

# Bayesian Loss for Crowd Count Estimation with Point Supervision

Zhiheng Ma<sup>1\*</sup> Xing Wei<sup>1\*</sup> Xiaopeng Hong<sup>1,2†</sup> Yihong Gong<sup>1</sup>

<sup>1</sup>Faculty of Electronic and Information Engineering, Xi'an Jiaotong University

<sup>2</sup>Research Center for Artificial Intelligence, Peng Cheng Laboratory

mazhiheng@stu.xjtu.edu.cn xingxjtu@gmail.com {hongxiaopeng,ygong}@mail.xjtu.edu.cn

## Abstract

*In crowd counting datasets, each person is annotated by a point, which is usually the center of the head. And the task is to estimate the total count in a crowd scene. Most of the state-of-the-art methods are based on density map estimation, which convert the sparse point annotations into a “ground truth” density map through a Gaussian kernel, and then use it as the learning target to train a density map estimator. However, such a “ground-truth” density map is imperfect due to occlusions, perspective effects, variations in object shapes, etc. On the contrary, we propose Bayesian loss, a novel loss function which constructs a density contribution probability model from the point annotations. Instead of constraining the value at every pixel in the density map, the proposed training loss adopts a more reliable supervision on the count expectation at each annotated point. Without bells and whistles, the loss function makes substantial improvements over the baseline loss on all tested datasets. Moreover, our proposed loss function equipped with a standard backbone network, without using any external detectors or multi-scale architectures, plays favourably against the state of the arts. Our method outperforms previous best approaches by a large margin on the latest and largest UCF-QNRF dataset. The source code is available at <https://github.com/ZhihengCV/Baysian-Crowd-Counting>.*

## 1. Introduction

Counting dense crowds using computer vision techniques has attracted remarkable attentions in recent years. It has a wide range of applications such as estimating the scale of, and counting the number of participants in political rallies, civil unrest, social and sport events, etc. In addition, methods for crowd counting also have great potentials to handle similar tasks in other domains, includ-

ing estimating the number of vehicles in traffic congestion [29, 30, 56, 14, 28], counting the cells and bacteria from microscopic images [20, 42, 44, 45, 8], and animal crowd estimations for ecological survey [27, 1, 18], to name a few.

Crowd counting is a very challenging task because: 1) dense crowds often have heavy overlaps and occlusions between each other; 2) perspective effects may cause large variations in human size, shape, and appearance in the image. In the past decade, a number of crowd counting algorithms [22, 58, 21, 12, 5, 35, 20, 11, 7, 31] have been proposed in the literature. Recently, crowd counting methods using Convolutional Neural Networks (CNNs) have made remarkable progresses [53, 46, 36, 6, 57, 59, 38, 9, 4, 32, 16]. The best performing methods are mostly based on the density map estimation, which typically obtain the crowd count by predicting a density map for the input image and then summing over the estimated density map. Nowadays, publicly available datasets [15, 57, 16] for training crowd count estimators only provide point annotations for each training image, *i.e.*, only one pixel of each person is labeled (typically the center of the head). Currently, the most common approach for using these annotations is to first convert the point annotations for each training image to a “ground-truth” density map using the Gaussian kernel, and then train a CNN model by regressing the value at each pixel in this density map. With such pixel-level strict supervisions, the accuracy of a CNN model is highly dependent on the quality of the obtained “ground-truth” density maps.

Obviously, “ground-truth” density maps obtained by applying a hypothetical Gaussian kernel to the point annotations can hardly be of top quality, due to the occlusions, irregular crowd distributions, large variations in object size, shape, density, *etc.* On the contrary, we propose Bayesian loss, which constructs a density contribution probability model from the point annotations. Then the expected count at each annotated point is calculated by summing the product of the contribution probability and estimated density at each pixel, which can be reliably supervised by the ground-truth count value (apparently, *one*). Compared with previous loss functions that constrain the density value at every

\*Equal contribution.

†Corresponding author.

pixel, our proposed training loss supervises on the count expectation at each annotated point, instead.

Extensive experimental evaluations show that the proposed loss function substantially outperforms the baseline training loss on UCF-QNRF [16], ShanghaiTech [57], and UCF\_CC\_50 [15] benchmark datasets. Moreover, our proposed loss function equipped with the standard VGG-19 network [39] as backbone, without using any external detectors or multi-scale architectures, achieves the state-of-the-art performances on all the benchmark datasets, especially with a magnificent improvement on the UCF-QNRF dataset compared to other methods.

## 2. Related Work

We review related works in the literature on crowd count estimation from the following respects.

**Detection-then-counting.** Most of early works [22, 58, 21, 12] estimate crowd count by detecting or segmenting individual objects in the scene. This kind of methods has to tackle great challenges from two respects. Firstly, they produce more accurate results (*e.g.* bounding-boxes or masks of instances) than the overall count which is computational expensive and mostly suitable in lower density crowds. In overcrowded scenes, clutters and severe occlusions make it unfeasible to detect every single person, despite the progresses in related fields [19, 10, 47, 17, 33, 43, 61, 50, 48, 34, 49, 55, 60, 52]. Secondly, training object detectors require bounding-box or instance mask annotations, which is much more labor-intensive in dense crowds. Thus most of current counting datasets only provide a one-point label per object.

**Direct count regression.** To avoid the more complex detection problem, some researchers proposed to directly learn a mapping from image features to their counts [5, 35, 7, 23, 46, 36, 6]. Former methods [5, 35, 7] in this category rely on hand-crafted features, such as SIFT, LBP *etc.*, and then learn a regression model. Chan *et al.* [5] proposed to extract edge, texture and other low-level features of the crowds, and learn a Gaussian Process regression model for crowd counting. Chen *et al.* [7] proposed to transform low-level image features into a cumulative attribute space where each dimension has clearly defined semantic interpretation that captures how the crowd count value changes continuously and cumulatively. Recent methods [46, 36, 6] resort to deep CNNs for end-to-end learning. Wang *et al.* [46] adopted an AlexNet architecture where the final fully connected layer of 4096 neurons is replaced by a single neuron for predicting the scalar count value. Shang *et al.* [36] proposed to extract a set of high level image features via a CNN firstly, and then map the features to local counts using a Long Short-Term Memory (LSTM) unit. These direct

regression methods are more efficient than detection based methods, however, they do not fully utilized available point supervisions.

**Density map estimation.** This kind of methods [20, 11, 31] take advantage of the location information to learn a map of density values for each training sample and the final count estimation can be obtained by summing over the predicted density map. Lempitsky and Zisserman [20] proposed to transform the point annotations into a density map by the Gaussian kernel as “ground-truth”. Then they train their models using a least-square objective. This kind of training framework has been widely used in recent methods [11, 31]. Furthermore, thanks to the excellent feature learning ability of deep CNNs, CNN based density map estimation methods [53, 57, 51, 38, 26, 25, 4, 32, 16] have achieved the state-of-the-art performance for crowd counting. One major problem of this framework is how to determine the optimal size of the Gaussian kernel which is influenced by many factors. To make matters worse, the models are trained by a loss function which applies supervision in a pixel-to-pixel manner. Obviously, the performance of such methods highly depend on the quality of the generated “ground-truth” density maps.

