Learning Spatial Awareness to Improve Crowd Counting

Zhi-Qi Cheng[1,2]*, Jun-Xiu Li[1,3]*, Qi Dai[3], Xiao Wu[1], Alexander G. Hauptmann[2]
[1]Southwest Jiaotong University, [2]Carnegie Mellon University, [3]Microsoft Research
{zhiqic, alex}@cs.cmu.edu, {lijunxiu@my, wuxiaohk@home}.swjtu.edu.cn, qid@microsoft.com

Abstract

The aim of crowd counting is to estimate the number of people in images by leveraging the annotation of center positions for pedestrians' heads. Promising progresses have been made with the prevalence of deep Convolutional Neural Networks. Existing methods widely employ the Euclidean distance (i.e., L_2 loss) to optimize the model, which, however, has two main drawbacks: (1) the loss has difficulty in learning the spatial awareness (i.e., the position of head) since it struggles to retain the high-frequency variation in the density map, and (2) the loss is highly sensitive to various noises in crowd counting, such as the zero-mean noise, head size changes, and occlusions. Although the Maximum Excess over SubArrays (MESA) loss has been previously proposed by [16] to address the above issues by finding the rectangular subregion whose predicted density map has the maximum difference from the ground truth, it cannot be solved by gradient descent, thus can hardly be integrated into the deep learning framework. In this paper, we present a novel architecture called SPatial Awareness Network (SPANet) to incorporate spatial context for crowd counting. The Maximum Excess over Pixels (MEP) loss is proposed to achieve this by finding the pixel-level subregion with high discrepancy to the ground truth. To this end, we devise a weakly supervised learning scheme to generate such region with a multi-branch architecture. The proposed framework can be integrated into existing deep crowd counting methods and is end-to-end trainable. Extensive experiments on four challenging benchmarks show that our method can significantly improve the performance of baselines. More remarkably, our approach outperforms the state-of-the-art methods on all benchmark datasets.

1. Introduction

The problem of crowd counting is described in [16]. Different from visual object detection, it is impossible to provide bounding boxes for all pedestrians due to the extremely dense crowds. On the other side, when only the total crowd counts of the images are provided, the training process will become notably difficult since the spatial awareness is completely ignored. Therefore, to preserve as many spatial constraints as possible and reduce annotation cost, the previous work [16] started to only provide center points of heads and utilizes Gaussian distribution to generate ground truth density maps. It is worth noting that this annotation scheme is widely adopted by subsequent studies.

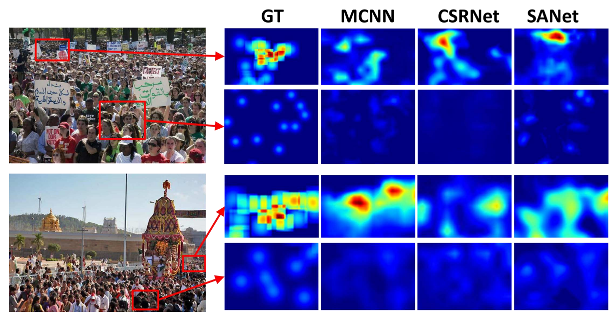

Figure 1: The L_2 loss function has difficulty in learning the spatial awareness and is sensitive to various noises in crowd counting, which will lead to a lower estimation in high-density regions (the first row of each example), and a higher estimation in low-density regions (the second row of each example). Note that the corresponding improvements of our method are shown in Figure 5.

Existing crowd counting approaches mainly focus on improving the scale invariance of feature representation, including the multi-column networks [13, 38, 39, 42, 52, 6], scale aggregation modules [3, 47], and scale-invariant networks [9, 17, 20, 39, 45]. Despite the architectures of these methods are different, the L_2 loss function is employed by most of them. As a result, the spatial awareness in crowd image is largely ignored, though more scale information is embedded into their features.

