Background Noise Filtering and Distribution Dividing for Crowd Counting

Hong Mo, *Student Member, IEEE*, Wenqi Ren, *Member, IEEE*, Yuan Xiong, Xiaoqi Pan, *Graduate Student Member, IEEE*, Zhong Zhou, *Member, IEEE*, Xiaochun Cao, *Senior Member, IEEE*, and Wei Wu

***Abstract*— Crowd counting is a challenging problem due to the diverse crowd distribution and background interference. In this paper, we propose a new approach for head size estimation to reduce the impact of different crowd scale and background noise. Different from just using local information of distance between human heads, the global information of the people distribution in the whole image is also under consideration. We obey the order of far- to near-region (small to large) to spread head size, and ensure that the propagation is uninterrupted by inserting dummy head points. The estimated head size is further exploited, such as dividing the crowd into parts of different densities and generating a high-fidelity head mask. On the other hand, we design three different head mask usage mechanisms and the corresponding head masks to analyze where and which mask could lead to better background filtering. Based on the learned masks, two competitive models are proposed which can perform robust crowd estimation against background noise and diverse crowd scale. We evaluate the proposed method on three public crowd counting datasets of ShanghaiTech, UCF_{QNRF} and UCF_{CC_50}. Experimental results demonstrate that the proposed algorithm performs favorably against the state-of-the-art crowd counting approaches.**

***Index Terms*— Crowd counting, head size estimation, density division, head mask.**

I. Introduction

CROWD counting aims to count the number of people appearing in the image. With its significant impact on socio-political and safety perspectives, crowd counting has recently achieved hot attention in computer vision. However, suffering from various crowd scale [5] and background noise [6], crowd counting remains a challenging problem.

The state-of-the-art methods overcome diverse crowd distribution by slicing the image into small patches [4], [5], [7] or parting the image into near- and far-region [8]. Although those methods perform meaningful progress, density map stitched from small pieces may not be equal to the original one due to scaling, as pointed out by Shen *et al.* [9]. On the other hand, the region-based approaches [8] are less effective when the robustness of the decision boundary is low. Additionally, Xu *et al.* [8] indicate that some critical regions in the image are prone to be counted twice since the dividing cline may pass through people.

[10]–[15] are dedicated to deal with the issue of the background noise. Early count-by-detection methods [10]–[13], removing background noise according to depth information or video sequence, are not suitable to static RGB images, which are hard to obtain the background or depth information. Liu *et al.* [14] and Zhu *et al.* [15] introduce an attention map to filter out the background information. Liu *et al.* [14] learn the attention map with searching background images from the Internet. Whereas, such a method is limited to the search scope, which may result in background images under-rich. Zhu *et al.* [15] generate the ground truth of attention map by truncating the density map. However, their attention map is low-fidelity and is not eligible for images with diverse scale crowds.

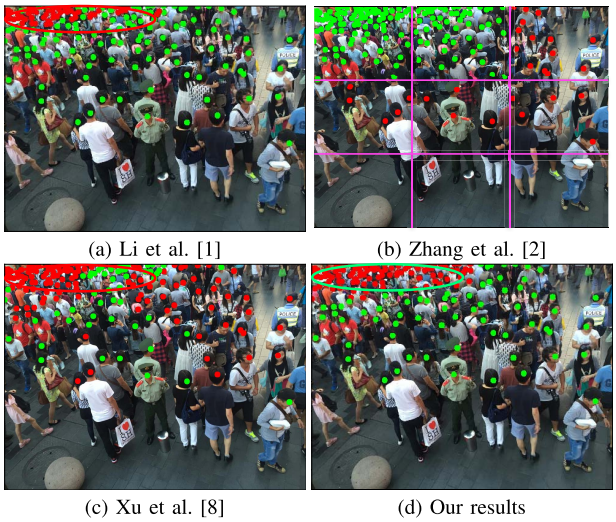
In this paper, we propose a new approach for head size estimation to reduce the impact of both the diverse crowd scale and background noise. We estimate the head size based on the assumption that people at the same distance far from the lens have similar head sizes [1]. Different from just using local information about the distance between human heads [1], the global information of the people distribution in the whole image is also under consideration. We obey the order of far- to near-region (small to large) to spread head size, and ensure that the propagation is uninterrupted by inserting dummy head points. The estimated head size is further exploited, such as dividing the crowd into parts of different densities and generating a high-fidelity head mask. Fig. 1 presents a set of distribution dividing results based on different methods. Red points indicate small heads (dense crowd), and green points indicate large heads (sparse crowd). The picture in the lower right corner of Fig. 1 shows the division result based on our

Manuscript received November 6, 2019; revised April 18, 2020 and June 3, 2020; accepted July 7, 2020. Date of publication August 6, 2020; date of current version August 12, 2020. This work was supported by Natural Science Foundation of China under Grant No. 61872024, National Key R&D Program of China under Grant No. 2018YFB2100603, Beijing Education Committee Cooperation Beijing Natural Science Foundation (No. KZ201910005007), National Natural Science Foundation of China (No. U1803264) and Key Program of the Chinese Academy of Sciences (No. QYZDB-SSW-JSC003). The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Ming-Ming Cheng. *(Corresponding author: Zhong Zhou.)*

Hong Mo, Yuan Xiong, Xiaoqi Pan, Zhong Zhou, and Wei Wu are with the State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, Beijing 100191, China (e-mail: mandymo@buaa.edu.cn; xiongyuanxy@buaa.edu.cn; panxiaoqi@buaa.edu.cn; zz@buaa.edu.cn; wuwei@buaa.edu.cn).

Wenqi Ren and Xiaochun Cao are with the State Key Laboratory of Information Security, Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100195, China (e-mail: renwenqi@iie.ac.cn; caoxiaochun@iie.ac.cn).

Digital Object Identifier 10.1109/TIP.2020.3009030



(a) Li et al. [1] (b) Zhang et al. [2]

(c) Xu et al. [8] (d) Our results

Fig. 1. Distribution dividing results (a) according to head size estimated by head distance [1], (b) of MCNN [2], (c) from depth information [8], and (d) of our method. Red points are crowd in dense, and green points are in sparse.

method, consistent with the law of perspective, where the small head is far from the lens, and the large head is close. Additionally, to grasp where (Image−Feature maps−Density map) and which mask (LowFidelity−MiddleFidelity−HighFidelity) could lead to better background filtering, we provide ablative experiments. Our contributions are summarized as follows:

- We present a simple but effective method to estimate the head size. We design an adaptive threshold to divide the crowd into different groups for each image. Additionally, we propose a high-fidelity head mask according to the estimated head size.
- We provide ablative experiments to grasp late vs. early filtering background noise with different masks. Combined with the learned mask, we propose two competitive models that perform robust crowd estimation against background noise and diverse crowd scale.
- Performances on three public datasets demonstrate the effectiveness of our proposed algorithm.

II. Related Work

There exist three kinds of methods for crowd counting: detection-based methods, regression-based approaches, and the combination of detection and regression.

A. Detection Based Methods

Generally, detection-based methods are a variant of the face and pedestrian detector technology. For example, [16]–[24] quantify the crowd by the number of face/head/head-shoulder. In early, [16]–[18] use traditional feature descriptors such as HOG, edge, and wavelet to represent picture features, and then exploit a classifier (*e.g.,* Bayesian [25] and SVM [26]) to discriminate person or non-person object. These methods are constrained by the accuracy of feature expression and counting speed. On the other hand, with the rapid development of deep learning, the detection technology has improved dramatically. The boosting advance of DNN [27] put object detection technology on track to faster and better performance. Such as Fast-RCNN [28], Faster-RCNN [29], and YOLO [30] have visible improvement on both speed and accuracy. However, detection-based methods are still fragile to the dense crowd since detectors cannot work well when people are too close.

B. Regression Based Approaches

Regression-based methods mainly contain two categories: counts regression and density map regression. Kumagai *et al.* [31] directly predict the count by combining a Gating CNN and Expert CNNs. Since a similar count may have different crowd distribution and different backgrounds, it is not wise to directly predict the exact count. Zhang *et al.* [2] convert the count problem to a density estimation problem by introducing a Gaussian [32] blurred density map. This method of returning the density map has evolved rapidly and into multiple versions. Such as Li *et al.* [1] introduce adaptive kernels to generate Gaussian density maps for dense crowds. Considering a fixed kernel is not suitable for all people, both Yan *et al.* [33] and Wan and Chan [34] use Gaussian with different variances to generate the density map. Moreover, to reduce the influence caused by the position deviation of point prediction, Olmschenk *et al.* [35] propose an inverse k-nearest neighbor map. Besides, [3], [36]–[38] employ multi-task with regressing count and density map at the same time to improve the count accuracy. However, density regression based methods are still affected by background noise and inconsistent crowd scales.