**Hybrid training.** Several works observed that crowd counting benefits from mixture training strategies, *e.g.*, multi-task, multi-loss, *etc.* Liu *et al.* [24] proposed DecideNet to adaptively decide whether to use a detection model or a density map estimation model. This approach takes the advantage of mixture-of-experts where a detection based model can estimate crowds accurately in low density scenes while the density map estimation model is good at handling crowded scenes. However, this method requires external pre-trained human detection models and is less efficient. Some researchers proposed to combine multiple losses to assist each other. Zhang *et al.* [53] proposed to train a deep CNN by alternatively optimizing a pixel-wise loss function and a global count regression loss. A similar training approach was adopted by Zhang *et al.* [54], in which they first train their model via the density map loss and then add a relative count loss in the last few epochs. Idrees *et al.* [16] proposed a composition loss, which consists of  $\ell_1$ ,  $\ell_2$ , and  $\ell_\infty$  norm losses for the density map and a count regression loss. Compared to these hybrid losses, our proposed single loss function is simpler and more effective.

## 3. The Proposed Method

### 3.1. Background and Motivation

Let  $\{\mathbf{D}(\mathbf{x}_m) \geq 0 : m = 1, 2, \dots, M\}$  be a density map, where  $\mathbf{x}_m$  denotes a 2D pixel location, and  $M$  is the number of pixels in the density map. Let  $\{(\mathbf{z}_n, y_n) : n =$

$1, 2, \dots, N$  denote the point annotation map for a sample image, where  $N$  is the total crowd count,  $\mathbf{z}_n$  is a head point position and  $y_n = n$  is the corresponding label. The point annotation map contains only one pixel for each person (typically the center of the head), which is sparse, and contains no information about the object size and shape. It is difficult to directly use such point annotation maps to train a density map estimator. A common remedy to this difficulty is to convert it to a ‘‘ground-truth’’ density map using the Gaussian kernel.

$$\begin{aligned} \mathbf{D}^{gt}(\mathbf{x}_m) &\stackrel{\text{def}}{=} \sum_{n=1}^N \mathcal{N}(\mathbf{x}_m; \mathbf{z}_n, \sigma^2 \mathbf{1}_{2 \times 2}) \\ &= \sum_{n=1}^N \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\|\mathbf{x}_m - \mathbf{z}_n\|_2^2}{2\sigma^2}\right), \end{aligned} \quad (1)$$

where  $\mathcal{N}(\mathbf{x}_m; \mathbf{z}_n, \sigma^2 \mathbf{1}_{2 \times 2})$  denotes a 2D Gaussian distribution evaluated at  $\mathbf{x}_m$ , with the mean at the annotated point  $\mathbf{z}_n$ , and an isotropic covariance matrix  $\sigma^2 \mathbf{1}_{2 \times 2}$ .

Many recent works use the above ‘‘ground-truth’’ density map as the learning target, and train a density map estimator using the following loss function:

$$\mathcal{L}^{baseline} = \sum_{m=1}^M \mathcal{F}(\mathbf{D}^{gt}(\mathbf{x}_m) - \mathbf{D}^{est}(\mathbf{x}_m)), \quad (2)$$

where  $\mathcal{F}(\cdot)$  is a distance function and  $\mathbf{D}^{est}$  is the estimated density map. If a fix-sized Gaussian kernel is adopted  $\sigma \stackrel{\text{def}}{=} const$ , it is assumed that all the people in a dataset have the same head size and shape, which is obviously not true due to the occlusions, irregular crowd distributions, perspective effects, *etc.* An alternative solution is to use an adaptive Gaussian kernel [57, 16] for each  $n$ :  $\sigma_n \propto d_n$ , where  $d_n$  is a distance that depends on its nearest neighbors in the spatial domain, which assumes that the crowd is evenly distributed. Some other methods utilize specific information such as camera parameters to get a more accurate perspective map, but in general such information is not available.

We argue that the point annotations in the available crowd counting datasets can rather be considered as weak labels for density map estimation. It is more reasonable to take such annotations as priors or likelihoods instead of the learning targets. Loss functions that impose such strict, pixel-to-pixel supervisions as Eq. (2) on density map are not always beneficial to enhance the count estimation accuracy when used to train a CNN model, because it forces the model to learn inaccurate, or even erroneous information.

### 3.2. Bayesian Loss

Let  $\mathbf{x}$  be a random variable that denotes the spatial location and  $y$  be a random variable that represents the annotated head point. Based on the above discussions, instead of

converting point annotations into the ‘‘ground-truth’’ density maps generated by Eq. (1) as the learning targets, we propose to construct likelihood functions of  $\mathbf{x}_m$  given label  $y_n$  from them,

$$p(\mathbf{x} = \mathbf{x}_m | y = y_n) = \mathcal{N}(\mathbf{x}_m; \mathbf{z}_n, \sigma^2 \mathbf{1}_{2 \times 2}). \quad (3)$$

To simplify the notations, we omit the random variable  $\mathbf{x}$  and  $y$  in the following formulations, *e.g.*, Eq. (3) becomes  $p(\mathbf{x}_m | y_n) = \mathcal{N}(\mathbf{x}_m; \mathbf{z}_n, \sigma^2 \mathbf{1}_{2 \times 2})$ . According to Bayes’ theorem, given a pixel location  $\mathbf{x}_m$  in the density map, the posterior probability of  $\mathbf{x}_m$  having the label  $y_n$  can be computed as:

$$\begin{aligned} p(y_n | \mathbf{x}_m) &= \frac{p(\mathbf{x}_m | y_n) p(y_n)}{p(\mathbf{x}_m)} = \frac{p(\mathbf{x}_m | y_n) p(y_n)}{\sum_{n=1}^N p(\mathbf{x}_m | y_n) p(y_n)} \\ &= \frac{p(\mathbf{x}_m | y_n)}{\sum_{n=1}^N p(\mathbf{x}_m | y_n)} = \frac{\mathcal{N}(\mathbf{x}_m; \mathbf{z}_n, \sigma^2 \mathbf{1}_{2 \times 2})}{\sum_{n=1}^N \mathcal{N}(\mathbf{x}_m; \mathbf{z}_n, \sigma^2 \mathbf{1}_{2 \times 2})}. \end{aligned} \quad (4)$$

In the above derivation, the third equality holds as we assume the equal prior probability  $p(y_n)$  for each class label  $y_n$ , *i.e.*  $p(y_n) = \frac{1}{N}$ , without loss of generality. In practice, if we know the prior that crowds are more or less tend to appear at certain places, a tailored  $p(y_n)$  can be applied here.

