18.6	-	-
CAN [27]	107.0	183	62.3	**100.0**	7.8	**12.2**	212.2	**243.7**
SDIHD [16]	112	173	-	-	-	-	-	-
SDSP [25]	115.2	175.7	-	-	-	-	229.4	325.6
BL [9]	88.7	**154.8**	62.8	101.8	**7.7**	12.7	229.3	308.2
CBL (Proposed)	**87.0**	155.8	**61.8**	101.7	**7.7**	13.1	**191.7**	283.0

improved in the region where people were sparsely distributed. Density regions that were not counted in the BL method were now expressed in detail, and the underestimated regions were improved. In the second column, which was a highly congested situation, our method more accurately counted the crowd compared to the BL method by improving the crowd's representation and reducing overestimations. In the fourth column, the accuracy was improved from the accurate separation of the foreground and the background. The BL method often failed to count people near the top of the image and on the railing with complex patterns because of the errors made in these regions. Our method accurately recognized not only the people on the railing but also people in the congested region.

• **ShanghaiTech Part B**: Fig. 8 shows the qualitative results from the ShanghaiTech Part B dataset. Because the number of people was smaller than with the other datasets and there were few crowded situation, performance improvement with the proposed method was limited. There was no meaningful difference between the BL method and the proposed method. This was because the proposed method learns various surveillance environments, while this dataset had almost the same scale distribution over the sample images. A slight improvement was achieved in partially crowded regions, such as in the first and the third columns of Fig. 8. Although the qualitative results looked similar, the counting accuracy was improved for the whole case in the sample images in the second and fourth column of Fig. 8.

• **UCF_CC_50**: Fig. 9 shows the qualitative results from the UCF_CC_50 dataset. Since the qualitative results can be slightly different depending on randomly selected samples in the cross-validation setting of the UCF_CC_50 dataset, only the qualitative results of the proposed method are presented. In the second column, the background area is clearly represented by a small density value close to zero. The positions of the people are accurately represented by a point-shape in the last column. In the high-congested scene, such as third column, the estimated density map is blurred, as opposed to the last column, where the estimated density is clearly represented.

F. DIFFERENCES FROM EXISTING PERSON-SCALE INFERENCE

We discuss distinctive aspects of the proposed person-scale inference in contrast to the existing methods as follows.

First, there are methods using the built-in person-scale inference module similar to the proposed method. These methods train the networks to infer the scale of a person along with a learning crowd density. In [12], an additional network module is used for data-driven person-scale inference that requires predefined scale-levels for training scale-level-wise branches in the network. Unlike ours, [12] has a limitation that the scale-level must be defined in advance. Also, the additional network module for person-scale inference requires additional computational overhead. In [13], 'scale' is defined by a value inversely proportional to the number of people in a local image patch. In addition, the 'scale' has to be learned as additional feature. Since the 'scale' in [13] is defined under the assumption that people are evenly distributed in the image patch, even the same scale can be measured differently depending on the sparsity of a local region, which leads to inaccurate scale estimation. In contrast, our person-scale estimation is based on the distance from the person to the camera and so the estimation is robust to the sparsity of a local region. Furthermore, [13] requires additional module for learning person-scale, which increases computational overhead.

Second, there are methods to obtain an accurate person-scale using external information such as head

detection results [10] or scene perspective information [11]. In [10], the scale inference module is based on head detection results. In this detection-based framework, the person-scale can be accurately inferred in the ideal case, but it is difficult to apply the Bayesian loss framework if humans are falsely detected or undetected. Even if detection performs well, scale inference can depend on the performance of detector to affect crowd density estimation performance. Reference [11] targets scenarios where we provide scene information such as region-of-interest and perspective information. However, in the single-image crowd density estimation settings, scene information is usually not accessible in the training phase.

To sum up, the proposed scale inference method can be applied to various crowd environments. The proposed person-scale inference method enables the scale to be inferred even if a small number of point annotations are given. Also, we consider the limitation of single-image crowd density estimation settings that only the position of the annotation is given.

V. CONCLUSION

In this research, we tackled the problem of accurately estimating crowd density in congested scenes for crowd counting. We proposed a novel congestion-aware loss method that considers the scale and sparsity of people. The scale of a person (*i.e.*, person-scale) was estimated from scene geometry. The sparsity of a local region (*i.e.*, crowd-sparsity) was then estimated from the difference between the estimated scale and the nearest neighbor distance. The estimated person-scale and crowd-sparsity was utilized for the proposed congestion-aware loss. We verified the effect of the proposed components through ablation experiments. From the analysis of the ablation study, the person-scale estimation helped to improve the localization accuracy of the crowd density; however, it degraded the counting performance. We found that utilizing the crowd-sparsity improved the counting performance while maintaining the localization accuracy. Based on the results from the ablation study, we conducted comparative experiments between the proposed method and the state-of-the-art methods. It was shown that the proposed method also demonstrated the state-of-the-art performance. The proposed method illustrated that the person-scale and crowd-sparsity were important for crowd density estimation. In addition, if these two properties were dealt with in a unified way, we could show that both the counting performance and the localization accuracy could be improved. In future works, additional performance improvement is expected if a unified method is developed.

