Adaptive Dilated Network with Self-Correction Supervision for Counting

Shuai Bai[1], Zhiqun He[2], Yu Qiao[3], Hanzhe Hu[4], Wei Wu[2], Junjie Yan[2]
[1]Beijing University of Posts and Telecommunications [2]SenseTime Group Limited
[3]Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences [4]Peking University
baishuai@bupt.edu.cn {hezhiqun, wuwei, yanjunjie}@sensetime.com
yu.qiao@siat.ac.cn huhz@pku.edu.cn

Abstract

The counting problem aims to estimate the number of objects in images. Due to large scale variation and labeling deviations, it remains a challenging task. The static density map supervised learning framework is widely used in existing methods, which uses the Gaussian kernel to generate a density map as the learning target and utilizes the Euclidean distance to optimize the model. However, the framework is intolerable to the labeling deviations and can not reflect the scale variation. In this paper, we propose an adaptive dilated convolution and a novel supervised learning framework named self-correction (SC) supervision. In the supervision level, the SC supervision utilizes the outputs of the model to iteratively correct the annotations and employs the SC loss to simultaneously optimize the model from both the whole and the individuals. In the feature level, the proposed adaptive dilated convolution predicts a continuous value as the specific dilation rate for each location, which adapts the scale variation better than a discrete and static dilation rate. Extensive experiments illustrate that our approach has achieved a consistent improvement on four challenging benchmarks. Especially, our approach achieves better performance than the state-of-the-art methods on all benchmark datasets.

1. Introduction

The counting task is an important topic in computer vision. There are many practical applications, such as traffic management and congestion estimation under video surveillance. In recent years, the methods using convolutional neural network (CNN) have achieved remarkable progress. However, this task remains challenging, which is mainly faced with two challenges: how to effectively supervise the learning process and address the large scale variation problem?

Firstly, compared to the bounding-box annotation, the dotted annotation is less labour-intensive, which is widely used in most of the counting datasets[11, 50, 15, 14, 43]. However, as shown in Fig. 1 (a), the dotted annotations are not consistent on different targets because of subjective deviation. Most of the existing state-of-the-art methods [48, 20, 23, 37, 26, 30] use Gaussian distribution to generate a density map as the learning target. The model is optimized by comparing the Euclidean (L_2) distance between the target density map and the model estimation. However, there are three limitations in this supervised method: 1) The labeling deviation makes the target density map not accurate. 2) The variance of the Gaussian density map does not match the scale of the target. 3) L_2 loss is sensitive to the deviation of position and the change of variance. As a result, the model cannot learn consistent mapping relationships between density maps and features, which greatly limits the upper bound of the performance. Recently, some works [51, 36, 39, 22] have been proposed to alleviate the inconsistency, including introducing additional loss function (e.g., adversarial loss [34] and structural similarity (SSIM) loss [2]) and fusing the density maps of different variances [42]. These methods mainly focus on the loss function while ignoring the deviation of the target density map.

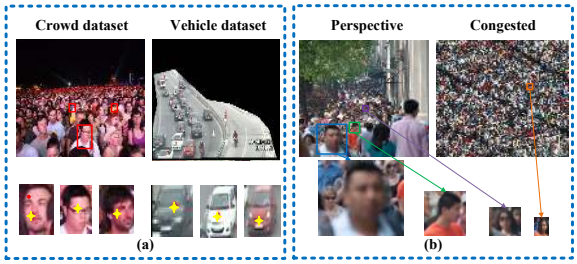

Figure 1. *Two challenges for the counting problem. a) The location of the dotted annotation (yellow points) is inconsistent, whether it is a vehicle or a person. b) There is large scale variation in the same scene and different scenes.*

Secondly, as shown in Fig. 1 (b), there is large scale variation in different scenes. Even in the same scene, the scale still changes dramatically due to perspective phenomenon.

In order to address large scale variation problem, some previous methods use multi-column network[50, 33, 17, 39, 40, 41], stacked multi-branch blocks [2, 25], or multi-dilated decoders [12, 24] to extract features with different receptive fields. Other methods [29, 13, 35, 45] apply different resolution features to estimate the density maps, and then fuse the density maps with the attention maps [13] or perspective maps [35, 8] to obtain the final result. In fact, the value of the scale is continuous, but the aforementioned methods only consider several discrete scales or receptive fields, and there is no way to adapt to a wider range of continuous scale variation. At the same time, extracting multi-scale features will bring more computation load.

Towards the aforementioned issues, we propose a novel supervised learning framework. The framework utilizes the model estimation to correct the annotation with an expectation-maximization (EM) manner, which effectively alleviates the effect of labeling deviations. We consider the density map as a Gaussian mixture model, which consists of K two-dimensional Gaussian distributions, where K is the number of objects in the image. In this setting, the dotted annotation is used to initialize the Gaussian mixture model. The expectation (E) step works as correcting and estimating the responsibility between each position and Gaussian distribution. Then, the maximization (M) step functions work as updating the parameters (e.g., means, the covariances, and the mixing coefficients) of the Gaussian mixture by maximizing the complete data likelihood. The E step and the M step execute alternately. Instead of using L_2 loss to optimize the model, the self-correction (SC) loss is proposed to optimize both the whole and individuals. For the whole, we generate a new density map with the re-estimated parameters as the GT. For the individuals, we introduce the supervision to the mixing coefficient of each Gaussian.

Furthermore, the scale variation is continuous, which means that the continuous receptive field matches the target scale better than the discrete. To distinguish the different regions, the specific receptive fields at different locations are more effective than the same. Based on the analysis above, we have designed an adaptive dilated convolution module. Instead of static and discrete dilation rate, each location has a specific dilation rate to match the scale variation. Moreover, the range of dilation is continuous and it is learned from the preceding feature maps, which costs less computation load than extracting multi-scale feature.

- We propose a novel supervised learning framework, which effectively utilizes results of model learning to progressively correct the labeling deviations. Besides, the self-correction loss is proposed to simultaneously optimize the model from both the whole and the individuals perspective.
- We propose an adaptive dilated convolution, which learns a specific continuous dilation rate to effectively match the scale variation at different locations. Moreover, it gains better performance than multi-scale features fusion or multi-column networks with less computation load.
- Extensive experiments illustrate that our approach has achieved a consistent improvement on four challenging benchmarks. Especially, our approach achieves better performance than the state-of-the-art methods on all benchmark datasets.