With the posterior label probability  $p(y_n | \mathbf{x}_m)$  and estimated density map  $\mathbf{D}^{est}$ , we derive the Bayesian loss as follows. let  $c_n^m$  denotes the count that  $\mathbf{x}_m$  contributes to  $y_n$ , and  $c_n$  be the total count associated with  $y_n$ , we have the expectation of  $c_n$  as:

$$\begin{aligned} E[c_n] &= E\left[\sum_{m=1}^M c_n^m\right] = \sum_{m=1}^M E[c_n^m] \\ &= \sum_{m=1}^M p(y_n | \mathbf{x}_m) \mathbf{D}^{est}(\mathbf{x}_m). \end{aligned} \quad (5)$$

Obviously, the ground-truth count  $c_n$  at each annotation point is one, therefore we have the following loss function:

$$\mathcal{L}^{Bayes} = \sum_{n=1}^N \mathcal{F}(1 - E[c_n]), \quad (6)$$

where  $\mathcal{F}(\cdot)$  is a distance function and we adopt  $\ell_1$  distance in our experiments. A special case should be handled when there is no object in a training image. In such scenario we directly force the sum of the density map to zero. Our proposed loss function is differentiable and can be readily applied to a given CNN using the standard back propagation training algorithm.

At the inference stage, we do not have to know the posterior label probability  $p(y_n | \mathbf{x}_m)$  in advance, because

when we sum over the estimated density map, we eliminate  $p(y_n|\mathbf{x}_m)$  as follows:

$$\begin{aligned} C &= \sum_{n=1}^N E[c_n] = \sum_{n=1}^N \sum_{m=1}^M p(y_n|\mathbf{x}_m) \mathbf{D}^{est}(\mathbf{x}_m) \\ &= \sum_{m=1}^M \sum_{n=1}^N p(y_n|\mathbf{x}_m) \mathbf{D}^{est}(\mathbf{x}_m) \quad (7) \\ &= \sum_{m=1}^M \mathbf{D}^{est}(\mathbf{x}_m). \end{aligned}$$

### 3.3. Background Pixel Modelling

For background pixels that are far away from any of the annotation points, it makes no sense to assign them to any head label  $y_n$ . To better model the background pixels, we introduce an additional background label  $y_0 = 0$ , in addition to the head labels  $\{y_n = n : n = 1, 2, \dots, N\}$ . Then, the posterior label probability can be rewritten as:

$$\begin{aligned} p(y_n|\mathbf{x}_m) &= \frac{p(\mathbf{x}_m|y_n)p(y_n)}{\sum_{n=1}^N p(\mathbf{x}_m|y_n)p(y_n) + p(\mathbf{x}_m|y_0)p(y_0)} \\ &= \frac{p(\mathbf{x}_m|y_n)}{\sum_{n=1}^N p(\mathbf{x}_m|y_n) + p(\mathbf{x}_m|y_0)}. \quad (8) \end{aligned}$$

The last equation is simplified with the assumption  $p(y_n) = p(y_0) = \frac{1}{N+1}$ , without loss of generality. Similarly, we have:

$$p(y_0|\mathbf{x}_m) = \frac{p(\mathbf{x}_m|y_0)}{\sum_{n=1}^N p(\mathbf{x}_m|y_n) + p(\mathbf{x}_m|y_0)}. \quad (9)$$

And the expected counts for each person and for the entire background are defined as:

$$E[c_n] = \sum_{m=1}^M p(y_n|\mathbf{x}_m) \mathbf{D}^{est}(\mathbf{x}_m), \quad (10)$$

$$E[c_0] = \sum_{m=1}^M p(y_0|\mathbf{x}_m) \mathbf{D}^{est}(\mathbf{x}_m). \quad (11)$$

In this case, the summation over the whole density map  $\sum_{m=1}^M \mathbf{D}^{est}(\mathbf{x}_m)$  consists of the foreground counts  $\sum_{n=1}^N E[c_n]$  and the background count  $E[c_0]$ . Obviously, we would like the background count to be zero and the foreground count at each annotation point equals to one, thus we have the following enhanced loss function,

$$\mathcal{L}^{Bayes+} = \sum_{n=1}^N \mathcal{F}(1 - E[c_n]) + \mathcal{F}(0 - E[c_0]). \quad (12)$$

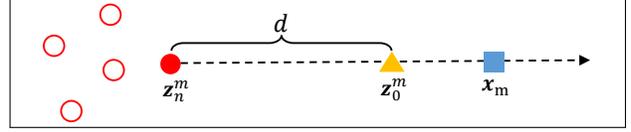


Figure 1: Geometrical illustration of the dummy background point, where  $\mathbf{x}_m$  denotes a pixel in the density map,  $\mathbf{z}_n^m$  is its nearest head point and  $\mathbf{z}_0^m$  is the defined dummy background point.

To define the background likelihood, we construct a dummy background point for each pixel,

$$\mathbf{z}_0^m = \mathbf{z}_n^m + d \frac{\mathbf{x}_m - \mathbf{z}_n^m}{\|\mathbf{x}_m - \mathbf{z}_n^m\|_2}, \quad (13)$$

where  $\mathbf{z}_n^m$  denotes the nearest head point of  $\mathbf{x}_m$ , and  $d$  is a parameter that controls the margin between the head and the dummy background points. As illustrated in Fig. 1, with the defined dummy background point  $\mathbf{z}_0^m$ , for a pixel  $\mathbf{x}_m$  that is far away from head points, it can be assigned to the background label instead. Here we also use the Gaussian kernel to define the background likelihood,

$$\begin{aligned} p(\mathbf{x}_m|y_0) &\stackrel{\text{def}}{=} \mathcal{N}(\mathbf{x}_m; \mathbf{z}_0^m, \sigma^2 \mathbf{1}_{2 \times 2}) \\ &= \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(d - \|\mathbf{x}_m - \mathbf{z}_n^m\|_2)^2}{2\sigma^2}\right). \quad (14) \end{aligned}$$

### 3.4. Visualization and Analysis

We build the entropy map  $\mathbf{Ent}$  of label assignment for visualization and analysis, which is calculated for each pixel  $\mathbf{x}_m$  as follows,

$$\mathbf{Ent}(\mathbf{x}_m) = - \sum_{n=0}^N p(y_n|\mathbf{x}_m) \ln p(y_n|\mathbf{x}_m). \quad (15)$$

The entropy measures the uncertainty on the label a pixel  $\mathbf{x}_m$  in the density map belongs to. We display entropy maps with different settings in Fig. 2 and have the following summarizes:

- The posterior could find the boundary between persons roughly.
- Dense crowd areas have higher entropy values than sparse areas.
- The parameter  $\sigma$  controls the softness of the posterior label probability, comparing (b) and (c).
- Pixels far from crowds are handled better via background pixel modelling, comparing (b) and (e).
- The parameter  $d$  controls the margin between foreground and background, comparing (e) and (f).