We have examined three state-of-the-art approaches (i.e., MCNN [52], CSRNet [17], and SANet [3]) on four crowd counting datasets (i.e., ShanghaiTech [52], UCF_CC_50 [11], WorldExpo'10 [48], and UCSD [4]). Two examples are shown in Figure 1. Similar to [3, 19, 20], we observe that dense-crowd regions are usually underestimated, while sparse-crowd regions are overestimated. Such phenomenon is due to two main factors. First, the pixel-wise L_2 loss struggles to retain the high-frequency variation

*indicates equal contribution. This work was done when Zhi-Qi Cheng and Jun-Xiu Li were visiting at Microsoft Research. Xiao Wu is the corresponding author.

in the density map: minimizing L_2 loss encourages finding pixel-wise averages of plausible solutions which are typically overly-smooth and thus have poor spatial awareness [15]. Second, L_2 loss is highly sensitive to typical noises in crowd counting, including the zero-mean noise, head size changes, and head occlusions. We take a simple statistics and show that the co-occurrence of zero-mean noise and overestimation could reach 96% (6,776 out of 7,044 testing images). We further find that almost all estimated density maps inaccurately predict the head positions or sizes when occlusion occurs, which could result in underestimation in high-density areas. Moreover, the generated ground truth density could also be imprecise due to the annotation error and the fixed variance in Gaussian kernel. It is noted that the corresponding improvements of our method are illustrated in Figure 5.

To fully utilize the spatial awareness, previous work [16] proposes a loss named Maximum Excess over SubArrays (MESA) to handle the above problems. Generally speaking, MESA loss attempts to find the rectangular subregion whose predicted density map has the maximum difference from the ground truth. It directly optimizes the counts of this subregion instead of the pixel-level density. Since the set of subregions could include the full image, MESA loss is an upper bound for the count estimation of the entire image. Besides, this loss is only sensitive to the spatial layout of pedestrians and is robust to various noises. However, the complexity of MESA loss function is extremely high. [16] utilizes Cutting-Plane optimization to obtain an approximate solution. Since this method cannot be solved by the conventional gradient descent, MESA loss has not been employed in any existing CNN-based approach.

Motivated by the MESA loss, in this paper we present a novel deep architecture called SPatial Awareness Network (SPANet) to retain the high-frequency spatial variations of density. Instead of finding the mismatched rectangular subregion as in MESA, the Maximum Excess over Pixels (MEP) loss is proposed to optimize the pixel-level subregion which has high discrepancy to the ground truth density map. To obtain such pixel-level subregion, the weakly-supervised ranking information [23] is exploited to generate a mask indicating the pixels with high discrepancies. We further devise a multi-branch architecture to leverage the full image for discrepancy detection by imitating the salience region detection [33, 50, 54], where patches with increasing areas are used for ranking. The proposed framework could be easily integrated into existing CNN-based methods and is end-to-end trainable.

The main contribution of this work is the proposed Spatial Awareness Network and Maximum Excess over Pixels loss for addressing the issue of crowd counting. The solution also provides the elegant views of what kind of spatial context should be exploited and how to effectively utilize such spatial awareness in crowd images, which are problems not yet fully understood in the literature.

2. Related Work

2.1. Detection-based Methods

The methods in this category use object detector to locate people in images. Given the individual localization of each people, crowd counting becomes trivial. There are two directions in this line, i.e., detection on 1) whole pedestrians [2, 7, 53] and 2) parts of pedestrians [8, 12, 18, 43]. Typically, local features [7, 18] are first extracted and then are exploited to train various detectors (e.g., SVM [18] and AdaBoost [41]). Though spatial information is well learned in these methods, they are not applicable in challenging situations, such as the high-density clogging crowds.

2.2. Regression-based Methods

Different from detection-based methods, regression-based approaches avoid the hard detection problem and estimate crowd counts from image features. Earlier methods [4, 5, 11, 28] usually predict the counts directly from the features, which will lead to poor performance as the spatial awareness is completely ignored. Later methods try to estimate the density map for counting [16, 26, 29], where the crowd count is obtained by integrating all pixel values over the density map. Though learning the density map somewhat provides the spatial information, their models still have difficulties in preserving the high-frequency variation in the density map.

2.3. CNN-based Methods

Deep CNN based crowd counting methods have shown very strong performance improvements over the shallow learning counterparts. Existing methods mainly focus on coping with the large variation in pedestrian scales, where many multi-column networks are extensively studied. A dual-column network is proposed by [1] to combine shallow and deep layers for estimating the count. Inspired by this work, a famous three-column network MCNN is proposed by [52], which employs different filters on separate columns to obtain features with various scales. Many works have improved MCNN [13, 38, 39, 42] to further enhance the scale adaptation. Sam *et al.* [32] introduce a switching structure, which uses a classifier to assign input image patches to appropriate columns. Recently, Liu *et al.* [19] propose a multi-column network to simultaneously estimate crowd density by detection and regression based models. Ranjan *et al.* [27] utilize a two-column network to iteratively train their model with images of different resolution.