To overcome the effect of background noise, Sam and Babu [6] propose a top-down feedback mechanism to correct spurious detections. They treat the high-level context information as a correcting signal to rectify the wrong density prediction caused by occlusions or background interferences. Although this method obtains favorable results, it is not an end-to-end framework, and its performance is limited as the lack of foreground/background labels. Unlike [6], Sindagi and Patel [39] introduce a self-attention based fusion module to suppress the background noise in the bottom-top path and avoid over-suppression of detailed information due to the top-bottom path. However, the self-attention mechanism is limited to itself in suppressing background noise, that is, if the feature map itself contains a large amount of background noise, the corresponding attention map still contains background noise. Zhu *et al.* [15] propose to cascade feature maps extracted from different convolution layers to vibrant feature expression. They use two branches to regress head-mask and density maps, respectively. Then, they multiply the density map with the predicted head-mask to filter out the background noise. However, their head-mask is not high-fidelity, which may make noise reduction less than expected. Liu *et al.* [40] collect background images from the Internet to distinguish the foreground and background area of the image. Whereas, this algorithm depends heavily on an accurate threshold setting and the sufficiency of the background images. In this paper, we provide a simple way to achieve a high-fidelity head mask

and learn its importance through a set of ablative experiments with the head mask in different degrees of fidelity.

Another line of work tries to solve the problem of perspective change and density inconsistency. To reduce the difference of crowd distribution, Zhang *et al.* [2] slice image into small patches and introduce a multi-column network structure to learn the crowd in different densities. Sam *et al.* [5] make an improvement by adding a switcher to choose the best branch for each patch. However, these branch structures are susceptible to the number of branches [41]. Besides, both Shen *et al.* [9] and Li *et al.* [1] point out that each branch is working in a competitive mechanism, not a cooperative mechanism. Our models work in a "divide and conquer" manner, similar to Zhang *et al.* [42], to narrow the gap of various crowd distribution.

C. Combination of Detection and Regression

Based on an observation that detection-based methods perform better on the sparse crowd while regression-based methods work better on the dense crowd, some works try to combine detection-based and regression-based methods. With a Gaussian convolution, Liu *et al.* [40] convert the detected bounding box into a density map and combine it with the predicted density map. Lian *et al.* [43] improve the head/non-head classification by treating the density map as a probability of a pixel being head. Xu *et al.* [8] partition the image into near- and far-region and use detection and regression to count the number of people in the two regions, respectively. A new deep detection network with only point supervision required is proposed by Liu *et al.* [44], which can simultaneously detect the size and location of human heads and count them in crowds. Considering the complementarity of the detection-based methods and regression-based methods, the combination of the two methods is still worth investigation.

III. Proposed Method

In this section, we introduce our approach from four aspects, the proposed method to estimate head size in both sparse and dense, our division mechanism, ablative experiments of head mask and the frameworks we proposed.

A. Head Size Estimation

We approximate the human head as a circle, the same as Lempitsky and Zisserman [32], and estimate head size with considering both the local head distance and the whole crowd distribution. The proposed head size estimation process can sort out into the following three steps:

1) Radius of Inserted Points: As described in Sec. I, we obtain the head size of a sparse crowd by transmitting the dense head size from far- to near-region and ensure the propagation is uninterrupted by inserting dummy head points into the image. We use the nearest neighbor head spacing to initialize the head size, and use the average of its maximum and minimum value as the head size of the inserted virtual point. However, the estimated head size of the sparse area according to the head distance is much larger than the truth value. Therefore, we set lower and upper limits of the head radius to ensure a reasonable average head size. It is worth investigation that an appropriate largest head radius is vital. If the average size is too large, the inserted points may be insufficient and may cause the propagation to fail. On the contrary, if the average size is too small, there is a possibility that the propagation result does not conform to the principle of perspective. We set the head size range to [5], [50], which is a reasonable range derived from the laws of perspective and still is functional when the data distribution is different from the datasets in this paper. Besides, replacing the maximum value 50 with other values (e.g., 40 and 60) has little effect on the head size estimation. The maximum value within a reasonable range has a particular effect on the speed of propagation. A smaller maximum makes the propagation process converges faster, while larger maximums lead to a more slowly converges. The radius of the dummy points is set as,

$$A_{vg}R = \frac{max(R_{min}, 5) + min(R_{max}, 50)}{2}, \tag{1}$$

where R_{min} and R_{max} mean the minimum and the maximal head size, respectively. The initial head size (R) in this paper is the head distance of the nearest neighbor measured according to the Euler equation.

2) Inserting Points: We consider that the head size is subject to the perspective principle, that is, a linear correlation to the distance far from the lens. Relying on this linear correlation, we learn head size in sparse regions by propagating the head size of dense areas. However, the crowd in the image is not all dense, which may make a jump propagation. Therefore, we propose to insert virtual points to ensure the linear relation. Since the head is considered to be a circle, we can obtain the number of points to be inserted according to the ratio of the image area to the head area, see in Eq. 2.

$$Num = \frac{H \times W}{\pi \times A_{vg}R \times A_{vg}R}, \tag{2}$$

H and W represent the height and width of the image, respectively. Generally, we can insert points into the image with a random inserting mechanism. However, a random insertion mechanism may lead to a failure of the head radius propagation and non-reproducibility of the head size. Therefore, we adopt a fixed insertion method. We treat the ideal population distribution as a dot matrix with $A_{vg}R$ spacing and discard the points that appear in the area, centering on the real head marker and $A_{vg}R$ as the radius. The remaining are the points to be inserted.

3) Iteratively Update the Head Size: Before iteration, we use the nearest neighbor's Euclidean distance to re-estimate the head size of the real head after inserting sufficient points. In addition, in each iteration, we first sort the head points in ascending order according to the head radius and Y-axis and then update the head size according to the following formula.

$$R'_i = \sum_{j=1}^{4} R_j \times \beta_j, \tag{3}$$

Fig. 2. The estimation process of the head radius on three datasets.

where β_j means the weight of the head radius of the j-th people closest to P_i, which is a function with an inverse relationship with distance, as shown in Eq. 4 and 5.

$$\beta_j = \frac{\sigma_j}{\sum_{j=1}^{4} \sigma_j} \tag{4}$$

$$\sigma_j = 1/\sqrt{(x_j - x)^2 + (y_j - y)^2} \tag{5}$$

In this paper, we set the number of iteration to 5, which is statistical information obtained through visual observation. We observe that the head size after five iterations can reach a reasonable value, and too many iterations may cause the head size to increase and decrease suddenly.

In Fig. 2, we show a set of the radius correction process. Each row from top to bottom represents a sample from SHA, SHB, UCF$_{QNRF}$, and UCF$_{CC_50}$, respectively. SHA and SHB are two sub-datasets of ShanghaiTech dataset. Each red area in the figure is a circle with the head label point as the center and the head size as the radius. Each column from left to right indicates source image, the head radius before inserting virtual points, the average head radius, the head radius without the iteration update after inserting virtual points, and the head radius after five iterations. From the head radius in the second column, we can conclude that the method of measuring the head size only by the distance to the nearest head is not suitable for the sparse crowd. With the insertion of virtual points, the crowd's distribution tends to be uniform, as shown in the 3rd column, the head size as shown in the 4th column is mostly in line with the real situation. However, because we insert the virtual points fixedly and uniformly, there are still some unreasonable situations in the estimated head radius. For example, the head size of people who are close to each other is significantly different, or the head radius of people far from the camera is larger than near to the camera. The estimated head size is more in line with the perspective law after adding the information on the distribution of people in the whole image, as shown in the 5th column.