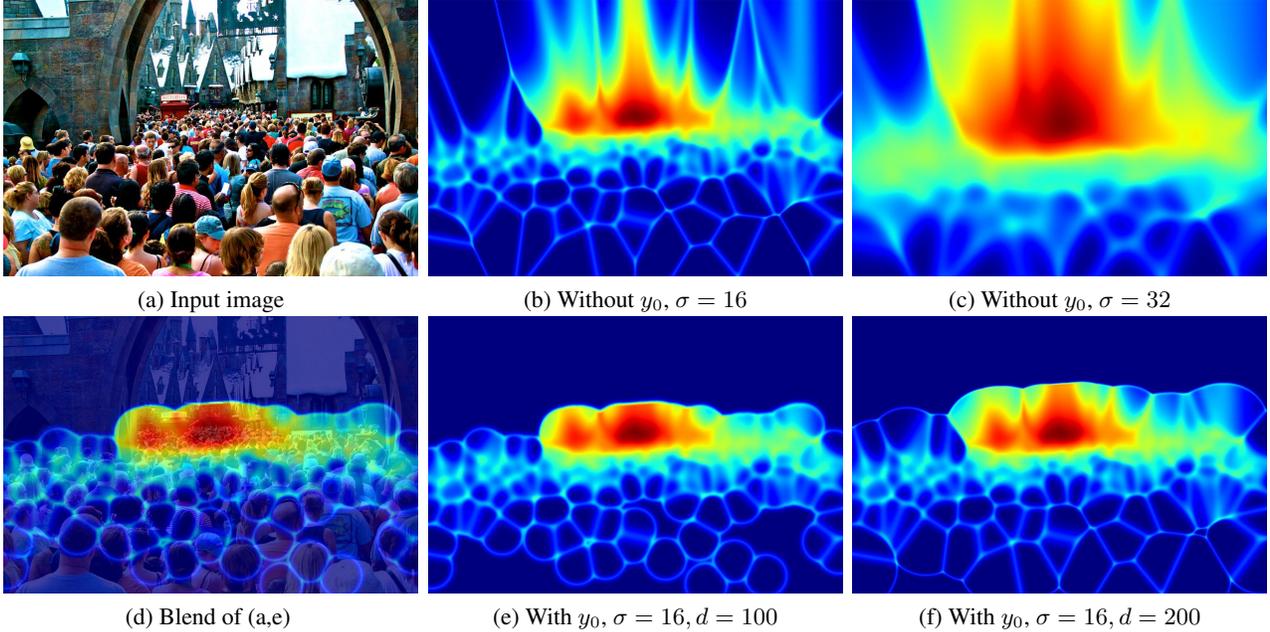


Figure 2: Visualization of the posterior label probability. We construct an entropy map using Eq. (15), which measures the uncertainty on the label a pixel in the density map belongs to. The color is warmer, the value is larger. (a) Input image. (b)-(c): Entropy maps with different  $\sigma$ , without background pixel modelling. (e)-(f): Entropy maps with different  $d$ , with background pixel modelling. (d): Blend of the input image and the entropy map in (e).

## 4. Experiments

### 4.1. Evaluation Metrics

Crowd count estimation methods are evaluated by two widely used metrics: Mean Absolute Error (MAE) and Mean Squared Error (MSE), which are defined as follows:

$$MAE = \frac{1}{K} \sum_{k=1}^K |N_k - C_k|, \quad (16)$$

$$MSE = \sqrt{\frac{1}{K} \sum_{k=1}^K |N_k - C_k|^2}, \quad (17)$$

where  $K$  is the number of test images,  $N_k$  and  $C_k$  are the ground-truth count and the estimated count for the  $k$ -th image, respectively.

### 4.2. Datasets

Experimental evaluations are conducted using four widely used crowd counting benchmark datasets: UCF-QNRF [16], UCF\_CC\_50 [15], ShanghaiTech [57] part A and part B. These datasets are described as follows.

**UCF-QNRF** [16] is the latest and largest crowd counting dataset including 1535 images crawled from Flickr with 1.25 million point annotations. It is a challenging dataset because it has a wide range of counts, image resolutions, light conditions and viewpoints. The training set has 1,201 images and the remaining 334 images are used for testing.

**ShanghaiTech** [57] consists of part A and part B. In part A, there are 300 images for training and 182 images for testing. All the images are crawled from the Internet, and most of them are images of very crowded scenes such as rallies and large sport events. Part B has 400 training images and 316 testing images captured from busy streets in Shanghai. Part A has a significantly higher density than part B.

**UCF\_CC\_50** [15] contains 50 gray images with different resolutions. The average count for each image is 1,280, and the minimum and maximum counts are 94 and 4,532, respectively. Since this is a small-scale dataset and no data split is defined for training and testing, we perform five-fold cross validations to get the average test result.

### 4.3. Implementation Details

**Network structure.** We use a standard image classification network as our backbone, with the last pooling and the subsequent fully connected layers removed. In our experiments, we test two networks which are VGG-19 [39] and AlexNet [17]. We upsample the output of the backbone to  $1/8$  of the input image size by bilinear interpolation, and then feed it to a regression header, which consists of two  $3 \times 3$  convolutional layers with 256 and 128 channels respectively, and a  $1 \times 1$  convolutional layer, to get the density map. The regression header is initialized by the MSRA initializer [13] and the backbone is pre-trained on ImageNet. The Adam optimizer with an initial learning rate  $10^{-5}$  is used to update the parameters.

Datasets Methods	UCF-QNRF		ShanghaiTechA		ShanghaiTechB		UCF_CC_50	
	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE
CROWD-CNN [53]	-	-	181.8	277.7	32.0	49.8	467.0	498.5
MCNN [57]	277	426	110.2	173.2	26.4	41.3	377.6	509.1
CMTL [40]	252	514	101.3	152.4	20.0	31.1	322.8	341.4
SWITCH-CNN [3]	228	445	90.4	135.0	21.6	33.4	318.1	439.2
CP-CNN [41]	-	-	73.6	106.4	20.1	30.1	295.8	320.9
ACS-CNN [37]	-	-	75.7	102.7	17.2	27.4	291.0	404.6
D-CONVNET [38]	-	-	73.5	112.3	18.7	26.0	288.4	404.7
IG-CNN [2]	-	-	72.5	118.2	13.6	21.1	291.4	349.4
IC-CNN [32]	-	-	68.5	116.2	10.7	16.0	260.9	365.5
SANET [4]	-	-	67.0	104.5	8.4	13.6	258.4	334.9
CL-CNN [16]	132	191	-	-	-	-	-	-
BASELINE	106.8	183.7	68.6	110.1	8.5	13.9	251.6	331.3
Our BAYESIAN	92.9	163.0	64.5	104.0	7.9	13.3	237.7	320.8
Our BAYESIAN+	<b>88.7</b>	<b>154.8</b>	<b>62.8</b>	<b>101.8</b>	<b>7.7</b>	<b>12.7</b>	<b>229.3</b>	<b>308.2</b>

Table 1: Benchmark evaluations on four benchmark crowd counting datasets using the MAE and MSE metrics. The baseline and our methods are trained using VGG-19.