REFERENCES

[1] J. Shao, K. Kang, C. C. Loy, and X. Wang, "Deeply learned attributes for crowded scene understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 4657–4666.

[2] D. B. Sam, N. N. Sajjan, R. V. Babu, and M. Srinivasan, "Divide and grow: Capturing huge diversity in crowd images with incrementally growing CNN," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3618–3626.

[3] C. C. Loy, K. Chen, and S. Gong, "Crowd counting and profiling: Methodology and evaluation," in *Modeling, Simulation and Visual Analysis of Crowds*. New York, NY, USA: Springer, 2013, pp. 347–382.

[4] K. Chen, C. C. Loy, S. Gong, and T. Xiang, "Feature mining for localised crowd counting," in *Proc. BMVC*, 2012, p. 3.

[5] D. Onoro-Rubio and R. J. López-Sastre, "Towards perspective-free object counting with deep learning," in *Proc. ECCV*, 2016, pp. 615–629.

[6] A. B. Chan, Z.-S. J. Liang, and N. Vasconcelos, "Privacy preserving crowd monitoring: Counting people without people models or tracking," in *Proc. CVPR*, Jun. 2008, pp. 1–7.

[7] K. Chen, S. Gong, T. Xiang, and C. C. Loy, "Cumulative attribute space for age and crowd density estimation," in *Proc. CVPR*, Jun. 2013, pp. 2467–2474.

[8] V. A. Sindagi and V. M. Patel, "A survey of recent advances in CNN-based single image crowd counting and density estimation," *Pattern Recognit. Lett.*, vol. 107, pp. 3–16, May 2018.

[9] Z. Ma, X. Wei, X. Hong, and Y. Gong, "Bayesian loss for crowd count estimation with point supervision," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6142–6151.

[10] S. D. Khan, H. Ullah, M. Uzair, M. Ullah, R. Ullah, and F. A. Cheikh, "Disam: Density independent and scale aware model for crowd counting and localization," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2019, pp. 4474–4478.

[11] C. Zhang, H. Li, X. Wang, and X. Yang, "Cross-scene crowd counting via deep convolutional neural networks," in *Proc. CVPR*, Jun. 2015, pp. 833–841.

[12] Z. Ma, X. Wei, X. Hong, and Y. Gong, "Learning scales from points: A scale-aware probabilistic model for crowd counting," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 220–228.

[13] M. Hossain, M. Hosseinzadeh, O. Chanda, and Y. Wang, "Crowd counting using scale-aware attention networks," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2019, pp. 1280–1288.

[14] R. Muñoz-Salinas, E. Aguirre, and M. García-Silvente, "People detection and tracking using stereo vision and color," *Image Vis. Comput.*, vol. 25, no. 6, pp. 995–1007, Jun. 2007.

[15] M. Li, Z. Zhang, K. Huang, and T. Tan, "Estimating the number of people in crowded scenes by MID based foreground segmentation and head-shoulder detection," in *Proc. 19th Int. Conf. Pattern Recognit.*, Dec. 2008, pp. 1–4.

[16] S. D. Khan and S. Basalamah, "Scale and density invariant head detection deep model for crowd counting in pedestrian crowds," *Vis. Comput.*, vol. 37, pp. 2127–2137, Aug. 2021.

[17] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 28, 2015, pp. 91–99.

[18] V. Lempitsky and A. Zisserman, "Learning to count objects in images," in *Proc. NIPS*, 2010, pp. 1324–1332.

[19] V.-Q. Pham, T. Kozakaya, O. Yamaguchi, and R. Okada, "COUNT forest: CO-voting uncertain number of targets using random forest for crowd density estimation," in *Proc. ICCV*, Dec. 2015, pp. 3253–3261.

[20] Y. Wang and Y. Zou, "Fast visual object counting via example-based density estimation," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2016, pp. 3653–3657.

[21] Y. Zhang, D. Zhou, S. Chen, S. Gao, and Y. Ma, "Single-image crowd counting via multi-column convolutional neural network," in *Proc. CVPR*, Jun. 2016, pp. 589–597.

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

[23] V. A. Sindagi and V. M. Patel, "Generating high-quality crowd density maps using contextual pyramid CNNs," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1861–1870.

[24] J. Jeong, H. Jeong, J. Lim, J. Choi, S. Yun, and J. Y. Choi, "Selective ensemble network for accurate crowd density estimation," in *Proc. 24th Int. Conf. Pattern Recognit. (ICPR)*, Aug. 2018, pp. 320–325.

[25] S. D. Khan and S. Basalamah, "Sparse to dense scale prediction for crowd couting in high density crowds," *Arabian J. Sci. Eng.*, vol. 46, no. 4, pp. 3051–3065, Apr. 2021.