TABLE I
OPTIMAL UPPER AND LOWER HEAD SIZE THRESHOLD OF EACH DATASET

	SHA	SHB	UCF$_{QNRF}$	UCF$_{CC_50}$
T_{lower}	15	20	10	10
T_{upper}	20	30	15	15

B. Sparse & Dense Division

Considering that images have different resolutions and crowd distributions, we propose an adaptive head size threshold to divide the crowd. We first count the average occurrence frequency of head sizes in each data set in the range of [0, 50] at intervals of 5 pixels (as shown in Fig. 3) to figure out the upper and lower limits of the threshold (as shown in Table I). Our method can ensure that the threshold not only obeys the crowd distribution on the whole dataset but also has the characteristics of crowd distribution in a single image. T_{lower} and T_{upper} are set to the head size, which divides the crowd by about one to four and one to one, respectively, to balance the

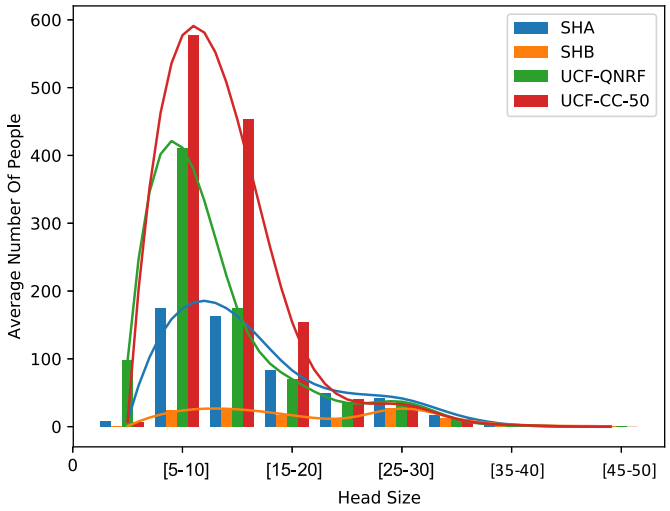


Fig. 3. The average frequency of different head sizes in each picture on the three datasets. SHA and SHB are two sub-datasets of ShanghaiTech.

division area and the people count. For example, [10], [15] can bisect the SHA dataset according to the people count, which causes an unbalanced area division of about 5 : 1. In contrast, our defined threshold range from T_{lower} to T_{upper} can lead to the ratio of people counting about 7 : 20 with an area ratio of 1.7 : 1. The head size threshold (T) for each image can be formulated as follows,

$$T = \begin{cases} \text{HZ} & T_{\text{lower}} < \text{HZ} < T_{\text{upper}}, \\ \dfrac{T_{\text{lower}} + T_{\text{upper}}}{2} & \text{else.} \end{cases} \tag{6}$$

When the average head size (HZ) falls within the division interval, we use this value for division; otherwise, we use the average of the upper and lower limits.

We compare the methods of dividing the crowd based on Li *et al.* [1], MCNN [2], depth information [8], and our method, and show a set of division results in Fig. 1. Both [1] and our method divide the crowd according to head size. Different from [1], we estimate head size with not only considering the local information about the head distance but also employing the global information of the people distribution in the whole image. For comparison, 20 is used as the threshold for both [1] and our method to divide the crowd. According to [2], iimages are sliced into small patches to learn separately. Besides, the crowd in the patch, which is more accurately predicted by the branch with a small kernel size, is considered dense. Considering that the crowd is divided into two groups: sparse and dense, we only employ two branches of MCNN with kernel size in 5×5 and 9×9 to learn the crowd in each patch is dense or sparse. We use the latest monocular image depth estimation method [45] to obtain depth information and then cluster the crowd into two groups by combining the coordinates and depth information. The crowd divide results of these four methods are shown in Fig. 1 $(a) \sim (d)$ in turn. Red points represent the dense crowd, and green points mean the sparse crowd. In theory, according to the perspective rule, the dense crowd (red dots) should be farther from the camera than the sparse crowd (green dots). Overall, among the four division results, only (a) and (d) are more in line with this law, and (b) and (c) are seriously inconsistent. Moreover, (b) rigidly divide the crowd, which leads to a situation where a person is divided into two or even four subgraphs. We believe that the confusion dividing result of (c) may be caused by that the depth information estimated by [45] is not wholly accurate. On the other hand, we consider green points in the large red ellipse of (a) are misjudged for the head sizes are incorrectly estimated. However, our head radius estimation method corrects the head radius of these misjudged points, as shown in the green ellipse of (d). Therefore, our proposed method has better performance in dividing the crowd.

C. Ablative Experiments of Head Mask

Our ablative experiments of head mask mainly consider two variables, the occasion of filtering and the fidelity of the head mask. Considering the two steps of crowd counting, feature representation and density map prediction, we design three kinds of filtering mechanisms: in early (image), in mid-term (feature maps), and in late (density map), as shown in Fig. 4. We use the top ten layers of VGG [46] to extract feature maps and employ a set of dilated convolution with a fixed kernel of 3 and a fixed stride of 2 to generate a density map. The number of channels in each layer is shown in Fig. 4. Besides, a sigmoid activation is following the density map to generate the head mask, and element-wise multiply on Image or Feature Maps or Density Map to implement background filtering. To validate the role of the high-fidelity head mask ($\text{Mask}_{\text{ours}}$), we introduce $\text{Mask}_{0.001}$ and Mask_0 for comparison, which has low fidelity and mid-fidelity, respectively. $\text{Mask}_{0.001}$ and Mask_0 are the image thresholding results of 0.001 and 0, respectively. $\text{Mask}_{\text{ours}}$ is a high-fidelity head mask, which covers the head area as accurately as possible. In this paper, we treat people head as a circle. Through the estimated head size, we set pixels in head area to 1, otherwise 0.

We conduct our ablation experiments on the ShanghaiTech dataset [2], which covers SHA and SHB two parts. SHA has a relatively dense crowd, and SHB has a relatively sparse crowd. Fig. 5 shows the MAE curves on SHA (top row) and SHB (bottom row). We use the green, orange, and blue lines to denote the noise filtering on the input, feature maps, and density maps, respectively. The three columns from left to right are the results of the head mask with low fidelity, medium-fidelity, and high fidelity, respectively. The legend in each sub-figure exhibits the best MAE-MSE value of each model. The blue curve in the 1st column of Fig. 5 has a significant rebound and the orange and green curves in the 2nd∼3rd columns of the 2nd row show significant jitter. Values in the legend of the first column indicate that low fidelity ($\text{Mask}_{0.001}$) works better in mid-term (feature maps) than in early (image). Therefore, we can conclude that using a low-fidelity head mask in feature maps and a high-fidelity mask in the density maps are more stable and reliable.

Besides, we show a sample of post-filtering mechanism framework with different masks in Fig. 6. From the first two rows and the last three columns in Fig. 6, we can see that the mask and density map work collaboratively. Specifically,

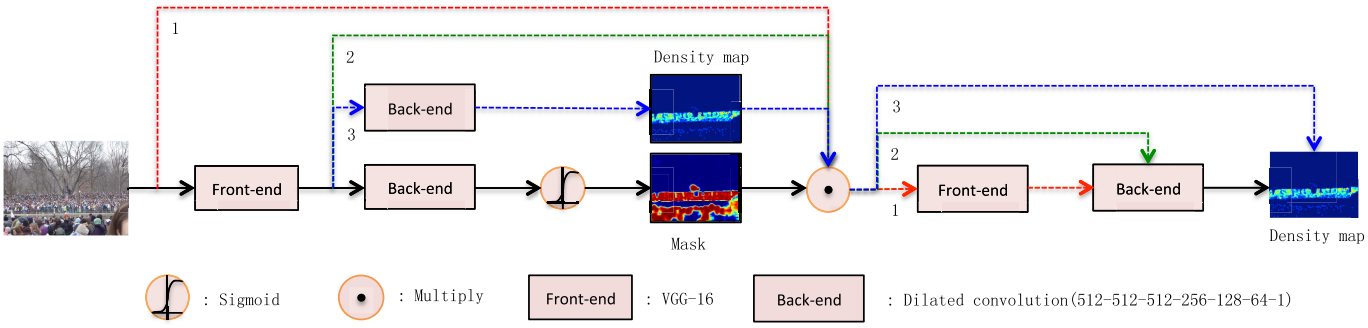
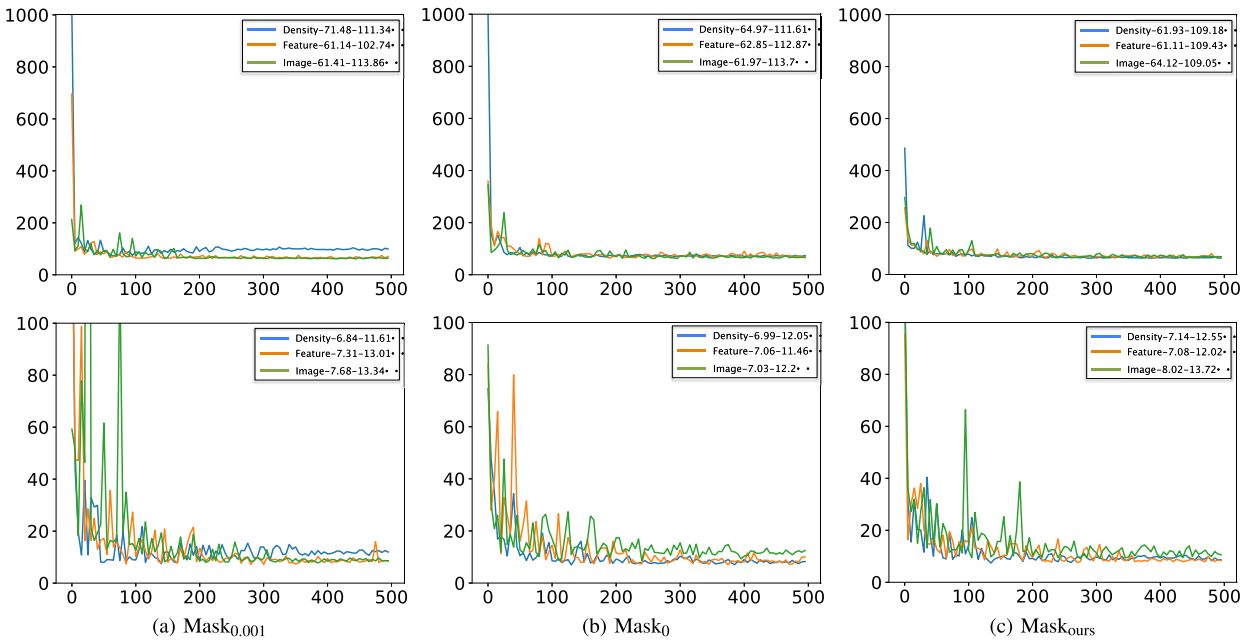