There are a lot of other attempts to further improve the scale invariance, including 1) study on the fusion of various scale information [22, 40, 45, 46], 2) study on multi-blob based scale aggregation networks [3, 47], 3) design of scale-invariant convolutional or pooling layers [9, 17, 20, 39, 45], and 4) study on the automated scale adaptive networks [30, 31, 49]. Typically, Li *et al.* [17] propose CSRNet that exploits dilated convolutional layers to enlarge receptive fields for boosting performance. Cao *et al.* [3] propose SANet to aggregate multi-scale features for more accurate crowd count. These two approaches have achieved state-of-the-art performance. Additionally, there also exist studies devoted to utilization of perspective maps [35], geometric constraints [21, 51], and region-of-interest (ROI) [20] to improve the counting accuracy.

The aforementioned methods utilize the Euclidean distance, i.e. L_2 loss to optimize the model. Although these methods can obtain scale-invariant features, their performances are still unsatisfactory since the spatial awareness is largely ignored. Note that, SANet [3] also tries to solve the problem of L_2 loss and adds local pattern consistency (L_c loss) in the training phase. However, we find that L_c still cannot learn the spatial context well. In our experiment, when integrating our MEP loss (L_{mep}) into SANet, we achieve significant performance improvement. Our proposed MEP loss could fully utilize the spatial awareness, which is a key factor for the task of crowd counting.

3. Our Method

In this section, we first review the problem of crowd counting and two loss functions (i.e., MESA loss and L_2 loss). Then we present the proposed SPANet and MEP loss in details. It is worth noting that our method can be directly applied to all CNN-based crowd counting networks.

3.1. Problem Formulation

Recent technologies define the crowd counting task as a density regression problem [3, 16, 52]. Given N images $\mathbf{I} = \{I_1, I_2, \cdots, I_N\}$ as the training set, each image I_i is annotated with a total of c_i center points of pedestrians' heads $\mathbf{P}_i^{gt} = \{P_1, P_2, \cdots, P_{c_i}\}$. Typically, the ground truth density map for each pixel p in image I_i is defined as $D^{gt,i}$,

$$\forall p \in I_i, D^{gt,i}(p) = \sum_{P \in \mathbf{P}_i^{gt}} \mathcal{N}^{gt}(p; \mu = P, \sigma^2), \tag{1}$$

where $\mathcal{N}^{gt}$ is a Gaussian distribution. The number of people c_i in image I_i is equal to the sum of density values over all pixels as $\sum_{p \in I_i} D^{gt,i}(p) = c_i$. With these training data, the aim of crowd counting task is to learn the predicted density map D^{pr} towards the ground truth density map D^{gt}.

MESA loss. To make use of the spatial awareness in annotations (i.e., center head positions $\mathbf{P}^{gt}$), the previous work [16] has proposed the Maximum Excess over SubArrays (MESA) loss L_{mesa} as follows,

$$L_{mesa}\left(D^{pr}, D^{gt}\right) = \frac{1}{N} \sum_{i=1}^{N} \max_{B \in \mathbf{B}} \left| \sum_{p \in B} D^{pr,i}(p) - \sum_{p \in B} D^{gt,i}(p) \right|, \tag{2}$$

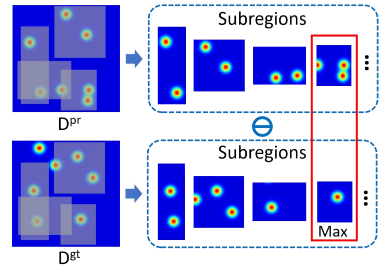

Figure 2: Computation process of MESA loss. It is required to traverse all possible subregions and calculate the differences between their predicted density maps and the ground truth. Then the subregion with maximum difference is selected for optimization.

where $\mathbf{B}$ is the set of all potential rectangular subregions in image. As illustrated in Figure 2, MESA loss tries to find the box subregion whose predicted density map has the maximum difference from the ground truth. It can be treated as an upper bound for the count estimation of the entire image, as $\mathbf{B}$ could include the full image. Besides, this loss is directly related to the counting objective instead of the pixel-level density, and is only sensitive to the spatial layout of pedestrians. In the 1D case, Kolmogorov-Smirnov distance [24] can be seen as a special case of L_{mesa}.