2. Related Works

As a crucial topic in computer vision, the counting problem has been researched for many years. The early methods [19, 52, 10, 1, 9, 21] regard it as a detection problem, but it is difficult to detect all targets in congested areas. To improve the accuracy of counting in some extremely dense cases, the methods [3, 4, 5, 32] of direct regression are proposed. Recently, CNN-based methods have achieved remarkable progress. These methods mainly concentrate on solving two challenging problems: large scale variation and lack of effective supervision. We review related works about the counting problem from the aforementioned two aspects.

Methods of alleviating large scale variation. One way to cope with large scale variation is to obtain more rich feature representation. MCNN [50] designs a multi-column convolutional neural network, in which different branches use different kernel sizes to control the size of the receptive fields. Switch-CNN [33] introduces a switch classifier is trained to relay the crowd scene patch to the best branch. SANet [2] applies stacked multi-branch blocks to extract features with different receptive fields. Using a single CNN, CSRNet [20] employs dilated convolution to expand the receptive field, which improves accuracy and proves the effectiveness of dilated convolution. DADNet [12] applies multi-dilated convolution to capture rich spatial context and utilizes deformable convolution to generate a high-quality density map. Another way is to effectively combine features of different resolutions. Hydra-CNN [29] uses a pyramid of image patches extracted at multiple scales to perform the final density prediction. With only one scale, SPN [44] extracts multi-scale features from different layers by the scale pyramid module. SAAN [28] employs the attention mechanism to fuse the density maps estimated by multi-resolution features. PACNN [35] introduces the additional branch to predict the perspective map, which is used to fuse the multi-resolution density maps. Obviously, these methods use discrete receptive fields, which limits their ability to better adapt to continuous scale changes. Moreover, extracting multi-scale features adds more computation load.

Methods of effective supervison. The mainstream state-of-the-art approaches are based on density map super-

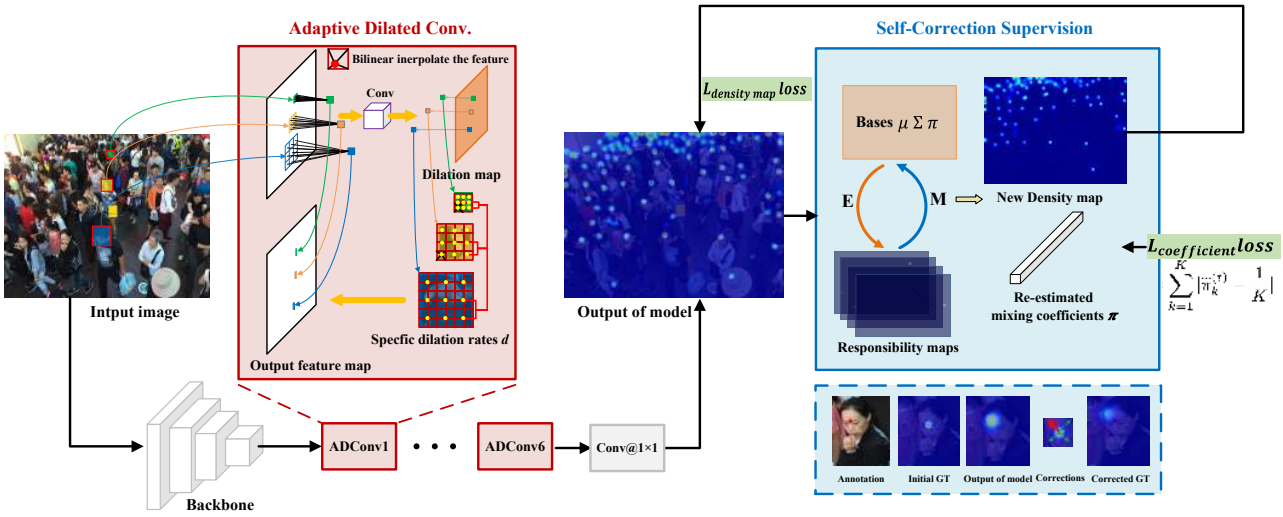

Figure 2. ***Overview of our counting framework.*** *The input image is first fed into the backbone network to obtain feature representation. The decoder consists of six adaptive dilated convolutions and outputs the estimated density map. Each adaptive dilated convolution estimates a specific dilation rate for each location over the input feature. Then, the sampled locations are determined with the dilation rates and the feature is sampled by bilinear interpolation. The current network estimation is used to correct the annotation with an EM-like manner. Furthermore, the SC loss simultaneously optimizes the model from both the whole and the individuals perspective.*

vision. Lempitsky *et al*. [18] start to use Gaussian distribution to generate a density map as the learning target, which is widely adopted by subsequent methods [15, 23, 48, 37]. L_2 loss is commonly used in these CNN based methods. However, this kind of supervised learning framework is intolerable to the inconsistency mapping relationships caused by labeling deviation between density maps and features. Some methods [49, 48, 15, 51, 36, 39] introduce the supervision of additional tasks (e.g., depth maps, segmentation graphs, quantity estimation) to mitigate the effects of inconsistency. SANet [2] adds the local pattern consistency loss to reduce the sensitivity of L_2 loss. CODA [34] introduces an adversarial loss to attenuate the blurry effects of density map. ADMG [42] uses a learned refinement network to fuse the density maps of different variances as a new density map. DSSINet [22] utilizes a dilated multi-scale structural similarity loss to learn the consistency within regions of various sizes. However, most of these methods [2, 34, 47] primarily focus on the design of the loss function and employ hand-craft variance and scale settings.

3. Methodology

We propose a framework for objects counting, which is shown in Fig. 2. It consists of the adaptive dilated convolution network and the self-correction supervision. In this section, firstly, we revisit the conventional target density map from the Gaussian Mixture Model (GMM). Then we will present a novel supervised learning framework, which utilizes the network estimation to correct the annotation with an expectation-maximization manner. Furthermore, we describe the architecture and the operation details of the adaptive dilated convolution.