Fig. 4. Frameworks of using mask in three different ways. The network along the red dotted line indicates that background filtering is performed in the early stage, the green dotted line indicates filtered in middle, and the blue dotted line indicates filtered in late. The black arrows connect the collaborative operations. Different colored dotted lines connect the operation for filtering on different occasions. Red dotted line means filtering the background on the image, green means filtering on feature maps, and the blue one means on density map.



(a) $\text{Mask}_{0.001}$ (b) Mask_0 (c) $\text{Mask}_{\text{ours}}$

Fig. 5. MAE evaluation results on SHA (the upper row) and SHB (the lower row). Text in legend means Framework-Mask. Such as "Density-0.001" means using $\text{Mask}_{0.001}$ to filter the background information on density map.

the density map working with the low-fidelity head mask has high ambiguity, while the density map with the high-fidelity head mask has high clarity. If the head mask of the dense area is too large, it is difficult to filter the noise to overestimate the number of people; and for the sparse area, if the head mask is too small, it is easy to lose information, thereby underestimating the number of people. For example, the leftmost block in the middle of the 2nd to 4th columns, in ascending order, the size of the head mask is $\text{Mask}_{0001} < \text{Mask}_{\text{ours}} < \text{Mask}_0$, the number of people without noise filtering is $\text{Mask}_{\text{ours}} < \text{Mask}_0 < \text{Mask}_{0001}$, and the final count is $\text{Mask}_{0001} < \text{Mask}_{\text{ours}} < \text{Mask}_0$. Alos, both $\text{Mask}_{\text{ours}}$ and Mask_{0001} underestimate the number of people, while Mask_0 overestimate. This shows that as a mask, the $\text{Mask}_{\text{ours}}$ is too small, and Mask_0 is too large. Because our head mask is positively related to the size of the head in the distance, the size of the head in near-region without perspective correction will be smaller. As shown in the most densely populated sub-region in the figure, in ascending order, the size of the hood is $\text{Mask}_{\text{ours}} < \text{Mask}_{0001} < \text{Mask}_0$, the sum of density map without noise filtering is $\text{Mask}_0 < \text{Mask}_{0001} < \text{Mask}_{0001}$. While the sum of density map shows anomalies, that is, $\text{Mask}_{0001} < \text{Mask}_{\text{ours}}$. Nonetheless, all the three counts are higher than the ground truth. On the one hand, because we limited the minimum head size to 5 pixels, the head mask of the dense area is too large. On the other hand, the predicted density map is 1/64 of the original image, with the value has been increased by 64 times, increasing the requirement for the accuracy of the head mask. Besides, from the left low-corner and the right low-corner images of Fig. 6, the total difference of left is smaller than the right one, with more significant difference in local. In summary, we conclude that 'High fidelity GT head mask' is essential and the role of the head mask is only limited by its accuracy.

D. The Proposed Frameworks

Based on the learned head mask and the division result, we propose two competitive models that perform robust crowd estimation against background noise and diverse

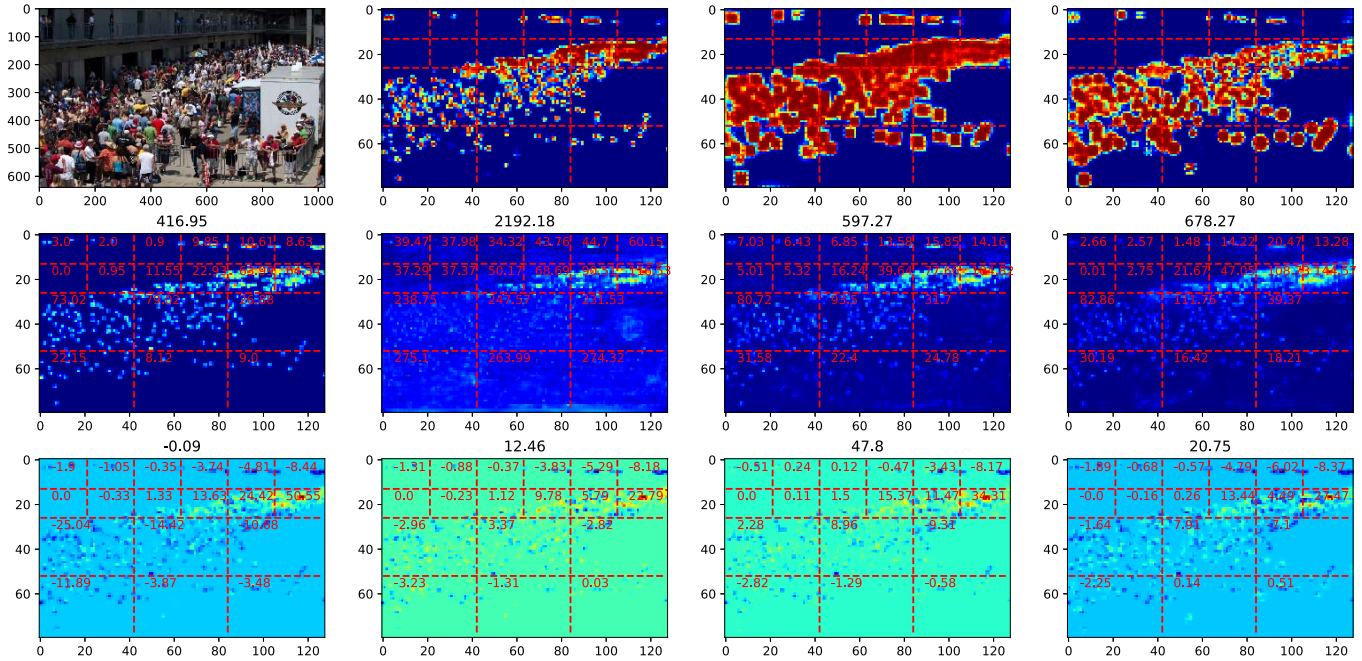


Fig. 6. A sample of post-filtering mechanism frameworks with different masks. From left to right, the 2nd ∼ 4th columns are results of $\text{Mask}_{0.001}$, Mask_{0}, and $\text{Mask}_{\text{ours}}$, respectively. The title on the last two rows is the sum of each sub-image. The 2nd row, 1rst column, is ground truth. The last row is the difference between the predicted density map and ground truth. The lower left corner shows the difference between the density map generated by multiplying the mask in the 1st row, 2nd column with the density map in the 2nd row, 4th column and the ground truth.