Despite the above merits, it is difficult to optimize the MESA loss due to the hard process of finding such subregion. One has to traverse all potential subregions to achieve this, which is obviously an impossible task in practical application. To solve it, previous approach [16] converts the optimization of MESA loss to a convex quadratic program problem with limited constraints and utilizes Cutting-Plane optimization to obtain an approximate solution. However, since this method cannot be solved by the traditional gradient descent, MESA loss has not been exploited in any existing CNN-based crowd counting methods.

***L_2* loss**. To facilitate the computation in deep frameworks, existing CNN-based methods [17, 27, 52] all directly use L_2 loss to minimize the difference between the estimated and ground truth density maps,

$$L_2\left(D^{pr}, D^{gt}\right) = \frac{1}{2N} \sum_{i=1}^{N} \sum_{p \in D^{pr,i}} \left|\left|D^{pr,i}(p) - D^{gt,i}(p)\right|\right|_2^2. \tag{3}$$

However, as discussed in Sec. 1, we reveal that L_2 loss can hardly retain the high-frequency variation in the density map, leading to the poor spatial awareness. Furthermore, it is also highly sensitive to typical noises in crowd counting, including the zero-mean noise, head size changes, and head occlusions. For example, existing methods always overestimate the density value in low-density areas and underestimate it in high-density regions.

3.2. Spatial Awareness Network

The proposed Spatial Awareness Network (SPANet) aims to leverage the spatial context for accurately predicting the density values. Instead of searching the mismatched

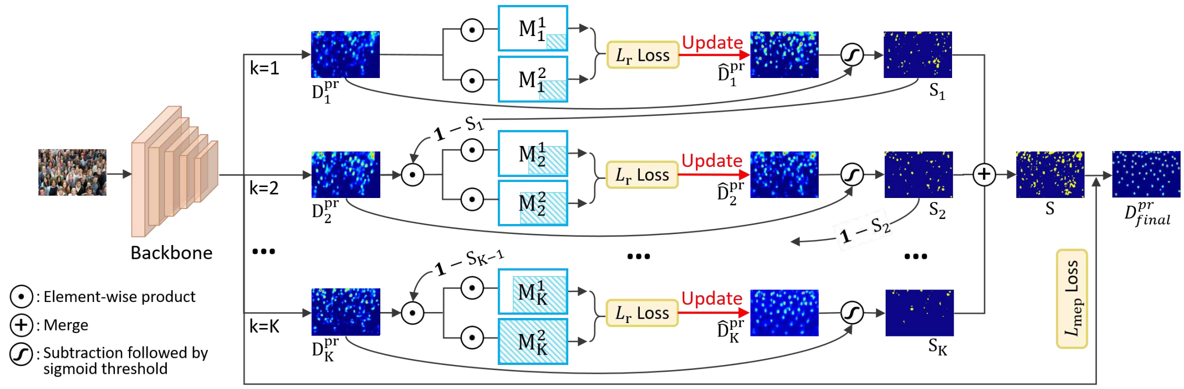

Figure 3: The framework of our proposed SPatial Awareness Network (SPANet). The input images are first fed into the backbone network to extract feature representations and output the estimated density maps D^{pr}. A K-branch architecture is devised. In each branch k, the network is optimized with the ranking objective by sampling two patches (one is sub-patch of the other) and outputs a new density map $\hat{D}_k^{pr}$. Then the two density maps are utilized to produce the subregion S_k which has high discrepancy to the ground truth. The density values within the generated S_k is erased in next branch to facilitate the latter optimization. In the end, K subregions from K branches are fused to form the final pixel-level subregion S, which is exploited to calculate the Maximum Excess over Pixels (MEP) loss.

rectangular subregion as in MESA loss, which is the main obstacle for optimization, we try to find the pixel-level subregion S which has high discrepancy to the ground truth density map. Since there is not any annotation of such region, this problem is unsupervised and will still be significantly difficult to solve. Inspired by the recent weakly-supervised method [23], we exploit an obvious ranking relation to achieve this, i.e., one patch of a crowded scene image is guaranteed to contain the same number or fewer persons than the original image. By sampling a pair of patches (where one is the sub-patch of the other), the network is optimized with the ranking objective and outputs a new density map, which is in turn utilized to produce the subregion with high discrepancy, together with the previous one. We further devise a multi-branch architecture to leverage the full image by sampling multiple pairs of patches. Note that the whole SPANet could be end-to-end trained.