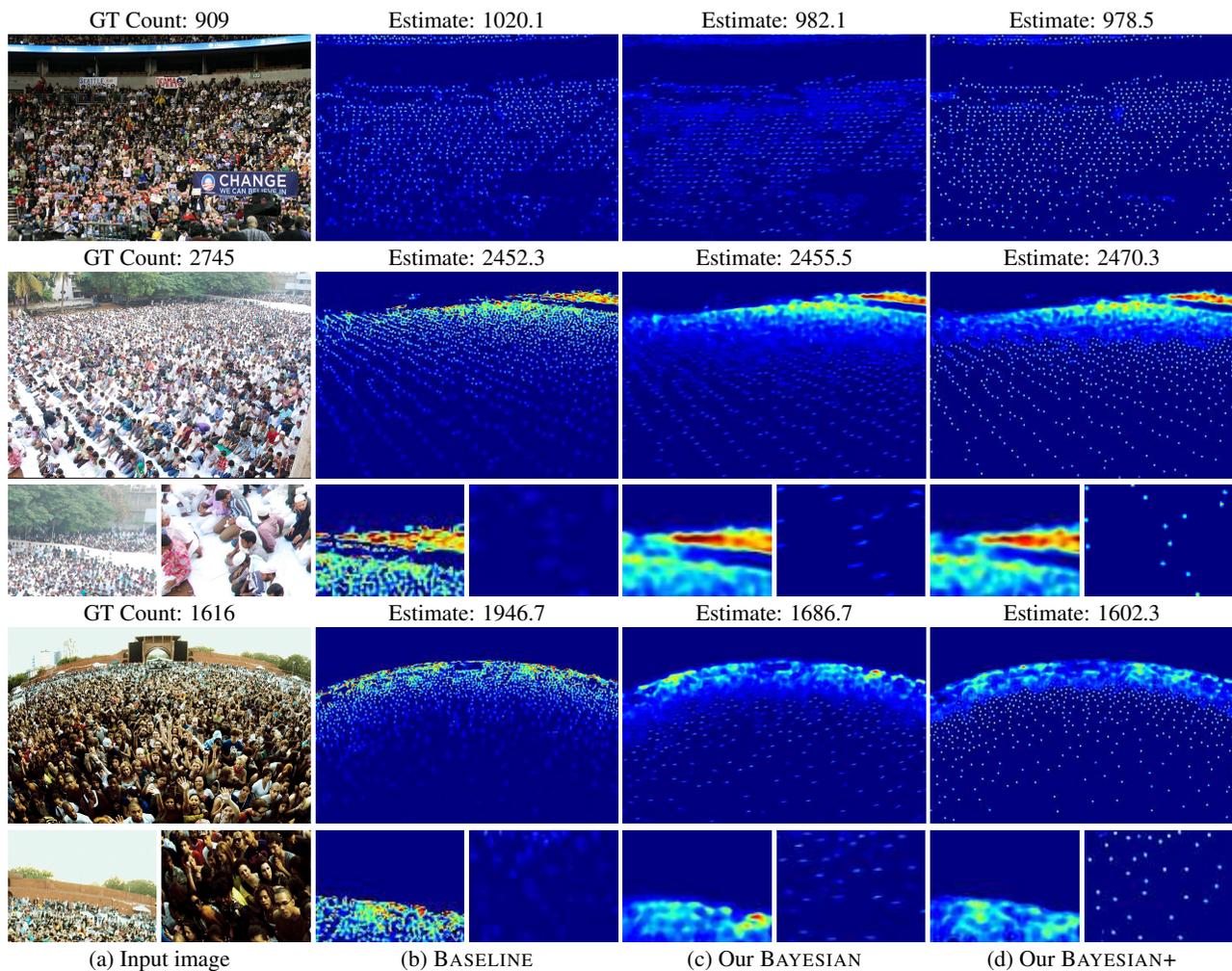


Figure 3: Density maps generated by (b) BASELINE, our (c) BAYESIAN, and (d) BAYESIAN+. The color is warmer, the density is higher. Note that in dense crowds, BASELINE often produces abnormal values, while in sparse areas, it can not localize each person well. In contrast, our methods give more accurate count estimate and localization, respectively.

**Training details.** We augment the training data using random crop and horizontal flipping. We note that image resolutions in UCF-QNRF vary widely from 0.08 to 66 megapixels. However, a regular CNN can not deal with images with all kinds of scales due to its limited receptive field. Therefore, we limit the shorter side of each image within 2048 pixels in UCF-QNRF. Images are then randomly cropped for training, the crop size is  $256 \times 256$  for ShanghaiTechA and UCF\_CC\_50 where image resolutions are smaller, and  $512 \times 512$  for ShanghaiTechB and UCF-QNRF. We set the Gaussian parameter  $\sigma$  in Eqs. (3) and (14) to 8 and the distance parameter  $d$  in Eq. (13) to 15% of the shorter side of image. The parameters are selected on a validation set (120 images randomly sampled from the training set) of UCF-QNRF.

#### 4.4. Experimental Evaluations

**Quantitative results.** We compare our proposed method with the baseline and the state-of-the-art methods on the benchmark datasets described in Sec. 4.2. To make a fair comparison, the baseline method (BASELINE) shares the same network structure (VGG-19) and training process as ours. We use Eq. (1) to generate the “ground-truth” density maps for the baseline method and follow previous works [57, 16] to select parameters for the Gaussian kernel. Specifically, the geometry-adaptive kernels are adopted for UCF-QNRF, ShanghaiTechA and UCF\_CC\_50, while a fixed Gaussian kernel with  $\sigma = 15$  is used for ShanghaiTechB. We study both our basic Bayesian loss (BAYESIAN) and the enhanced Bayesian loss with the background pixel modelling (BAYESIAN+). We show the experimental results in Table 1 and the highlights can be summarized as follows:

- BAYESIAN+ achieves the state-of-the-art accuracy on all the four benchmark datasets. On the latest and the toughest UCF-QNRF dataset, it reduces the MAE and MSE values of the best method (CL-CNN) by 43.3 and 36.2, respectively. It is worth mentioning that our method does not use any external detection models or multi-scale structures.
- BAYESIAN+ consistently improves the performance of BAYESIAN by around 3% on all the four datasets.
- Both BAYESIAN and BAYESIAN+ outperform BASELINE significantly on all the four datasets. BAYESIAN+ makes 15% improvements on UCF-QNRF, 9% on ShanghaiTechA, 8% on ShanghaiTechB, and 8% on UCF\_CC\_50, respectively.