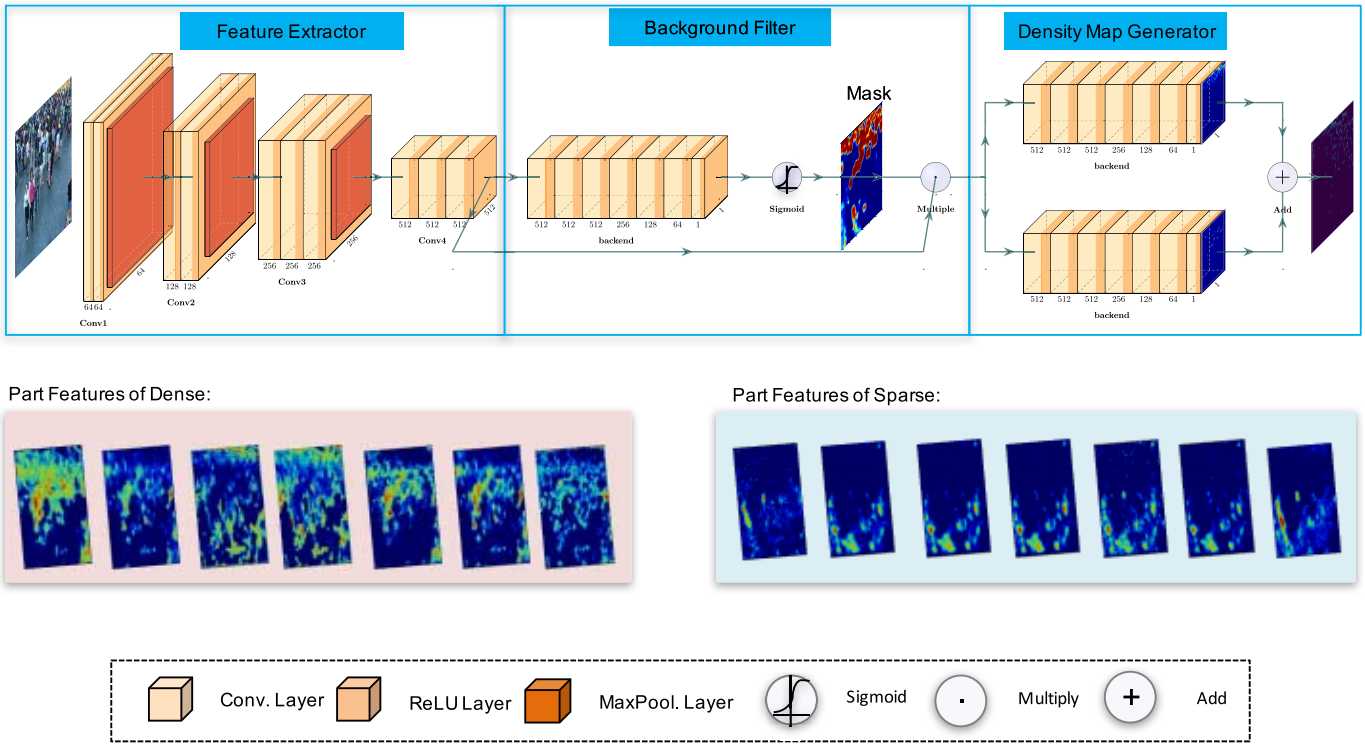


Fig. 7. Framework of "CSR+$\text{mask}_{0.001}^{\text{Feature}}$ + OurDivision". Number under block is the channel of each convolution. "·" means element-wise product, and "+" means element-wise add.

crowd scale. Both of the two models use two branches of structure to reduce the diverse crowd scale and use Mask0.001 in mid-term and Maskours'in the late-term to filter out the background noise, respectively. Besides, CSR-2 [1] is used as the basic structure of our network to show the effectiveness of our method more clearly. We simplify the two models as CSR + $\text{Mask}_{0.001}^{\text{Feature}}$ + OurDivision and CSR + $\text{Mask}_{\text{ours}}^{\text{DensityMap}}$ + OurDivision, respectively.

1) CSR + $Mask_{0.001}^{Feature}$ + OurDivision: This framework is shown in Fig. 7 includes three parts: Feature Extractor, Background Filter, and Density Map Generator. We use the first ten layers of VGG [46] to extract feature maps, and use a set

of dilated convolutions with dilated stride two as the backend in Background Filter and Density Map Generator. A backend structure with a Sigmoid function is just right behind the extracted feature maps to generate the head mask. Then we multiply the extracted feature maps with the predicted head mask to achieve background noise reduction. Following the multiplication are two parallel backends to learn the crowd dividing results generated by our methods separately. The multiply is element-wise production; the GroundTruth of the head mask is Mask0001.

2) CSR + $Mask_{ours}^{DensityMap}$ + OurDivision: The submodules, including the feature extractor and backend, are the same as the submodules in the previous network. Due to the similar structure, we do not present the illustration of this model. Different from the previous framework, three backends are used in parallel after the feature extractors to learn the head mask and the divided crowd density maps, respectively. Besides, $\text{Mask}_{\text{ours}}$ is used as the GroundTruth of the head mask. And then, we multiply the predicted density maps with the generated head mask to reduce the background noise.

For simplicity, in the following chapters, we use Ours-1 to represent the framework of using $\text{Mask}_{0.001}$ to filter the background noise on feature maps, and Ours-2 to represent the framework of using $\text{Mask}_{\text{ours}}$ to filter the background noise on the density map.

IV. Implementation Details

In this section, we introduce our training method in details. The GroundTruth of the head mask is generated based on the rectified head radius estimated by our method.

A. GroundTruth of Density Map

According to the method of generating density maps in [1], we use Gaussian distribution with a deviation of 5 for all the three datasets to blurring each head point. We define the gaussian blurring process as follows,

$$F(x) = \sum_{i=1}^{N} \delta(x - x_i) * G_{\sigma=5}(x), \tag{7}$$

where x is the position of the pixel in the image, and x_i is the position with head annotation, $*$ is the convolution operator, and G denotes a Gaussian distribution with sigma is 5.

B. GroundTruth of Head Mask

To achieve a high-fidelity head mask, we introduce the method to get a more accurate head size by inserting dummy points and passing head size from far- to the near region. Our mask is a hard attention map. We set pixels in the head area to 1, otherwise 0. The head mask can be formulated as follows,

$$\text{Mask}_{\text{ours}} = \begin{cases} 1 & (x - x_i)^2 + (y - y_i)^2 < R_i'^2. \\ 0 & else. \end{cases} \tag{8}$$

C. Data Augmentation

In the training process, we first randomly select a scaling factor from [0.5, 2.0]. If this scaling factor makes the short side of the image less than 512, we set the scaling factor to the ratio of 512 to the length of the short side of the original image. Then, we randomly cropped out 512×512 image patches from the resized images and randomly horizontally flipped them. Besides, to match the size of the network output, we use bilinear interpolation to rescale mask and density map to 1/64 of the original label.

D. Training Details

Our loss includes two parts, mask loss and density map loss. Since the mask is a binary label, we choose the Binary Cross Entropy Loss (BCELoss) to learn mask, shown as follows,

$$l_m(\theta) = -\sum_i P_i^m(\theta) \log Y_i^m + (1 - P_i^m) \log(1 - Y_i^m). \tag{9}$$

To density map, we use Mean Square Error Loss (MSELoss),

$$l_d(\theta) = \sum_i (P_i^d(\theta) - Y_i^d)^2. \tag{10}$$

In Eq. 9 and 10, θ means parameters of our model; P_i^m and and P_i^d means the predicted mask and density map, respectively, Y_i^m and Y_i^d mean the ground truth of head mask and density map, respectively. In addition, Y_i^m in Ours-1 means $\text{Mask}_{0.001}$, while in Ours-2 represents $\text{Mask}_{\text{ours}}$. We show the loss function as follows,

$$\text{Loss} = l_d^S(\theta) + l_d^D(\theta) + \eta l_m(\theta) \tag{11}$$

Although the outputs of the two models are different, Eq. 11 can express the optimization object of the two models. It should be noted that, for Ours-2, l_d^S and l_d^D represent the predicted sparse and dense density map before being multiplied by head mask, respectively. To balance the magnitude of the BCELoss and MSE loss, we multiply BCE loss by an equalization coefficient. We carry out experiments to evaluate the robustness of the parameter by setting η as 1, 0.1, 0.01, 0.001 and 0.0001, respectively. The corresponding results are reported in Table V. As shown, the proposed model is insensitive to the parameter of η within a certain range.

Our model is trained in an end-to-end manner. We initiate the first ten layers of our model with a well-trained VGG-16 [46] and use Gaussian kernels with 0.01 standard deviation to initialize other layers. Adam is chosen as our optimizer with learning rate starting at 1e-4, decaying 0.995 every epoch. We set the batch size to 10, and train our models in 500 epochs.

V. Experiments

Three public datasets, including ShanghaiTech [2], UCF_{QNRF} [3], and $\text{UCF}_{\text{CC\_50}}$ [4], are used to evaluate the performance of our model. We use the standard criterion MAE (mean absolute error), and MSE (mean square error) as

TABLE II
RESULTS ON SHANGHAITECH DATASET

Methods	SHA		SHB	
	MAE	MSE	MAE	MSE
CSR [1]	68.2	115.0	10.6	16.0
SANet [41]	67.0	104.5	8.4	13.6
ASD [48]	65.6	98.0	8.5	13.7
SFCN [49]	64.8	107.5	7.6	13.0
TED [50]	64.2	109.1	8.2	12.8
ADC [14]	63.2	98.9	7.7	12.9
PACNN [52]	66.3	106.4	8.9	13.5
CAN [51]	62.3	100.0	7.8	12.2
SPN [53]	61.7	**99.5**	9.4	14.4
MBTTBF-SCFB [39]	60.2	**94.1**	8.0	15.5
Ours-1	**58.99**	106. 99	6.64	10.93
Ours-2	59.11	106. 21	**6.44**	**10.33**

our assessment, shown as follows,

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|C_i^S + C_i^D - C_i^{GT}|, \tag{12}$$

$$\text{MSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}|C_i^S + C_i^D - C_i^{GT}|^2}. \tag{13}$$

where C_i^S and C_i^D mean the estimated counts of the sparse and dense density map of the i-th image, respectively; C_i^{GT} means the true count of the i-th image.