Figure 3 illustrates the framework of our proposed SPANet. Input images I are first fed into the backbone network to generate the predicted density maps D^{pr}. The desired ***pixel-level subregion generation***, i.e., S_k, is conducted by branch k using a pair of patches sampled from density maps D^{pr}. To leverage the full image for discrepancy detection, a ***multi-branch architecture*** with K branches is devised to produce multiple subregions by imitating the salience region detection [50, 54]. Finally, K subregions $(S_1, S_2, ..., S_K)$ are combined to produce the final S, which is then exploited to compute our proposed ***Maximum Excess over Pixels (MEP) loss***. We will present these three sub-modules in details below.

Pixel-level Subregion Generation. The subregion S indicates the area with high density discrepancy to the ground truth. Unfortunately, directly subtracting the predicted D^{pr} from the ground truth D^{gt} would make the problem go round in circles – the bias is usually large enough to prevent it from providing accurate region. Consequently, we turn to find the region with high changes along with the network training. It is natural that one can pick two density maps of the same image from different iterations. However, the obtained area only reflects the region that is already "revised", which still seriously suffers from the poor spatial perception of the original L_2 loss. To this end, we exploit the weakly supervised ranking clues to produce the subregion. Instead of considering the pixel-level density, the ranking clue is directly related to the comparison of crowd counts.

In each branch k, two parallel image patches are first sampled. As the feature maps of deep convolutional layers already contain rich location information, we treat the sampling process as the mask pooling operation on the density map. The strategy of selecting patches will be described later. Without loss of generality, suppose the two masks M_k^1 and M_k^2 are the 2-dimensional matrix with 0 or 1 (1 indicates the patch area), and M_k^1 is the sub-patch of M_k^2. The crowd counts $C(M_k^1)$ and $C(M_k^2)$ under the masks M_k^1 and M_k^2 could be obtained by integrating the values of density map over individual mask, which could be implemented as the mask pooling as follows,

$$
\begin{aligned}
C\left(M_k^1\right) &= \sum_{p \in D_k^{pr}} \left(D_k^{pr} \odot M_k^1\right), \\
C\left(M_k^2\right) &= \sum_{p \in D_k^{pr}} \left(D_k^{pr} \odot M_k^2\right),
\end{aligned}
\tag{4}
$$

where $\odot$ is the element-wise product, and p indicates the pixel on density map D_k^{pr}. It is worth noting that we utilize the same predicted density map D_k^{pr} when calculating the counts for two masks, rather than generating individual maps at two consecutive iterations. The reason is that the density map D_k^{pr} is not restricted to be positive, thus pooling on the pair of patches could also provide the ranking information. We have conducted an experiment showing that the two schemes have similar results. Besides, directly pooling on the same map is more efficient than the other.

With the assumption that M_k^1 is the sub-patch of M_k^2, the explicit constraint is that the number of people in M_k^1 is fewer than that in M_k^2. Therefore, we employ a pairwise ranking hinge loss L_r to model such relationship, which is formulated as

$$L_r\left(C(M_k^1), C(M_k^2)\right) = \max\left(0,\ C(M_k^1) - C(M_k^2) + \xi\right), \tag{5}$$

where ξ is a margin value that is set to the upper bound of the difference in the ground truth. The gradient of L_r loss is calculated as

$$\triangledown_\theta L_r = \begin{cases} 0, & \text{if } C\left(M_k^1\right) - C\left(M_k^2\right) + \xi \leqslant 0, \\ \triangledown_\theta C\left(M_k^1\right) - \triangledown_\theta C\left(M_k^2\right), & \text{otherwise.} \end{cases} \tag{6}$$

Once the network parameters θ are updated with L_r by back-propagation, the renewed density map $\hat{D}_k^{pr}$ estimated by the network is computed by

$$\hat{D}_k^{pr} = \text{Conv}\left(I, \theta\right), \tag{7}$$

where I is the input image, and Conv$(\cdot)$ refers to a forward pass of the network. Given the updated density map $\hat{D}_k^{pr}$ and the old one D_k^{pr}, the desired subregion S_k is obtained by thresholding the difference $\triangledown D_k^{pr}$ between them, where $\triangledown D_k^{pr} = |\hat{D}_k^{pr} - D_k^{pr}|$. To make it differentiable, we utilize a Sigmoid thresholding function, and S_k is given by

$$S_k = \frac{1}{1 + \exp\left(-\delta\left(\triangledown D_k^{pr} - \mathbf{\Sigma}\right)\right)}, \tag{8}$$

where $\mathbf{\Sigma}$ is a threshold matrix with all elements being σ. δ is the parameter to ensure that the value of S_k is approximately equal to 1 when $\triangledown D_k^{pr}(p) > \sigma$, otherwise 0.