3.1. Self-Correction Supervision

Gaussian density maps are widely used as the learning target in CNN-based methods. It is formulated as:

$$\mathbf{D}_{gt}(x_n) = \sum_{k=1}^{K} \mathcal{N}(x_n \,|\, \mu_k, \Sigma_k), \tag{1}$$

where $\mathbf{D}$ represents the density map of size $H \times W$. $\mathbf{D}_{est}$ and $\mathbf{D}_{gt}$ denote the estimated and target density maps. x_n denotes the n_{th} two-dimensional location (h_n, w_n) in the image, and X represents the 2D location map of size $2 \times N$. $N = H \times W$ is the number of locations, and K is the number of objects in the image. $\mathcal{N}(x \,|\, \mu_k, \Sigma_k)$ denotes the k_{th} 2D Gaussian distribution. We use μ_k to indicate the k_{th} dotted annotation. Σ_k represents the variance of the k_{th} Gaussian. Commonly, the variance is pre-defined or calculated by the K-means algorithm [50, 15].

Considering the density map divided by K, it can be regard as a Gaussian Mixture Model:

$$p(x_n) = \frac{\mathbf{D}_{gt}(x_n)}{K} = \sum_{k=1}^{K} \pi_k \mathcal{N}(x_n \,|\, \mu_k, \Sigma_k)\,. \tag{2}$$
$$s.t. \sum_{k=1}^{K} \pi_k = 1, 0 \leq \pi_k \leq 1$$

The Gaussian mixture model has K hidden variables and

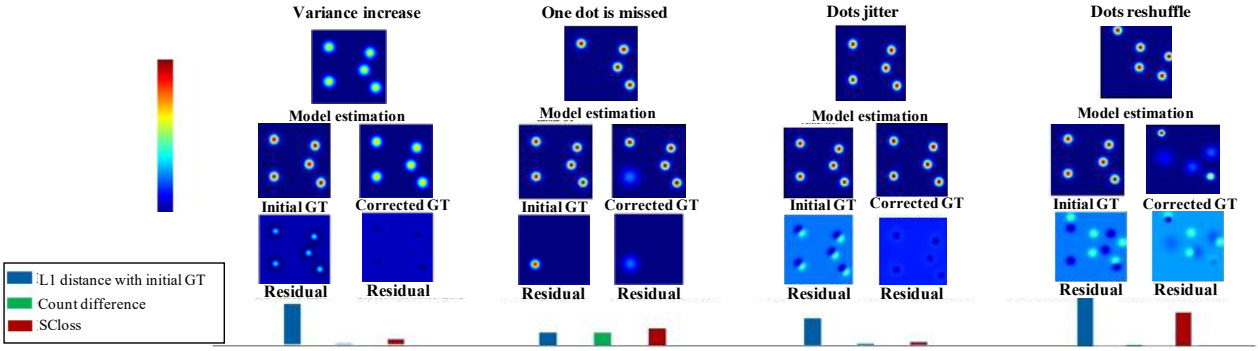

Figure 3. ***Comparison of different supervision methods.*** *Here are four common situations in density map estimation (jitter, reshuffle, missing and the change of Gaussian kernel). In the top row, we visualize the model estimation. In the second row, the initial GT and the corrected GT are shown. The residual between the GT and model estimation is compared in the third row. In the bottom, we compare the sum of the per-pixel L_1 distance without self-correction, the absolute difference of overall counts, and the SC loss. The SC supervision has a unique property that it tolerates the local bias but reacts strongly to the change of the number of objects.*

the mixing coefficient $\pi_k = \frac{1}{K}$, and the dotted annotation μ_k is used as the mean. Due to the subjective bias and the scale matching problem mentioned in section 1, this Gaussian mixture model is not an optimal probability distribution. As the model learns, the model predicts more accurate density maps, but inaccurate annotations limit the upper bound of the performance. Benefitting from features of images, the density map D_{est} estimated by the current network could be better than annotation or at least yiled complementary information, in terms of the consistency of the response locations and the matching of the response range to the target scale to some extent.

This fact inspires us to utilize the current network estimation D_{est} to correct the annotation D_{gt} for generating a more reliable density map for training network in the next time. We propose an EM-like iterative algorithm for this goal. Specially, the annotation is used as the initial parameters of GMM. In the E step, we introduce a responsibility estimation setup, in which the responsibility between each position and each latent distribution are estimated by the current parameters and corrected by the network estimation. In the M step, the parameters of GMM is re-estimated with the current responsibility by maximizing the complete data likelihood. The E step and the M step execute alternately. At each training iteration, the corrected GT will be regenerated. Furthermore, the supervision to the estimated mixing coefficient is introduced to balance the individuals.

Overall, the proposed SC supervision has three key parts, including responsibility estimation, likelihood maximization and self-correction (SC) loss. To simplify the symbols, we reshape $\mathbf{D}_{est}$ into $1 \times N$. It is noteworthy that the probability is non-negative. So we add a $ReLU$ layer behind the output layer. Firstly, the probability matrix $\mathbf{Z}$ of size $K \times H \times W$ is initialized, which represents the conditional probability of x_n belonging to the k_{th} Gaussian (object). The k_{th} matrix of size $H \times W$ in $\mathbf{Z}$ is initialized with a Gaussian distribution with the mean of the dotted annotation and pre-defined variance. Similar with $\mathbf{D}$, we reshape $\mathbf{Z}$ into $K \times N$. Here, we use 0.5 as the initialization value of the variance. As the iteration proceeds, $\mathbf{Z}^t$ is re-generated with new Gaussian distributions based on re-estimated parameters (e.g., $\mu^{(t-1)}$, $\Sigma^{(t-1)}$).

Responsibility estimation. Responsibility estimation works as the E step in the EM algorithm. From the view of dotted annotations, we use the posterior probability to evaluate the responsibility of the n_{th} position and the k_{th} Gaussian distribution. It is formulated as:

$$\mathbf{\Gamma}_{kn}^{(t)} = \frac{\mathbf{Z}_{kn}^{(t-1)}}{\sum_{j=1}^{k} \mathbf{Z}_{kn}^{(t-1)}}, \tag{3}$$

where t denotes the t_{th} iteration, and $\mathbf{Z}_{kn}$ denotes the value of $\mathbf{Z}$ at the position of (k, n). However, the aforementioned responsibility can not reflect the actual data distribution. The corrected responsibility is given by:

$$\mathbf{R}_{kn}^{(t)} = \mathbf{\Gamma}_{kn}^{(t)} \times \mathbf{D}_{est}(x_n). \tag{4}$$

Finally, using Eq. (4), the responsibility matrix is obtained as $\mathbf{R}$ of size $K \times N$.