A. ShanghaiTech

We first evaluate the proposed algorithm on the ShanghaiTech dataset [2], which contains two parts SHA and SHB. The SHA set contains 482 images crawled from the Internet with different resolutions, while SHB contains 716 images with fixed resolution (768 × 1024). The average number of people in each image in Part A is 501, and the number of people varies from 33 to 3138. Part B has a 123 average number of people, and a count ranges in [9, 976]. (We remove the points with negative coordinates.) Intuitively, the crowd in SHB is relatively more sparse than the crowd in SHA. Table II reports the performance of our model and the state-of-the-art methods of [1], [47]–[52]. Our model achieves the best performance among those methods. As shown, Ours-1 and Ours-2 achieve 2.0% and 1.8% improvement on the SHA dataset in MAE compared with MBTTBF-SCFB [39], respectively. Compared with the MAE and MSE on the SHB dataset of SFCN [48], Ours-1 gets 12.6% and 15.9% improvements and Ours-2 gets 15.3% and 20.5% promotion, respectively.

B. UCF_{QNRF}

Recently, Idress *et al.* [3] collect the UCF_{QNRF} dataset with various resolutions and various scenarios. The count of this dataset ranges from 65 to 12863, with an average of 816 annotations in each image. Table III shows the performance of different models on the UCF_{QNRF} dataset. Although SFCN [48] gets improvement on both MAE and MSE in each dataset by pre-training on a large number of synthetic datasets, our model still gets 4.3% MAE improvement than SFCN without using any other additional data.

TABLE III
RESULTS ON UCF_{QNRF} DATASET

Methods	UCF_{QNRF}	
	MAE	MSE
MCNN [2]	277	-
CMTL [38]	252	512
SwithcingCNN [5]	228	445
Resnet101 [54]	190.0	277.0
Densenet201 [55]	163.0	226.0
CL-CNN [3]	132	191
TED [50]	113	188
CAN [51]	107	183
SFCN [49]	102.0	**171.4**
Ours-1	**97.58**	198.79
Ours-2	101.1	191.74

TABLE IV
RESULTS ON $\text{UCF}_{\text{CC\_50}}$ DATASET

Methods	$\text{UCF}_{\text{CC\_50}}$	
	MAE	MSE
ADC [14]	257.1	363.5
TED [50]	249.4	354.5
PACNN [52]	267.9	357.8
SFCN [49]	214.2	318.2
CAN [51]	212.1	243.7
ASD [48]	196.2	270.9
Ours-1	**174.28**	**240.86**
Ours-2	184.18	241.27

C. UCF_{CC_50}

$\text{UCF}_{\text{CC\_50}}$ dataset, proposed by Idress *et al.* [4] in 2013, contains 50 images with high resolutions (average 2101 × 2888) and huge counts (average 1279). Besides, due to the small number of images, we use standard 5-fold cross-validation to verify the accuracy of our model. Each of our training includes 40 training images and 10 test pictures. We conduct five rounds of this training. The 10 test pictures in each group together make up the entire dataset. The results of different methods on this dataset are shown in Table IV. Compared to ASD [47], our model gets 11.2% and 11.1% enhancement in MAE and MSE, respectively.

Frameworks of Ours-1 and Ours-2 consider both the noise filtering mechanism and the crowd division strategy. Results on most of the three datasets show that Ours-1 is better than Ours-2 in MAE, which may demonstrate that when combined with the crowd grouping strategy, filtering on feature maps is more suitable than filtering on the density map.

D. Visual Presentation

We visually present part results of framework Ours-1 and Ours-2 in Fig. 8 and 9. For both of the two figures, the first

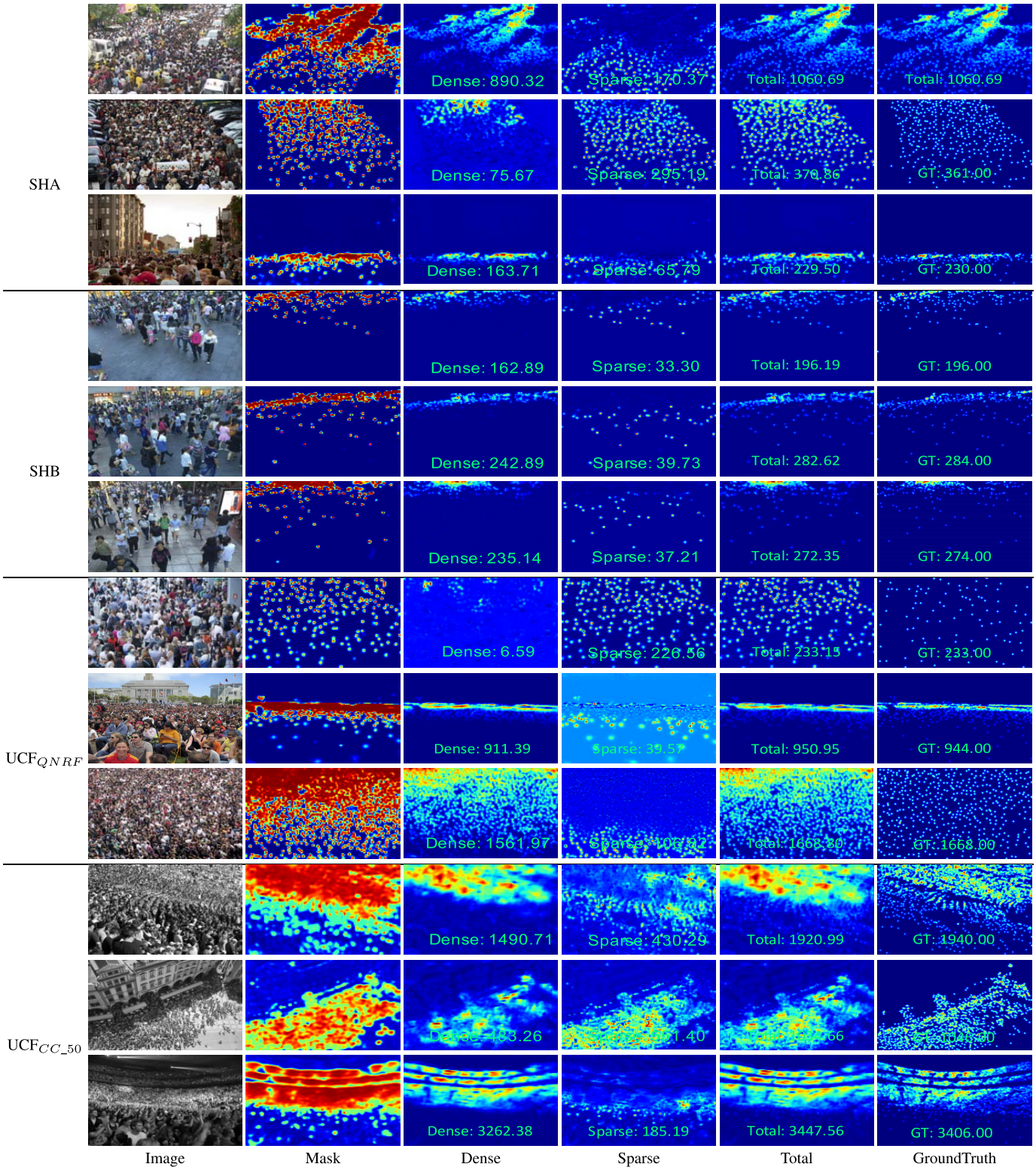


Fig. 8. Some sample results of framework Ours-1 on the three datasets. 'Dense' and 'Sparse' mean the predicted dense and sparse density map, respectively. 'Total' is the sum of 'Dense' and 'Sparse'. As shown in the third and fourth columns, the crowd in different densities can be estimated separately.

column represents the input, and the last column means the Ground Truth of the density map. The rest of the columns in Fig. 8, from left to right, represent the predicted head mask, the predicted dense crowd, the predicted sparse crowd, and the prediction of the whole density map. In Fig. 9, we also show the predicted density map before and after multiplying with the head mask. Columns of the sparse/dense in both Fig. 8 and 9 show that the predicted dense crowd is relatively far from the lens than the sparse crowd, which satisfying the phenomenon of the perspective rule. Moreover, our branch model learns the