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

[27] W. Liu, M. Salzmann, and P. Fua, "Context-aware crowd counting," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5099–5108.

[28] X. Liu, J. van de Weijer, and A. D. Bagdanov, "Leveraging unlabeled data for crowd counting by learning to rank," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7661–7669.
[29] Z. Shen, Y. Xu, B. Ni, M. Wang, J. Hu, and X. Yang, "Crowd counting via adversarial cross-scale consistency pursuit," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5245–5254.
[30] Z.-Q. Cheng, J.-X. Li, Q. Dai, X. Wu, and A. Hauptmann, "Learning spatial awareness to improve crowd counting," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6152–6161.
[31] M. Zhao, J. Zhang, C. Zhang, and W. Zhang, "Leveraging heterogeneous auxiliary tasks to assist crowd counting," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 12736–12745.
[32] J. Wan and A. Chan, "Adaptive density map generation for crowd counting," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1130–1139.
[33] 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.
[34] V. A. Sindagi and V. M. Patel, "CNN-based cascaded multi-task learning of high-level prior and density estimation for crowd counting," in *Proc. 14th IEEE Int. Conf. Adv. Video Signal Based Surveill. (AVSS)*, Aug. 2017, pp. 1–6.
[35] J. Liu, C. Gao, D. Meng, and A. G. Hauptmann, "DecideNet: Counting varying density crowds through attention guided detection and density estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5197–5206.
[36] H. Idrees, I. Saleemi, C. Seibert, and M. Shah, "Multi-source multi-scale counting in extremely dense crowd images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 2547–2554.
[37] Z. Shi, L. Zhang, Y. Liu, X. Cao, Y. Ye, M.-M. Cheng, and G. Zheng, "Crowd counting with deep negative correlation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5382–5390.
[38] V. Ranjan, H. Le, and M. Hoai, "Iterative crowd counting," in *Proc. ECCV*, 2018, pp. 270–285.
[39] X. Cao, Z. Wang, Y. Zhao, and F. Su, "Scale aggregation network for accurate and efficient crowd counting," in *Proc. ECCV*, 2018, pp. 734–750.
[40] 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 *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 532–546.
[41] Z. Zou, X. Su, X. Qu, and P. Zhou, "DA-Net: Learning the fine-grained density distribution with deformation aggregation network," *IEEE Access*, vol. 6, pp. 60745–60756, 2018.
[42] J. Sang, W. Wu, H. Luo, H. Xiang, Q. Zhang, H. Hu, and X. Xia, "Improved crowd counting method based on scale-adaptive convolutional neural network," *IEEE Access*, vol. 7, pp. 24411–24419, 2019.

JIYEOUP JEONG received the B.S. degree in electrical engineering from Korea University, South Korea, in 2014. He is currently pursuing the M.S. and Ph.D. degrees in electrical and computer engineering with Seoul National University, Seoul, South Korea. His current research interests include crowd counting, visual tracking, and deep learning.

JONGWON CHOI (Member, IEEE) received the B.S. and M.S. degrees in electrical engineering from KAIST, Daejeon, South Korea, in 2012 and 2014, respectively, and the Ph.D. degree in electrical engineering from Seoul National University, Seoul, South Korea, in 2018. From 2018 to 2020, he was with the Research Intelligence Research Center, Samsung SDS, Seoul, as a Research Engineer. In 2020, he joined the Department of Advanced Imaging, Chung-Ang University, Seoul, where he is currently working as an Assistant Professor. His research interests include the surveillance system with deep learning, the architecture of deep learning, and low-level computer vision algorithms.

DAE UNG JO received the B.S. degree in electrical and computer engineering from Seoul National University, Seoul, South Korea, in 2016, where he is currently pursuing the M.S. and Ph.D. degrees in electrical and computer engineering. His research interests include machine learning, computer vision, and the architecture of deep learning.

JIN YOUNG CHOI (Member, IEEE) received the B.S., M.S., and Ph.D. degrees in control and instrumentation engineering from Seoul National University, Seoul, South Korea, in 1982, 1984, and 1993, respectively. From 1984 to 1989, he was with the Electronics and Telecommunications Research Institute (ETRI), Daejeon, South Korea, where he was involved in the Project of Switching Systems. From 1992 to 1994, he was with the Basic Research Department, ETRI, where he was a Senior Member of Technical Staff involved in the neural information processing system. From 1998 to 1999, he was a Visiting Professor with the University of California at Riverside, Riverside, CA, USA. Since 1994, he has been with Seoul National University, where he is currently a Professor with the School of Electrical Engineering. He is also with the Automation and Systems Research Institute, Engineering Research Center for Advanced Control and Instrumentation; and the Automatic Control Research Center, Seoul National University. His current research interests include adaptive and learning systems, visual surveillance, motion pattern analysis, object detection, object tracking, and pattern recognition.