Likelihood maximization. Likelihood maximization works as the M step of EM algorithm. The parameters are re-estimated with the responsibility matrix:

$$\mu_k^{(t)} = \frac{1}{\mathbf{N}_k^{(t)}} \mathbf{R}_k^{(t)} \times \mathbf{X}^T, \tag{5}$$

$$\Sigma_k^{(t)} = \frac{1}{N_k^{(t)}} \mathbf{R}_k^{(t)} \times ((\mathbf{X} - \mu_k^{(t)}) \cdot (\mathbf{X} - \mu_k^{(t)}))^T, \tag{6}$$

$$\pi_k^{(t)} = \frac{N_k^{(t)}}{\sum_{n=1}^{N} \mathbf{D}_{est}(x_n)}, \tag{7}$$

where $N_k = \sum_{n=1}^{N} \mathbf{R}_{kn}^{(t)}$. Specially, in the counting problem, we can only determine the number of objects in the image and the fact that the probability between each target are same, which means the dimension of the hidden variable is K, and $\pi_k = \frac{1}{K}$. Therefore, we only update the mean and variance, and fix the π_k as $\frac{1}{K}$. In addition, if we consider the limit $\Sigma \to 0$, the log-likelihood function will also go to infinity when $K > 1$. It will cause a pathological solution, so the constraint is necessary. Here, we introduce the constraint $0.5 \leq \Sigma \leq 5$. As responsibility estimation and likelihood maximization executing alternately, $\mathbf{D}_{gt}^{(t)}$ becomes more compatible and reasonable than the initial density map.

Self-correction loss. In general, a more reasonable density map is obtained through online updating. Most methods use the Euclidean (L_2) distance to optimize the model:

$$\mathcal{L}_{L_2} = \sum_{n=1}^{N} |\mathbf{D}_{est}(x_n) - \mathbf{D}_{gt}(x_n)|^2. \tag{8}$$

Here, we use the L_1 distance of pixel-level subregion to supervise the density map:

$$\mathcal{L}_{density\ map} = \frac{1}{N} \sum_{n=1}^{N} |\mathbf{D}_{est}(x_n) - \mathbf{D}_{gt}^{(t)}(x_n)|. \tag{9}$$

However, the mixing coefficients π are not learned in the aforementioned process (Eq. (5) and (6)) since we fix it all the time. From another point of view, $\frac{\sum_{n=1}^{N} \mathbf{\Gamma}_{kn}^{(t)} \times \mathbf{D}_{est}(x_n)}{\sum_{n=1}^{N} \mathbf{D}_{est}(x_n)}$, the mixing coefficients represent the proportions of the targets assigned to the whole distribution. As mentioned above, the proportion of each target should be same, as well as $\pi_k = \frac{1}{K}$. But the re-estimated π_k is not a constant $\frac{1}{K}$ in the Eq. (7). Besides, the sum of estimated map is not accurate. Here, we set $\widetilde{\pi}_k = \frac{\sum_{n=1}^{N} \mathbf{R}_{kn}(x_n)}{\sum_{n=1}^{N} \mathbf{D}_{gt}(x_n)} = \frac{1}{K} \sum_{n=1}^{N} \mathbf{R}_{kn}$. To balance the individuals, we introduce a loss function as:

$$\mathcal{L}_{coefficient} = \sum_{k=1}^{K} |\widetilde{\pi}_k^{(t)} - \frac{1}{K}|. \tag{10}$$

Finally, the proposed SC loss is formulated as:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{density\ map} + \lambda_2 \mathcal{L}_{coefficient}. \tag{11}$$

Here, we simply set $\lambda_1 = \lambda_2 = 1$. Overall, the proposed SC supervision has a number of desirable properties. Firstly, it tolerates the labeling deviation. Dynamically updating the target density map corrects some labeling deviation and helps the model to learn the consistent feature representation. Secondly, it is robust to scale variation. The response area reflects the scale feature of images, and the variance is iteratively adjusted to adapt the response area. Thirdly, it is sensitive to the change of the number of objects. The fluctuation of the mixing coefficients effectively reflect the missed and false detection. These properties are illustrated in Fig. 3.

3.2. Adaptive Dilated Convolution

In order to address large scale variation, many network structures with rich receptive fields have been proposed. There is no doubt that a reasonable receptive field plays an important role in the counting problem. Here, we introduce two designs to the proposed adaptive dilated convolution. 1) From the aspect of scale variation, we use a continuous range of receptive fields to match the continuous scale variation. 2) To learn specific awareness, a specific receptive field is learned for each location.

In detail, a standard 2D convolution with the kernel 3×3 uses a regular grid to sample the input feature map, the grid $\mathbf{G}$ is defined as:

$$\mathbf{G} = \{(-1,-1), (-1,0), ..., (0,1), (1,1)\}. \tag{12}$$

The output feature $\mathbf{F_o}(x_n)$ is calculated as:

$$\mathbf{F_o}(x_n) = \sum_{\Delta x_i \in \mathbf{G}} w(\Delta x_i) \mathbf{F}_i(x_n + \Delta x_i \times d), \tag{13}$$

where d represents the static dilation rate. w denotes the parameters of the convolution.

Commonly, the dilation rate d is a pre-set integer value (e.g.,1,2 and 3) and static. In adaptive dilated convolution, the dilation is adjusted with $\widetilde{d}$, which is dramatic. Then, Eq. (13) becomes

$$\mathbf{F_o}(x_n) = \sum_{\Delta x_i \in \mathbf{G}} w(\Delta x_i) \mathbf{F_i}(x_n + \Delta x_i \times \widetilde{d}_n). \tag{14}$$

For the n_{th} location, the dilation rate is defined as $\widetilde{d}_n$, which is typically fractional. The value of $\mathbf{F_i}(x + \Delta x_i \times \widetilde{d}_n)$ is computed by bilinear interpolation.

As shown in the red box of Fig. 2, through a standard convolutional layer with the kernel size 3×3 and dilation 1 over the same input feature map, the specific dilations $\widetilde{d}$ are estimated. Particularly, we add a $ReLU$ layer to guarantee the dilations are no-negative. The output dilation maps have the same spatial resolution with the input feature maps. The channel dimension becomes 1. The gradients of the dilations are back-propagated through the bilinear operations.