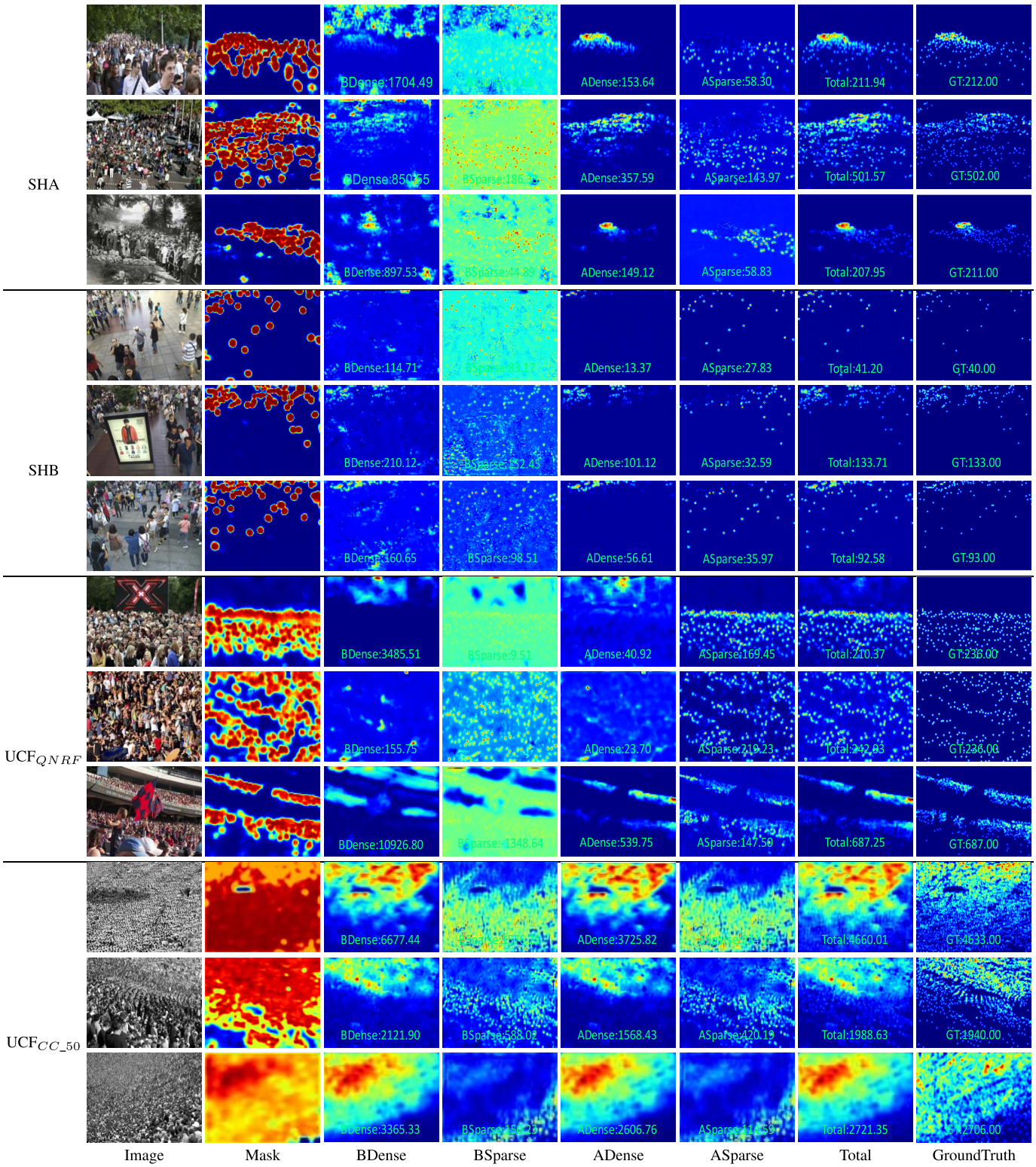


Fig. 9. Some sample results of the proposed framework Ours-2 on the three datasets. 'BDense' and 'BSparse' mean the dense and sparse density before filtering, respectively. 'ADense' and 'ASparse' denote dense and sparse density after filtering, respectively. Resolution of density after filtering is higher than density before filtering. 'Total' is the sum of 'ADense' and 'ASparse'.

crowd in different densities independently decreases the gap between crowd diversity. On the other hand, by comparing the predicted density maps before and after background filtering in Fig. 9, we can intuitively observe that the head mask can effectively filter background noise.

VI. Ablation Study

In this paper, we design two strategies to optimize the prediction. The first scheme uses the head mask to filter out the background noise, and the second one divides the crowd into groups with different densities. We learn the head mask in

TABLE V

"20DIVISION" MEANS DIVIDING SPARSE AND DENSE BY HEAD SIZE WITH 20, "OURDIVISION" REPRESENTS DENSITY DIVIDED ACCORDING TO OUR METHOD PROPOSED IN THIS PAPER. BLACK BOLD DATA IS THE BEST RESULT IN EACH GROUP OF COMPARISON EXPERIMENTS

Methods	SHA		SHB	
	MAE	MSE	MAE	MSE
CSR + Augmentation	62.3	**104.7**	8.0	12.7
CSR + 20Devision	61.48	108.43	7.48	12.41
CSR + OurDevision	**60.10**	106.16	**7.13**	**12.32**
CSR + $\text{Mask}_{\text{ours}}^{\text{FeatureMaps}}$ + 0.0001	60.82	107.66	7.15	**11.55**
CSR + $\text{Mask}_{\text{ours}}^{\text{FeatureMaps}}$ + 0.001	**60.36**	108	**6.83**	11.76
CSR + $\text{Mask}_{\text{ours}}^{\text{FeatureMaps}}$ + 0.01	60.41	112.43	6.99	12.02
CSR + $\text{Mask}_{\text{ours}}^{\text{FeatureMaps}}$ + 0.1	61.11	109.43	7.08	12.02
CSR + $\text{Mask}_{\text{ours}}^{\text{FeatureMaps}}$ + 1	64.63	108.05	7.77	13.51

Sec. III-C and validate the effectiveness of dividing the crowd into different parts in this section.

The ablation study is conducted on the ShanghaiTech dataset. Besides, we use the Front-end and the Back-end of CSR-2 [1] as the base block to verify the effectiveness of our method. For a fair comparison, we use the same data pre-processing and augmentation to retrain CSR-2 [1], and obtain 24.5% MAE and 20.6% MSE improvement on part B. Then we use two Back-ends after the Front-end to crowd in different density independently and use different crowd division results to prove whether it is necessary to divide the crowd according to the density, and which division result is better. The results are shown in Table V. "20Division" means estimating the head size according to [1], and use 20 as the threshold to divide the crowd. "ourDivision" means divide the crowd according to our method. The results of CSRNet with and without division show that any of the two kinds of crowd division method boom the performance. Additionally, the performance of the "20Division" gets 1.3% and 6.5% MAE improvement on SHA and SHB, respectively, while "OurDivision" gets 3.5% and 10.8% MAE enhancement, which demonstrate the superiority of our crowd division method.

VII. CONCLUSION

In this paper, we propose two competitive models that perform robust crowd estimation against background noise and diverse crowd scale. We first present a new approach for head size estimation, which not only use local information but also consider the whole crowd distribution to better handle both sparse and dense crowd. Additionally, we further utilize the estimated head size to divide the crowd into different density parts and generate high-fidelity head masks. By utilizing the crowd partition results, our proposed method learning crowd in different density separately can effectively narrow the gap of various crowd distribution. We believe that learning crowd in different density independently is essential, and 'High fidelity GT head mask' is valuable for filtering background noise. We have conducted extensive experiments to validate the effectiveness of each module. Results show that our method can accomplish high-quality crowd counting quantitatively. More materials will be released at http://nave.vr3i.com.

REFERENCES

[1] 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)*, Jun. 2018, pp. 1091–1100.

[2] 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.*, Jun. 2016, pp. 589–597.

[3] 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.

[4] 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)*, Jun. 2013, pp. 2547–2554.

[5] 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.

[6] D. B. Sam and R. V. Babu, "Top-down feedback for crowd counting convolutional neural network," in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2018.

[7] C. Zhang, H. Li, X. Wang, and X. Yang, "Cross-scene crowd counting via deep convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 833–841.

[8] M. Xu *et al.*, "Depth information guided crowd counting for complex crowd scenes," *Pattern Recognit. Lett.*, vol. 125, pp. 563–569, Jul. 2019.

[9] Z. Shen, Y. Xu, B. Ni, M. Wang, J. Hu, and X. Yang, "Crowd counting via adversarial cross-scale consistency pursuit," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 5245–5254.

[10] D. Song, Y. Qiao, and A. Corbetta, "Depth driven people counting using deep region proposal network," in *Proc. IEEE Int. Conf. Inf. Autom. (ICIA)*, Jul. 2017, pp. 416–421.

[11] S. Gupta, R. Girshick, P. Arbeláez, and J. Malik, "Learning rich features from RGB-D images for object detection and segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2014, pp. 345–360.