Multi-branch Architecture. Note that in above section, only a pair of patches are sampled for generating the subregion. In principle, we hope that the full density map could be leveraged to provide more information. Instead of only sampling a small-large pair of patches, which may involve large bias error due to the large difference between two patches, we adopt a multi-branch architecture as shown in Figure 3. The bottom right corners of all patches are located at the same position, i.e., the bottom right corner of the density map. The area of patch is gradually enlarged along with the branches, until it reaches the size of full density map. Such design guarantees both the small bias error in each branch and the full utilization of training images.

To eliminate the influence of the detected subregion S_k for better optimization in latter branches, we imitate the salience region detection [50] to erase the density values within S_k in next branch, which is formulated as

$$D_{k+1}^{pr} = D_{k+1}^{pr} \odot (\mathbf{1} - S_k), \tag{9}$$

where $\mathbf{1}$ is the matrix with all elements being 1, and $\odot$ is the element-wise product.

Maximum Excess over Pixels (MEP) loss. In the end, K subregions $(S_1, S_2, ..., S_K)$ are generated by the K branches. The final desired pixel-level subregion S is computed by simply combining them together as

$$S = \sum_{k=1}^{K} \{S_k\}, \tag{10}$$

where $\sum$ indicates merging pixels with values close to 1 in all subregion masks $\{S_k\}$, rather than the direct summation. In practice, we take the maximum value at each pixel position from all masks. The final output S is the mask that indicates the pixels which should be optimized. Based on that, our proposed MEP loss is then given by

$$L_{mep}\left(D^{pr}, D^{gt}\right) = \frac{1}{N}\sum_{i=1}^{N}\left|\sum_{p\in S} D^{pr,i}(p) - \sum_{p\in S} D^{gt,i}(p)\right|. \tag{11}$$

3.3. Model Learning

Our SPANet could be easily integrated into existing crowd counting methods, which is equivalent to adding a pooling layer with different masks on the final convolutional layer. It is trained by sequentially optimizing the K times ranking loss, MEP loss, and the original loss of existing methods. When calculating the original loss, the mask pooling layer is removed. The overall training objective is formulated as

$$L_{global} = \sum_{k=1}^{K} L_r + L_{mep} + L_{vanilla}, \tag{12}$$

where $L_{vanilla}$ refers to the original loss of existing approach. In most cases, $L_{vanilla}$ is the L_2 loss. More details of the ground truth generation and data augmentation are described in supplementary material.

4. Experiment

4.1. Experiment Settings

Networks. We evaluate our method by combining it with three networks, i.e., MCNN [52], CSRNet [17], and SANet [3]. The implementations of MCNN[1] and CSRNet[2] are from others, while SANet is implemented by us. In general, there are four main differences between them: (1) Different size of networks. Specifically, MCNN, SANet, and CSRNet are corresponding to small, medium, and large crowd counting networks. (2) Different architectures. MCNN and SANet are multi-column/multi-blob networks, while CSRNet is a single column network. In addition, SANet uses the Instance Normalization (IN) layer and the deconvolutional layer, while CSRNet utilizes the dilated convolutional layer. (3) Different size of density maps. Density maps of MCNN and CSRNet are 1/4 and

[1] https://github.com/svishwa/crowdcount-mcnn
[2] https://github.com/leeyeehoo/CSRNet-pytorch/tree/master

Table 1: Performance comparison with the state-of-the-art methods on ShanghaiTech [52], UCF_CC_50 [11], and UCSD [48] datasets.