Why is deformable convolution not used here? The deformable convolution [7] introduces unsymmetrical offsets for every position in the sampling grid, which causes the extracted feature has spatial deviations. In the task of object detection, the estimated boxes are corrected by regression to alleviate the deviations. However, the counting problem is position-sensitive task, in which the density and feature of each location need strong consistency. The feature with

Layer	Size	Type
1	$3 \times 3 \times 512$	adconv. + bn + relu
2	$3 \times 3 \times 512$	adconv. + bn + relu
3	$3 \times 3 \times 512$	adconv. + bn + relu
4	$3 \times 3 \times 256$	adconv. + bn + relu
5	$3 \times 3 \times 128$	adconv. + bn + relu
6	$3 \times 3 \times 64$	adconv. + bn + relu
7	$1 \times 1 \times 1$	adconv. + relu

Table 1. *The architecture of the decoder of ADNet. The adaptive dilated convolution is represented as "adconv.".*

spatial deviations will lead to wrongful learning, so adaptive dilated convolution is more reasonable than deformable convolution in the counting problem. Moreover, Compared with predicting offsets for every kernel weight, only predicting one value as the dilation rate is more lightweight.

4. Experiments

4.1. Implementation Details

Network structure. The first ten convolutional layers of VGG16_bn [38, 16] (pretrained on ImageNet [31]) are used as our backbone. The decoder structure is shown in Table 1. The stochastic gradient descent optimizer with an initial learning rate of 0.005 is used to update the parameters, and the learning rate is decayed by gamma 0.2 once the number of epoch reaches one of the milestones. **ADNet** denotes our adaptive dilated network with the conventional supervision. **ADSCNet** represents our adaptive dilated network with the SC supervision.

Training details. We augment the training data using horizontal flipping, random cropping and resizing. Without upsampling, the size of the density map is $\frac{1}{8}$ of the original image, and the batch size of each iteration is 32. Specially, to get a better initialization model, we pre-train our model with conventional L_2 supervision method for 20 epochs. In the SC iterative process, the iterations is set to 2, and we set the initial variance of each Gaussian to 0.5 at the output size.

Evaluation details. The mean absolute error (MAE) and the root mean squared error (MSE) is commonly used as evaluation metrics, which are defined as follows:

$$MAE = \frac{1}{M}\sum_{i=1}^{M}|C_i^{est} - C_i^{gt}|, MSE = \sqrt{\frac{1}{M}\sum_{i=1}^{N}(C_i^{est} - C_i^{gt})^2}, \tag{15}$$

where M is the number of test images. C_i^{est} and C_i^{gt} are the estimated and labeling count number of the i_{th} image. The lower MAE and MSE, the better performance.

Strong baseline. Due to the small datasets and dramatic scene changes, many state-of-the-art methods [20, 25] still train the model with batch size 1, which is time-consuming. As illustrated in Fig. 4, the performance of CSRNet [20] declines with the batch size increasing on ShanghaiTechA with 400 training images, but the CSRNet* can achieve an effective boost as the batch size increases after we introduce the batch normalization layers and the data augmentation (random cropping and resizing). Therefore, our baseline uses VGG16_bn [38, 16] as the backbone and the same decoder with CSRNet* but dilation as 1. In particular, our baseline achieves MAE of 66.5 on ShanghaiTech A, which outperforms the best performance [2] in 2018.

	ADNet	1	2	3	4
MAE (↓)	61.3	56.2	55.4	57.6	58.3
MSE (↓)	103.9	94.8	97.7	98.7	101.3

Table 2. *Performances of models with the different iteration number on the ShanghaiTechA.*

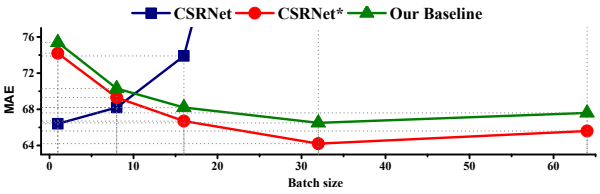

Figure 4. *The effect of batch size and batch normalization layer.*

4.2. Ablation Studies

4.2.1 Self-Correction Supervision

Iteration Number. As illustrated in Table 2, when SC supervision is introduced, MAE has achieved a significant decline. With the increase of the iteration number, MAE first declines and then increases. It gets the best performance when the iteration number reaches 2. Besides, the overall fluctuations are not very large. With the increase of the iteration number, the generated density map will over-fit the estimated result. Especially, the variation σ will be too large which will exceed the target scale. Then the interference from background information is introduced. So an appropriate iteration number 2 is used in our experiment.

Robustness to annotation error. As mentioned in section 1, subjective label deviation is a very common phenomenon in the dotted annotation. In order to further evaluate the robustness of our method to label deviation. We further introduce uniform random noise to the original labeling results. In Fig. 6, as the proportion of noised annotations increases, the MAE of SC supervision is not significantly affected, but the MAE of the conventional supervision method continuously declines. It proves the robustness of SC supervision to the annotation errors.

Expansion capability. In order to verify the expansion capability of the proposed method, we introduce our SC supervision to boost MCNN, CSRNet and our VGG_Baseline.