**Visualization of the estimated density maps.** We visualize the estimated density maps using different training losses in Fig. 3. From the close-ups we can see that BASELINE often predicts abnormally values in the congested areas, in contrast, our BAYESIAN and BAYESIAN+ give more

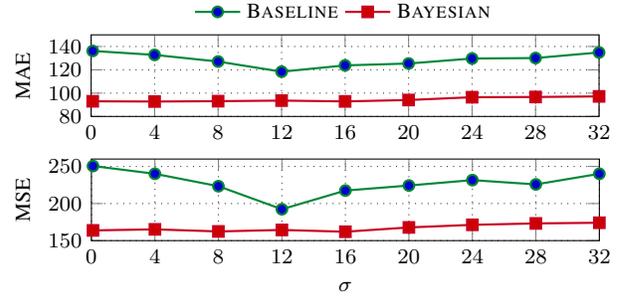


Figure 4: The curves of testing results for different losses w.r.t. the Gaussian parameter  $\sigma$  on UCF-QNRF.

accurate estimations. Our methods benefit from the proposed probability model which constructs soft posterior probabilities if the pixel is close to several head points. In sparse areas, on the other hand, BASELINE can not recognize each person well, while our methods predict more accurate results both on count estimation and localization.

#### 4.5. Ablation Studies

**Effect of  $\sigma$ .** Both the proposed BAYESIAN and the BASELINE methods use the parameter  $\sigma$  for the Gaussian kernel. In our loss function,  $\sigma$  controls the softness of the posterior label probability as shown in Fig. 2, while it determines the crowd density distribution directly in BASELINE. In this subsection, we study the effect of  $\sigma$  on the two methods by computing their MAE and MSE values w.r.t. different  $\sigma$  values on UCF-QNRF. As can be seen from the curves in Fig. 4:

- Our BAYESIAN performs well in a wide range of values of  $\sigma$ . Our MAE and MSE is less than 98.0 and 180.0 when  $\sigma$  changes from 0.1 to 32.0.
- BASELINE is more sensitive to this parameter and its MAE and MSE vary from 118.4 to 136.2 and from 192.3 to 250.6, respectively.

**Effect of  $d$ .** The proposed BAYESIAN+ method introduces an additional parameter  $d$  to control the margin between foreground and background. Fig. 5 shows the performance of BAYESIAN+ w.r.t.  $d$  where we can conclude that:

- Our BAYESIAN+ method performs well in a wide range of values of  $d$ . BAYESIAN+ consistently outperforms BAYESIAN when  $d$  is from 3% to 100% of the shorter side of the image.
- The parameter  $d$  has the meaning that the had size would not exceed  $d$  so that  $d$  should not be too small.

**Robustness to annotation error.** In this subsection, we discuss the robustness of different loss functions w.r.t. annotation error. Labeling a person by a single point is ambiguous, because the person occupies an area in the image.

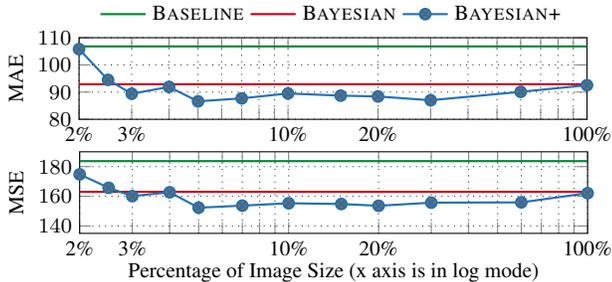


Figure 5: The performance of our BAYESIAN+ method w.r.t.  $d$ .

Although most of the datasets place the annotation point at the center of each head, small errors from human labeling is inevitable. In this experiment, we simulate human labeling errors by adding uniform random noises to the original head positions, and test the performance of different losses at several noise levels. Since  $\sigma$  is the spatial variance of Gaussian distribution, a larger  $\sigma$  is helpful to tolerate such spatial noises. Therefore, we evaluate our BAYESIAN with  $\sigma = 1$  and  $\sigma = 16$ , respectively, and BASELINE with  $\sigma = 16$ . As can be seen from Fig. 6, the proposed BAYESIAN performs better than BASELINE in different noise levels, even with a smaller  $\sigma$  value.

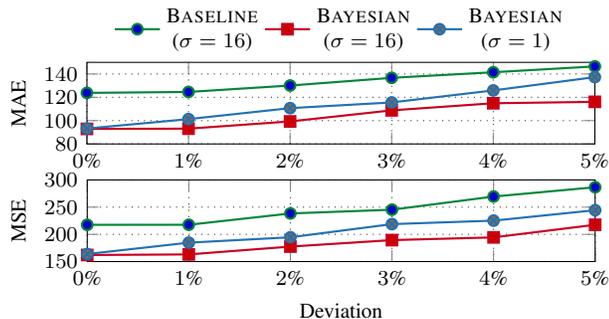


Figure 6: Robustness evaluations to annotation error. We simulate human labeling errors by adding uniform random noises to the annotated point locations (percentage of image height).

**Cross-dataset evaluation.** To further explore the generalization ability of different loss functions, we conduct cross-dataset experiments with the VGG-19 network. In this experiment, models are trained on one dataset and tested on the others without any further fine-tuning. More specifically, we train models on the largest UCF-QNRF dataset, and test them on UCF\_CC\_50, ShanghaiTechA and ShanghaiTechB, respectively. As can be seen from Table. 2, our methods have certain generalization ability and outperform BASELINE on all datasets.

**Limiting the image resolution.** We have found that image resolutions of UCF-QNRF vary widely, and a single CNN model can not handle such a large variation well. Therefore, we limit the image resolution to 2048 pixels and this experiment performs ablation study on this factor. As can be seen

UCF-QNRF → Methods	ShanghaiTechA		ShanghaiTechB		UCF_CC_50	
	MAE	MSE	MAE	MSE	MAE	MSE
BASELINE	73.4	136.3	18.5	30.9	323.9	558.0
Our BAYESIAN	71.7	124.3	16.3	27.8	312.6	540.3
Our BAYESIAN+	<b>69.8</b>	<b>123.8</b>	<b>15.3</b>	<b>26.5</b>	<b>309.6</b>	<b>537.1</b>

Table 2: Quantitative results for cross-dataset evaluation. The models are trained on UCF-QNRF while tested on other datasets.

from Table. 3, all methods benefit from resizing and our methods outperform the baseline method in both settings.