[12] H. Fu, H. Ma, and H. Xiao, "Real-time accurate crowd counting based on RGB-D information," in *Proc. 19th IEEE Int. Conf. Image Process.*, Sep. 2012, pp. 2685–2688.

[13] E. Bondi, L. Seidenari, A. D. Bagdanov, and A. Del Bimbo, "Real-time people counting from depth imagery of crowded environments," in *Proc. 11th IEEE Int. Conf. Adv. Video Signal Based Surveill. (AVSS)*, Aug. 2014, pp. 337–342.

[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.*, Jun. 2019, pp. 3225–3234.

[15] L. Zhu, Z. Zhao, C. Lu, Y. Lin, Y. Peng, and T. Yao, "Dual path multi-scale fusion networks with attention for crowd counting," 2019, *arXiv:1902.01115*. [Online]. Available: http://arxiv.org/abs/1902.01115

[16] V. B. Subburaman, A. Descamps, and C. Carincotte, "Counting people in the crowd using a generic head detector," in *Proc. IEEE 9th Int. Conf. Adv. Video Signal-Based Surveill.*, Sep. 2012, pp. 470–475.

[17] P. Viola, M. J. Jones, and D. Snow, "Detecting pedestrians using patterns of motion and appearance," *Int. J. Comput. Vis.*, vol. 63, no. 2, pp. 153–161, Jul. 2005.

[18] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1627–1645, Sep. 2010.

[19] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 1, Jun. 2005, pp. 886–893.

[20] B. Leibe, E. Seemann, and B. Schiele, "Pedestrian detection in crowded scenes," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 1, Jun. 2005, pp. 878–885.

[21] M. Enzweiler and D. M. Gavrila, "Monocular pedestrian detection: Survey and experiments," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 12, pp. 2179–2195, Dec. 2009.

[22] O. Tuzel, F. Porikli, and P. Meer, "Pedestrian detection via classification on Riemannian manifolds," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 10, pp. 1713–1727, Oct. 2008.

[23] S.-F. Lin, J.-Y. Chen, and H.-X. Chao, "Estimation of number of people in crowded scenes using perspective transformation," *IEEE Trans. Syst., Man, Cybern. A, Syst. Humans*, vol. 31, no. 6, pp. 645–654, Nov. 2001.

[24] B. Wu and R. Nevatia, "Detection and tracking of multiple, partially occluded humans by Bayesian combination of edgelet based part detectors," *Int. J. Comput. Vis.*, vol. 75, no. 2, pp. 247–266, Aug. 2007.

[25] J. M. Bernardo and A. F. Smith, *Bayesian Theory*, vol. 405. Hoboken, NJ, USA: Wiley, 2009.

[26] A. K. Menon, "Large-scale support vector machines: Algorithms and theory," Univ. California, San Diego, CA, USA, Tech. Rep., 2009.

[27] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[28] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 1440–1448.

[29] 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.*, 2015, pp. 91–99.

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

[31] S. Kumagai, K. Hotta, and T. Kurita, "Mixture of counting CNNs: Adaptive integration of CNNs specialized to specific appearance for crowd counting," 2017, *arXiv:1703.09393*. [Online]. Available: http://arxiv.org/abs/1703.09393

[32] V. Lempitsky and A. Zisserman, "Learning to count objects in images," in *Proc. Adv. Neural Inf. Process. Syst.*, 2010, pp. 1324–1332.

[33] Z. Yan *et al.*, "Perspective-guided convolution networks for crowd counting," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2019, pp. 952–961.

[34] J. Wan and A. Chan, "Adaptive density map generation for crowd counting," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2019, pp. 1130–1139.

[35] G. Olmschenk, H. Tang, and Z. Zhu, "Improving dense crowd counting convolutional neural networks using inverse k-nearest neighbor maps and multiscale upsampling," 2019, *arXiv:1902.05379*. [Online]. Available: http://arxiv.org/abs/1902.05379

[36] S. Huang *et al.*, "Body structure aware deep crowd counting," *IEEE Trans. Image Process.*, vol. 27, no. 3, pp. 1049–1059, Mar. 2018.

[37] M. Marsden, K. McGuinness, S. Little, and N. E. O'Connor, "ResnetCrowd: A residual deep learning architecture for crowd counting, violent behaviour detection and crowd density level classification," in *Proc. 14th IEEE Int. Conf. Adv. Video Signal Based Surveill. (AVSS)*, Aug. 2017, pp. 1–7.

[38] 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.

[39] V. Sindagi and V. Patel, "Multi-level bottom-top and top-bottom feature fusion for crowd counting," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2019, pp. 1002–1012.

[40] 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 Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5197–5206.

[41] 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.

[42] L. Zhang *et al.*, "Nonlinear regression via deep negative correlation learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Sep. 26, 2019, doi: 10.1109/TPAMI.2019.2943860.

[43] D. Lian, J. Li, J. Zheng, W. Luo, and S. Gao, "Density map regression guided detection network for RGB-D crowd counting and localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 1821–1830.

[44] Y. Liu, M. Shi, Q. Zhao, and X. Wang, "Point in, box out: Beyond counting persons in crowds," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 6469–6478.

[45] C. Godard, O. M. Aodha, M. Firman, and G. J. Brostow, "Digging into self-supervised monocular depth estimation," in *Proc. ICCV*, Oct. 2019, pp. 3828–3838.

[46] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: http://arxiv.org/abs/1409.1556

[47] X. Wu, Y. Zheng, H. Ye, W. Hu, J. Yang, and L. He, "Adaptive scenario discovery for crowd counting," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 2382–2386.

[48] 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.*, Jun. 2019, pp. 8198–8207.

[49] X. Jiang *et al.*, "Crowd counting and density estimation by trellis encoder-decoder networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 6133–6142.

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

[51] M. Shi, Z. Yang, C. Xu, and Q. Chen, "Revisiting perspective information for efficient crowd counting," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 7279–7288.

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

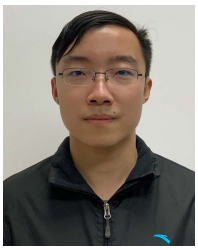
[53] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[54] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 4700–4708.

Hong Mo (Student Member, IEEE) received the M.S. degree in computer science from the Huazhong University of Science and Technology in 2016. She is currently pursuing the Ph.D. degree with the State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, China. Her research interests include deep learning and computer vision.

Wenqi Ren (Member, IEEE) received the Ph.D. degree from Tianjin University, Tianjin, China, in 2017. From 2015 to 2016, he was supported by China Scholarship Council and working with Prof. Ming-Husan Yang as a joint-training Ph.D. student with Electrical Engineering and Computer Science Department, University of California at Merced. He is currently an Assistant Professor with the Institute of Information Engineering, Chinese Academy of Sciences, China. His research interests include image processing and related high-level vision problems.

Yuan Xiong received the M.S. degree in computer science from Clemson University in 2014. He is currently pursuing the Ph.D. degree with the State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, China. His research interest includes multiple view geometry in computer vision and augmented virtuality.



Xiaoqi Pan (Graduate Student Member, IEEE) received the bachelor's degree from the University of Electronic Science and Technology of China in 2018. He is currently pursuing the master's degree with the State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, China. His research interests include deep learning and computer vision.

Zhong Zhou (Member, IEEE) received the B.S. degree from Nanjing University in 1999 and the Ph.D. degree from Beihang University, Beijing, China, in 2005. He is currently a Professor and the Ph.D. Adviser with the State Key Laboratory of Virtual Reality Technology and Systems, Beihang University. His main research interests include virtual reality/augmented reality/mixed reality, computer vision, and artificial intelligence.

Xiaochun Cao (Senior Member, IEEE) received the B.E. and M.E. degrees in computer science from Beihang University, Beijing, China, and the Ph.D. degree in computer science from the University of Central Florida, Orlando, FL, USA. He has spent about three years with ObjectVideo Inc., as a Research Scientist. From 2008 to 2012, he was a Professor with Tianjin University, Tianjin, China. He has been a Professor with the Institute of Information Engineering, Chinese Academy of Sciences, Beijing, since 2012. He has authored or coauthored over 120 journal articles and conference papers. He is a fellow of the IET. His dissertation was nominated for the University of Central Florida's University-level Outstanding Dissertation Award. In 2004 and 2010, he was a recipient of the Piero Zamperoni Best Student Article Award at the International Conference on Pattern Recognition. He is on the Editorial Board of the IEEE TRANSACTIONS OF IMAGE PROCESSING.

Wei Wu received the Ph.D. degree from the Harbin Institute of Technology, Harbin, China, in 1995. He is currently a Professor with the State Key Laboratory of Virtual Reality Technology and Systems with Beihang University. His current research interests include virtual reality, wireless networking, and distributed interactive systems. He is currently a Chair of the Technical Committee on Virtual Reality and Visualization, China Computer Federation.