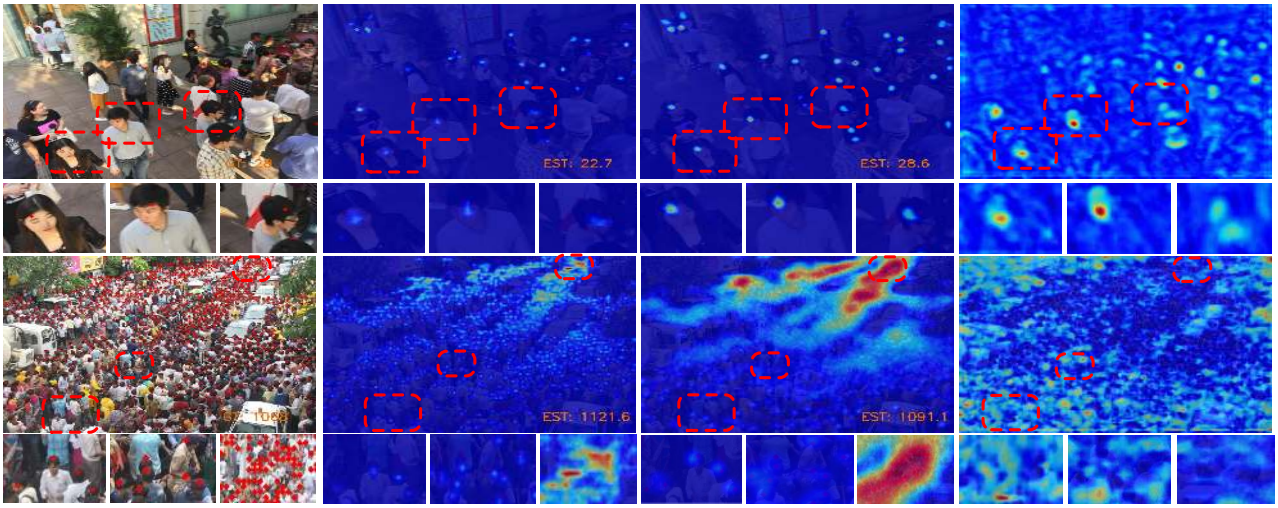

Figure 5. ***Visualization of estimated density maps and dilation maps.*** *Compared with the baseline, the results of SC supervision have consistent response locations (the upper left contours of the head) and uniform response intensity for each person whether in dense or sparse regions. From sparse regions to dense regions, the estimated dilation has obvious decline in (4).*

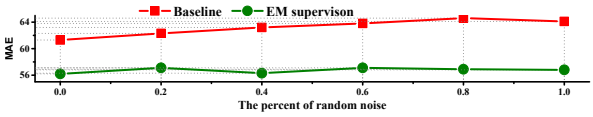

Figure 6. *Robustness evaluations to annotation error.*

Methods	Baseline MAE (↓)	Baseline MSE (↓)	Baseline+SC MAE (↓)	Baseline+SC MSE (↓)
MCNN*	108.2	167.5	101.5	152.4
CSRNet*	64.2	100.6	58.7	98.9
VGG	66.5	106.9	60.7	100.6

Table 3. *The effect of SC supervision on three different methods on the ShanghaiTechA.*

To exclude interference from other factors, we use the same experimental environment and introduce the normalization layer to reimplement MCNN (denoted as MCNN*), CSRNet (denoted as CSRNet*). As shown in Table 3, our SC supervision boosts all three baselines with consistent improvements. They gain relative MAE improvements of 6.19%, 8.57%, 8.72%, which verifies the effectiveness of our SC supervision method.

Visualization of the estimated density maps. We visualize the density maps with different supervision methods in Fig. 5. Firstly, compared with the traditional supervision method, the response positions of SC supervision are more consistent, which mainly concentrate on the upper left contours of the head. It means that the upper left contours of the head is an easily discernible annotation for crowd counting. The response positions of the conventional results are more random (e.g., face, eyes, or head). This shows that SC supervision enables the model to correct the human annotations itself. Secondly, in estimated density maps of the conventional method, the dense-crowd regions are usually underestimated, while sparse-crowd regions are usually overestimated. But the results of SC supervision have uniform response intensity whether in dense-crowd or sparse-crowd regions. It means that the SC loss effectively balances the proportion of the individuals. Thirdly, the density map has different response ranges for different objects, which reflects the scale variation.

4.2.2 Adaptive Dilated Convolution

Effect of dilation rate. In this section, we evaluate the effectiveness of adaptive dilated convolution. For the purpose of comparison, we train multiple variants of baseline. "Dilation-m" indicates the baseline has the static dilation rate m of decoder. "Adaptive-Dila." denotes that the adaptive dilated convolution is introduced. The multi-branch decoder with dilation rates $(1, 3, 5)$ is given as "Multi-Dila.(1,3,5)", and the DADNet [12] is a multi-dilated method. "Deformable-Dila." indicates that the deformable convolution is introduced into the decoder. As illustrated in Table 5, the size of the receptive field influences the performance greatly. The model with dilation 2 achieves the best performance among single static dilated networks. The concatenation of multiple dilated features has a slight improvement but brings heavy computation load. Our ADNet only replaces the dilated convolution in CSRNet [20]

Methods	UCF_QNRF MAE(↓)	UCF_QNRF MSE(↓)	ShanghaiTechA MAE(↓)	ShanghaiTechA MSE(↓)	ShanghaiTech B MAE(↓)	ShanghaiTech B MSE(↓)	UCF_CC_50 MAE(↓)	UCF_CC_50 MSE(↓)	TRANCOS MAE(↓)	TRANCOS MSE(↓)
MCNN(2016) [50]	277	426	110.2	173.2	26.4	41.3	377.6	509.1	-	-
Switch-CNN(2017)[50]	228	445	90.4	135.0	21.6	33.4	318.1	439.2	-	-
ACSCP(2018) [34]	-	-	75.7	102.7	17.2	27.4	291.0	404.6	-	-
CSRNet(2018) [20]	-	-	68.2	115.0	10.6	16.0	266.1	397.5	3.56	-
SANet(2018) [2]	-	-	67.0	104.5	8.4	13.6	258.4	334.9	-	-
CAN(2019) [25]	107	183	62.3	100.0	7.8	12.2	212.2	243.7	-	-
DSSINet(2019)[22]	99.1	159.2	60.6	96.0	6.9	10.3	216.9	302.4	-	-
BL(2019) [27]	88.7	154.8	62.8	101.8	7.7	12.7	229.3	308.2	-	-
SPN(2019)[44]	-	-	61.7	99.5	9.4	14.4	259.2	335.9	3.35	-
SPANet+SANet(2019) [6]	-	-	59.4	92.5	6.5	**9.9**	232.6	311.7	-	-
PGCNet(2019) [46]	-	-	57.0	**86.0**	8.8	13.7	-	-	-	-
Baseline	99.7	161.3	66.5	106.9	8.1	12.5	273.5	357.7	3.21	4.52
Our ADNet	90.1	147.1	61.3	103.9	7.6	12.1	245.4	327.3	2.99	4.28
Our ADSCNet	**71.3**	**132.5**	**55.4**	97.7	**6.4**	11.3	**198.4**	**267.3**	**2.60**	**3.89**

Table 4. *Comparisons with State-of-the-art methods on four datatsets.*

Methods	MAE(↓)	MSE(↓)
Dilation-1	66.5	106.9
Dilation-2	64.2	100.6
Dilation-3	65.7	100.5
Multi-Dila.(1,3,5)	63.6	98.8
DADNet [12]	64.2	99.9
Deformable-Dila.	62.6	97.0
Adaptive-Dila.	61.3	103.9

Table 5. *The effect of different dilation rates on the ShanghaiTechA.*

with adaptive dilated convolution and adds BN layer. The efficiency of the adaptive dilated convolution is between dilated convolution and deformable convolution. Our ADNet improves CSRNet [20] a lot with a small extra computational burden.