Methods	With Resize		Without Resize	
	MAE	MSE	MAE	MSE
BASELINE	106.8	183.7	128.7	193.8
Our BAYESIAN	92.9	163.0	115.1	187.0
Our BAYESIAN+	<b>88.7</b>	<b>154.8</b>	<b>112.6</b>	<b>181.1</b>

Table 3: The effect of limiting image resolution on UCF-QNRF.

**Different backbones.** Our proposed loss functions can be readily applied to any network structure to improve its performance on the crowd counting task. Here we apply the proposed losses to both VGG-19 and AlexNet and make comparisons with the baseline loss. The quantitative results from Table 4 indicate that our Bayesian loss functions outperform the baseline loss significantly on both the networks.

Backbones Methods	VGG-19 [39]		AlexNet [17]	
	MAE	MSE	MAE	MSE
BASELINE	106.8	183.7	130.8	221.0
Our BAYESIAN	92.9	163.0	121.2	202.5
Our BAYESIAN+	<b>88.7</b>	<b>154.8</b>	<b>116.3</b>	<b>191.7</b>

Table 4: Performances of models using different backbones on UCF-QNRF. Our proposed methods outperform the baseline method consistently.

## 5. Conclusions and Future Work

In this paper, we propose a novel loss function for crowd count estimation with point supervision. Different from previous methods that transform point annotations into the “ground-truth” density maps using the Gaussian kernel with pixel-wise supervision, our loss function adopts a more reliable supervision on the count expectation at each annotated point. Extensive experiments have demonstrated the advantages of our proposed methods in terms of accuracy, robustness and generalization. The current form of our formulation is fairly general and can easily incorporate other knowledge, *e.g.*, specific foreground or background priors, scale and temporal likelihoods, and other facts to further improve the proposed method.

**Acknowledgements.** This work was supported by the National Basic Research Program of China (Grant No. 2015CB351705) and National Major Project of China (Grant No. 2017YFC0803905).

## References

- [1] Carlos Arteta, Victor S. Lempitsky, and Andrew Zisserman. Counting in the wild. In *ECCV*, 2016. 1
- [2] Deepak Babu Sam, Neeraj N. Sajjan, R. Venkatesh Babu, and Mukundhan Srinivasan. Divide and grow: Capturing huge diversity in crowd images with incrementally growing cnn. In *CVPR*, 2018. 6
- [3] Deepak Babu Sam, Shiv Surya, and R. Venkatesh Babu. Switching convolutional neural network for crowd counting. In *CVPR*, 2017. 6
- [4] Xinkun Cao, Zhipeng Wang, Yanyun Zhao, and Fei Su. Scale aggregation network for accurate and efficient crowd counting. In *ECCV*, 2018. 1, 2, 6
- [5] Antoni B. Chan, Zhang-Sheng John Liang, and Nuno Vasconcelos. Privacy preserving crowd monitoring: Counting people without people models or tracking. In *CVPR*, 2008. 1, 2
- [6] Prithvijit Chattopadhyay, Ramakrishna Vedantam, Ramprasaath R. Selvaraju, Dhruv Batra, and Devi Parikh. Counting everyday objects in everyday scenes. In *CVPR*, 2017. 1, 2
- [7] Ke Chen, Shaogang Gong, Tao Xiang, and Chen Change Loy. Cumulative attribute space for age and crowd density estimation. In *CVPR*, 2013. 1, 2
- [8] Joseph Paul Cohen, Genevieve Boucher, Craig A. Glastonbury, Henry Z. Lo, and Yoshua Bengio. Count-caption: Counting by fully convolutional redundant counting. In *ICCV Workshops*, 2017. 1
- [9] Diptodip Deb and Jonathan Ventura. An aggregated multicolored dilated convolution network for perspective-free counting. In *CVPR Workshops*, 2018. 1
- [10] Pedro F Felzenszwalb, David A McAllester, Deva Ramanan, et al. A discriminatively trained, multiscale, deformable part model. In *CVPR*, 2008. 2
- [11] Luca Fiaschi, Ullrich Köthe, Rahul Nair, and Fred A. Hamprecht. Learning to count with regression forest and structured labels. In *ICPR*, 2012. 1, 2
- [12] Weina Ge and Robert T. Collins. Marked point processes for crowd counting. In *CVPR*, 2009. 1, 2
- [13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *ICCV*, 2015. 5
- [14] Meng-Ru Hsieh, Yen-Liang Lin, and Winston H. Hsu. Drone-based object counting by spatially regularized regional proposal network. In *ICCV*, 2017. 1
- [15] Haroon Idrees, Imran Saleemi, Cody Seibert, and Mubarak Shah. Multi-source multi-scale counting in extremely dense crowd images. In *CVPR*, 2013. 1, 2, 5
- [16] 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 *ECCV*, 2018. 1, 2, 3, 5, 6, 7
- [17] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *NIPS*. 2012. 2, 5, 8
- [18] Issam H. Laradji, Negar Rostamzadeh, Pedro O. Pinheiro, David Vazquez, and Mark Schmidt. Where are the blobs: Counting by localization with point supervision. In *ECCV*, 2018. 1
- [19] Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, 2006. 2
- [20] Victor Lempitsky and Andrew Zisserman. Learning to count objects in images. In *NIPS*, 2010. 1, 2
- [21] 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 *ICPR*, pages 1–4, 2008. 1, 2
- [22] Sheng-Fuu Lin, Jaw-Yeh Chen, and Hung-Xin Chao. Estimation of number of people in crowded scenes using perspective transformation. *IEEE Trans. Systems, Man, and Cybernetics, Part A*, 31(6):645–654, 2001. 1, 2
- [23] Bo Liu and Nuno Vasconcelos. Bayesian model adaptation for crowd counts. In *ICCV*, 2015. 2
- [24] Jiang Liu, Chenqiang Gao, Deyu Meng, and Alexander G. Hauptmann. Decidenet: Counting varying density crowds through attention guided detection and density estimation. In *CVPR*, 2018. 2
- [25] Lingbo Liu, Hongjun Wang, Guanbin Li, Wanli Ouyang, and Liang Lin. Crowd counting using deep recurrent spatial-aware network. In *IJCAI*, 2018. 2
- [26] Xialei Liu, Joost van de Weijer, and Andrew D. Bagdanov. Leveraging unlabeled data for crowd counting by learning to rank. In *CVPR*, 2018. 2
- [27] Zheng Ma, Lei Yu, and Antoni B. Chan. Small instance detection by integer programming on object density maps. In *CVPR*, 2015. 1
- [28] Mark Marsden, Kevin McGuinness, Suzanne Little, Ciara E. Keogh, and Noel E. O’Connor. People, penguins and petri dishes: Adapting object counting models to new visual domains and object types without forgetting. In *CVPR*, 2018. 1
- [29] T. Nathan Mundhenk, Goran Konjevod, Wesam A. Sakla, and Kofi Boakye. A large contextual dataset for classification, detection and counting of cars with deep learning. In *ECCV*, 2016. 1
- [30] Daniel Oñoro-Rubio and Roberto Javier López-Sastre. Towards perspective-free object counting with deep learning. In *ECCV*, 2016. 1
- [31] Viet-Quoc Pham, Tatsuo Kozakaya, Osamu Yamaguchi, and Ryuzo Okada. COUNT forest: Co-voting uncertain number of targets using random forest for crowd density estimation. In *ICCV*, 2015. 1, 2
- [32] Viresh Ranjan, Hieu Le, and Minh Hoai. Iterative crowd counting. In *ECCV*, 2018. 1, 2, 6
- [33] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*, 2015. 2
- [34] Weihong Ren, Di Kang, Yandong Tang, and Antoni B. Chan. Fusing crowd density maps and visual object trackers for people tracking in crowd scenes. In *CVPR*, 2018. 2