Visualization of the dilation maps. As illustrated in Fig. 5, large-scale targets and large-area backgrounds have larger receptive fields, while small-scale targets have smaller receptive fields. In particular, from the large-scale target center to the edge, the value of the dilation has a consecutive from high to low variation, which effectively reflects the scale variation. For the background, a large receptive field is necessary to effectively distinguish it. However, it is difficult for the static dilated network to address the scale variation problem and distinguish the background.

4.3. Comparisons with State-of-the-art

We evaluate our method on four datasets, including crowd datasets ShangHaiTech [50] , UCF_CC_50 [14] ,UCF_QNRF [43] and vehicle dataset TRANCOS [11]. The ShanghaiTech crowd counting dataset consists of two parts: PartA and PartB. PartA is more congested than PartB. UCF_CC_50 is a tiny crowd counting dataset with only 50 images, but it has extremely congested scenes with heavy background noise. The UCF_QNRF dataset is a large and high-resolution crowd counting dataset with 1.25 million head annotations. As an extension, TRANCOS is a vehicle counting dataset with various perspectives.

Table 4 reports the results of four challenging datasets. The proposed method achieves the consistent improvements. Furthermore, it performs better than existing state-of-the-art methods on all the four benchmark datasets. On UCF_QNRF, ADNet and ADSCNet gain relative MAE improvements of 9.6%, 28.5%. The EM supervision is benefit to the high-resolution images. On ShanghaiTech dataset, ADNet and ADSCNet improve the Baseline with relative MAE improvements of 7.8%, 16.7% on part A, and 6.2%,21.0% on part B. Since the labeling deviation in the sparse scenes is more serious, ADSCNet gets more improvement in the sparse scenes than the crowd. In addition, the adaptive dilated convolution bring similar improvement both the sparse and crowd scenes. ADNet and ADSCNet improve the Baseline with relative MAE improvements of 10.3%, 27.5% on UCF_CC_50, and 6.85%, 19.0% on TRANCOS, which indicates that our method has expansion capability to more congested scenes and other objects counting task.

5. Conclusion

In this paper, we present a novel supervised learning framework for the counting problem. It utilizes the model estimation to iteratively correct the annotation and introduces the SC loss to supervise the whole and individuals, which could be integrated into all CNN-based methods. To adapt the large scale variation, the adaptive dilated convolution is proposed, which learns a dynamic and continuous dilation rate for each location. Experiments on four datasets demonstrate that it significantly improve the performance of the baseline. Furthermore, the estimated density map shows the consistent response position and uniform intensity, which illustrates that using the model estimation to correct the annotation is an efficient way to obtain a suitable annotation for network learning.

References

[1] Gabriel J Brostow and Roberto Cipolla. Unsupervised bayesian detection of independent motion in crowds. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 1, pages 594–601. IEEE, 2006.

[2] Xinkun Cao, Zhipeng Wang, Yanyun Zhao, and Fei Su. Scale aggregation network for accurate and efficient crowd counting. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 734–750, 2018.

[3] Antoni B Chan, Zhang-Sheng John Liang, and Nuno Vasconcelos. Privacy preserving crowd monitoring: Counting people without people models or tracking. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–7. IEEE, 2008.

[4] Prithvijit Chattopadhyay, Ramakrishna Vedantam, Ramprasaath R Selvaraju, Dhruv Batra, and Devi Parikh. Counting everyday objects in everyday scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1135–1144, 2017.

[5] Ke Chen, Shaogang Gong, Tao Xiang, and Chen Change Loy. Cumulative attribute space for age and crowd density estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2467–2474, 2013.

[6] Zhi-Qi Cheng, Jun-Xiu Li, Qi Dai, Xiao Wu, and Alexander G. Hauptmann. Learning spatial awareness to improve crowd counting. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[7] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 764–773, 2017.

[8] Diptodip Deb and Jonathan Ventura. An aggregated multicolumn dilated convolution network for perspective-free counting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 195–204, 2018.

[9] Piotr Dollár, Boris Babenko, Serge Belongie, Pietro Perona, and Zhuowen Tu. Multiple component learning for object detection. In *European conference on computer vision*, pages 211–224. Springer, 2008.

[10] Weina Ge and Robert T Collins. Marked point processes for crowd counting. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2913–2920. IEEE, 2009.

[11] Ricardo Guerrero-Gómez-Olmedo, Beatriz Torre-Jiménez, Roberto López-Sastre, Saturnino Maldonado-Bascón, and Daniel Onoro-Rubio. Extremely overlapping vehicle counting. In *Iberian Conference on Pattern Recognition and Image Analysis*, pages 423–431. Springer, 2015.

[12] Dan Guo, Kun Li, Zheng-Jun Zha, and Meng Wang. Dadnet: Dilated-attention-deformable convnet for crowd counting. In *Proceedings of the 27th ACM International Conference on Multimedia*, MM '19, 2019.

[13] Mohammad Hossain, Mehrdad Hosseinzadeh, Omit Chanda, and Yang Wang. Crowd counting using scale-aware attention networks. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1280–1288. IEEE, 2019.

[14] Haroon Idrees, Imran Saleemi, Cody Seibert, and Mubarak Shah. Multi-source multi-scale counting in extremely dense crowd images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2547–2554, 2013.

[15] Haroon Idrees, Muhmmad Tayyab, Kishan Athrey, Dong Zhang, Somaya Al-Maadeed, Nasir Rajpoot, and Mubarak Shah. Composition loss for counting, density map estimation and localization in dense crowds. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 532–546, 2018.

[16] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

[17] Di Kang and Antoni B. Chan. Crowd counting by adaptively fusing predictions from an image pyramid. In *British Machine Vision Conference 2018, BMVC 2018, Northumbria University, Newcastle, UK, September 3-6, 2018*, page 89, 2018.

[18] Victor Lempitsky and Andrew Zisserman. Learning to count objects in images. In *Advances in neural information processing systems*, pages 1324–1332, 2010.

[19] Min Li, Zhaoxiang Zhang, Kaiqi Huang, and Tieniu Tan. Estimating the number of people in crowded scenes by mid based foreground segmentation and head-shoulder detection. In *2008 19th International Conference on Pattern Recognition*, pages 1–4. IEEE, 2008.

[20] Yuhong Li, Xiaofan Zhang, and Deming Chen. Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1091–1100, 2018.

[21] Sheng-Fuu Lin, Jaw-Yeh Chen, and Hung-Xin Chao. Estimation of number of people in crowded scenes using perspective transformation. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 31(6):645–654, 2001.

[22] Lingbo Liu, Zhilin Qiu, Guanbin Li, Shufan Liu, Wanli Ouyang, and Liang Lin. Crowd counting with deep structured scale integration network. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[23] Lingbo Liu, Hongjun Wang, Guanbin Li, Wanli Ouyang, and Liang Lin. Crowd counting using deep recurrent spatial-aware network. *arXiv preprint arXiv:1807.00601*, 2018.

[24] Ning Liu, Yongchao Long, Changqing Zou, Qun Niu, Li Pan, and Hefeng Wu. Adcrowdnet: An attention-injective deformable convolutional network for crowd understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3225–3234, 2019.

[25] Weizhe Liu, Mathieu Salzmann, and Pascal Fua. Context-aware crowd counting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5099–5108, 2019.

[26] Xialei Liu, Joost van de Weijer, and Andrew D Bagdanov. Leveraging unlabeled data for crowd counting by learning to

rank. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7661–7669, 2018.

[27] Zhiheng Ma, Xing Wei, Xiaopeng Hong, and Yihong Gong. Bayesian loss for crowd count estimation with point supervision. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[28] Hossain Mohammad, Hosseinzadeh Mehrdad, Chanda Omit, and Wang Yang. Crowd counting using scale-aware attention networks. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1280–1288, Jan 2019.

[29] Daniel Onoro-Rubio and Roberto J López-Sastre. Towards perspective-free object counting with deep learning. In *European Conference on Computer Vision*, pages 615–629. Springer, 2016.

[30] Viresh Ranjan, Hieu Le, and Minh Hoai. Iterative crowd counting. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 270–285, 2018.

[31] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

[32] David Ryan, Simon Denman, Clinton Fookes, and Sridha Sridharan. Crowd counting using multiple local features. In *2009 Digital Image Computing: Techniques and Applications*, pages 81–88. IEEE, 2009.

[33] Deepak Babu Sam, Shiv Surya, and R Venkatesh Babu. Switching convolutional neural network for crowd counting. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4031–4039. IEEE, 2017.

[34] Zan Shen, Yi Xu, Bingbing Ni, Minsi Wang, Jianguo Hu, and Xiaokang Yang. Crowd counting via adversarial cross-scale consistency pursuit. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5245–5254, 2018.

[35] Miaojing Shi, Zhaohui Yang, Chao Xu, and Qijun Chen. Revisiting perspective information for efficient crowd counting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7279–7288, 2019.

[36] Zenglin Shi, Pascal Mettes, and Cees G. M. Snoek. Counting with focus for free. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[37] Zenglin Shi, Le Zhang, Yun Liu, Xiaofeng Cao, Yangdong Ye, Ming-Ming Cheng, and Guoyan Zheng. Crowd counting with deep negative correlation learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5382–5390, 2018.

[38] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[39] Vishwanath A Sindagi and Vishal M Patel. Cnn-based cascaded multi-task learning of high-level prior and density estimation for crowd counting. In *2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–6. IEEE, 2017.

[40] Vishwanath A Sindagi and Vishal M Patel. Generating high-quality crowd density maps using contextual pyramid cnns. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1861–1870, 2017.

[41] Elad Walach and Lior Wolf. Learning to count with cnn boosting. In *European Conference on Computer Vision*, pages 660–676. Springer, 2016.

[42] Jia Wan and Antoni Chan. Adaptive density map generation for crowd counting. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[43] Xinlong Wang, Tete Xiao, Yuning Jiang, Shuai Shao, Jian Sun, and Chunhua Shen. Repulsion loss: Detecting pedestrians in a crowd. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7774–7783, 2018.

[44] Chen Xinya, Yanrui Bin, Nong Sang, and Changxin Gao. Scale pyramid network for crowd counting. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1941–1950, Jan 2019.

[45] Haipeng Xiong, Hao Lu, Chengxin Liu, Liang Liu, Zhiguo Cao, and Chunhua Shen. From open set to closed set: Counting objects by spatial divide-and-conquer. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[46] Zhaoyi Yan, Yuchen Yuan, Wangmeng Zuo, Xiao Tan, Yezhen Wang, Shilei Wen, and Errui Ding. Perspective-guided convolution networks for crowd counting. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[47] Jie Yang, Jiarou Fan, Yiru Wang, Yige Wang, Weihao Gan, Lin Liu, and Wei Wu. Hierarchical feature embedding for attribute recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[48] Cong Zhang, Hongsheng Li, Xiaogang Wang, and Xiaokang Yang. Cross-scene crowd counting via deep convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 833–841, 2015.

[49] Lu Zhang, Miaojing Shi, and Qiaobo Chen. Crowd counting via scale-adaptive convolutional neural network. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1113–1121, March 2018.

[50] Yingying Zhang, Desen Zhou, Siqin Chen, Shenghua Gao, and Yi Ma. Single-image crowd counting via multi-column convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 589–597, 2016.

[51] Muming Zhao, Jian Zhang, Chongyang Zhang, and Wenjun Zhang. Leveraging heterogeneous auxiliary tasks to assist crowd counting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12736–12745, 2019.

[52] Tao Zhao and Ramakant Nevatia. Bayesian human segmentation in crowded situations. In *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings.*, volume 2, pages II–459. IEEE, 2003.