- [35] David Ryan, Simon Denman, Clinton Fookes, and Sridha Sridharan. Crowd counting using multiple local features. In *DICTA*, 2009. 1, 2
- [36] Chong Shang, Haizhou Ai, and Bo Bai. End-to-end crowd counting via joint learning local and global count. In *ICIP*, 2016. 1, 2
- [37] Zan Shen, Yi Xu, Bingbing Ni, Minsi Wang, Jianguo Hu, and Xiaokang Yang. Crowd counting via adversarial cross-scale consistency pursuit. In *CVPR*, 2018. 6
- [38] Zenglin Shi, Le Zhang, Yun Liu, Xiaofeng Cao, Yangdong Ye, Ming-Ming Cheng, and Guoyan Zheng. Crowd counting with deep negative correlation learning. In *CVPR*, 2018. 1, 2, 6
- [39] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *ICLR*, abs/1409.1556, 2014. 2, 5, 8
- [40] Vishwanath A. Sindagi and Vishal M. Patel. Cnn-based cascaded multi-task learning of high-level prior and density estimation for crowd counting. In *AVSS*, 2017. 6
- [41] Vishwanath A. Sindagi and Vishal M. Patel. Generating high-quality crowd density maps using contextual pyramid cnns. In *ICCV*, 2017. 6
- [42] Korsuk Sirinukunwattana, Shan e Ahmed Raza, Yee-Wah Tsang, David R. J. Snead, Ian A. Cree, and Nasir M. Rajpoot. Locality sensitive deep learning for detection and classification of nuclei in routine colon cancer histology images. *IEEE Trans. Med. Imaging*, 35(5):1196–1206, 2016. 1
- [43] Russell Stewart, Mykhaylo Andriluka, and Andrew Y. Ng. End-to-end people detection in crowded scenes. In *CVPR*, 2016. 2
- [44] Matthias von Borstel, Melih Kandemir, Philip Schmidt, Madhavi K. Rao, Kumar T. Rajamani, and Fred A. Hamprecht. Gaussian process density counting from weak supervision. In *ECCV*, 2016. 1
- [45] Elad Walach and Lior Wolf. Learning to count with CNN boosting. In *ECCV*, 2016. 1
- [46] Chuan Wang, Hua Zhang, Liang Yang, Si Liu, and Xiaochun Cao. Deep people counting in extremely dense crowds. In *ACM MM*, 2015. 1, 2
- [47] Jinjun Wang, Jianchao Yang, Kai Yu, Fengjun Lv, Thomas Huang, and Yihong Gong. Locality-constrained linear coding for image classification. In *CVPR*, 2010. 2
- [48] Xinlong Wang, Tete Xiao, Yuning Jiang, Shuai Shao, Jian Sun, and Chunhua Shen. Repulsion loss: Detecting pedestrians in a crowd. In *CVPR*, 2018. 2
- [49] Xing Wei, Yue Zhang, Yihong Gong, Jiawei Zhang, and Nanning Zheng. Grassmann pooling as compact homogeneous bilinear pooling for fine-grained visual classification. In *ECCV*, 2018. 2
- [50] Xing Wei, Yue Zhang, Yihong Gong, and Nanning Zheng. Kernelized subspace pooling for deep local descriptors. In *CVPR*, 2018. 2
- [51] Feng Xiong, Xingjian Shi, and Dit-Yan Yeung. Spatiotemporal modeling for crowd counting in videos. In *ICCV*, 2017. 2
- [52] Yingyue Xu, Dan Xu, Xiaopeng Hong, Wanli Ouyang, Ji Rongrong, Xu Min, and Guoying Zhao. Structured modeling of joint deep feature and prediction refinement for salient object detection. In *ICCV*, 2019. 2
- [53] Cong Zhang, Hongsheng Li, Xiaogang Wang, and Xiaokang Yang. Cross-scene crowd counting via deep convolutional neural networks. In *CVPR*, 2015. 1, 2, 6
- [54] Lu Zhang, Miaoqing Shi, and Qiaobo Chen. Crowd counting via scale-adaptive convolutional neural network. In *WACV*, 2018. 2
- [55] Shifeng Zhang, Longyin Wen, Xiao Bian, Zhen Lei, and Stan Z. Li. Occlusion-aware r-cnn: Detecting pedestrians in a crowd. In *ECCV*, 2018. 2
- [56] Shanghang Zhang, Guanhang Wu, Joao P. Costeira, and Jose M. F. Moura. Fcn-rlstm: Deep spatio-temporal neural networks for vehicle counting in city cameras. In *ICCV*, 2017. 1
- [57] Yingying Zhang, Desen Zhou, Siqin Chen, Shenghua Gao, and Yi Ma. Single-image crowd counting via multi-column convolutional neural network. In *CVPR*, 2016. 1, 2, 3, 5, 6, 7
- [58] Tao Zhao and Ramakant Nevatia. Bayesian human segmentation in crowded situations. In *CVPR*, 2003. 1, 2
- [59] Zhuoyi Zhao, Hongsheng Li, Rui Zhao, and Xiaogang Wang. Crossing-line crowd counting with two-phase deep neural networks. In *ECCV*, 2016. 1
- [60] Sanping Zhou, Jinjun Wang, Deyu Meng, Yudong Liang, Yihong Gong, and Nanning Zheng. Discriminative feature learning with foreground attention for person re-identification. *IEEE Trans. Image Processing*, 2019. 2
- [61] Sanping Zhou, Jinjun Wang, Jiayun Wang, Yihong Gong, and Nanning Zheng. Point to set similarity based deep feature learning for person re-identification. In *CVPR*, 2017. 2