			ShanghaiTech A		ShanghaiTech B		UCF_CC_50		UCSD	
Method	**Venue & Year**		**MAE ↓**	**MSE ↓**	**MAE ↓**	**MSE ↓**	**MAE ↓**	**MSE ↓**	**MAE ↓**	**MSE ↓**
Idrees et al. [11]	CVPR	2013	-	-	-	-	419.5	541.6	-	-
Zhang et al. [48]	CVPR	2015	181.8	277.7	32.0	49.8	467.0	498.5	1.60	3.31
CCNN [25]	ECCV	2016	-	-	-	-	-	-	1.51	-
Hydra-2s [25]	ECCV	2016	-	-	-	-	333.7	425.3	-	-
C-MTL [38]	AVSS	2017	101.3	152.4	20.0	31.1	322.8	397.9	-	-
SwitchCNN [32]	CVPR	2017	90.4	135.0	21.6	33.4	318.1	439.2	1.62	2.10
CP-CNN [39]	ICCV	2017	73.6	106.4	20.1	30.1	295.8	320.9	-	-
Huang at al. [10]	TIP	2018	-	-	20.2	35.6	409.5	563.7	**1.00**	1.40
SaCNN [49]	WACV	2018	86.8	139.2	16.2	25.8	314.9	424.8	-	-
ACSCP [34]	CVPR	2018	75.7	102.7	17.2	27.4	291.0	404.6	-	-
IG-CNN [31]	CVPR	2018	72.5	118.2	13.6	21.1	291.4	349.4	-	-
Deep-NCL [36]	CVPR	2018	73.5	112.3	18.7	26.0	288.4	404.7	-	-
MCNN [52]	CVPR	2016	110.2	173.2	26.4	41.3	377.6	509.1	1.07	1.35
CSRNet [17]	CVPR	2018	68.2	115.0	10.6	16.0	266.1	397.5	1.16	1.47
SANet [3]	ECCV	2018	67.0	104.5	8.4	13.6	258.4	334.9	1.02	1.29
MCNN+SPANet	-	-	99.7	146.3	19.1	28.7	292.5	401.3	**1.00**	1.33
CSRNet+SPANet	-	-	62.4	99.5	8.4	13.2	245.8	333.1	1.12	1.42
SANet+SPANet	-	-	**59.4**	**92.5**	**6.5**	**9.9**	**232.6**	**311.7**	**1.00**	**1.28**

1/8 of original images, while SANet produces density maps with the same size as input images. (4) Different testing scheme. SANet is tested on image patches, while CSRNet and MCNN are tested on the whole images.

Learning settings. For MCNN and SANet, the parameters are randomly initialized by a Gaussian distribution with mean of 0 and standard deviation of 0.01. Adam optimizer [14] with a learning rate of $1\mathrm{e}{-5}$ is used to train the model. For CSRNet, the first ten convolutional layers are from pre-trained VGG-16 [37]. The other layers are initialized in the same way as MCNN. Stochastic gradient descent (SGD) with a fixed learning rate of $1\mathrm{e}{-6}$ is applied during the training.

Datasets. We evaluate our method on four datasets, including ShanghaiTech [52], UCF_CC_50 [11], WorldExpo'10 [48], and UCSD [4]. Typically, ShanghaiTech Part A is congested and noisy, while ShanghaiTech Part B is noisy but not highly congested. UCF_CC_50 consists of extremely congested scenes with heavy background noises. WorldExpo'10 and UCSD contain sparse crowd scenes. The scenes in WorldExpo'10 are noisier than UCSD.

Evaluation details. MCNN and CSRNet are tested on the whole images, while SANet is tested on image patches. Following previous works [17, 27, 52], Mean Absolute Error (MAE) and Mean Square Error (MSE) are used to evaluate the performance by

$$MAE = \frac{1}{N}\sum_{i=1}^{N}\left|C_i - C_i^{gt}\right|, \quad MSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(C_i - C_i^{gt}\right)^2}, \tag{13}$$

where C_i is the estimated crowd count and C_i^{gt} is the ground truth count of the i-th image. N is the number of test images. Additionally, PSNR (Peak Signal-to-Noise Ratio)[3] and SSIM (Structural Similarity)[4] [44] are utilized to measure the quality of density maps. For fair comparison, similar to [17], bilinear interpolation is employed to resize estimated density maps to the same size as input images.

[3]https://en.wikipedia.org/wiki/Peak_signal-to-noise_ratio

[4]https://en.wikipedia.org/wiki/Structural_similarity

4.2. Comparisons with State-of-the-art

Table 1 and 2 report the results of four challenging datasets. As a summary, our method significantly improves all baselines and outperforms the other state-of-the-art methods. This result fully demonstrates the effectiveness of our SPANet, which could provide accurate density estimation on both dense and sparse crowd scenes, and can be applied to all CNN-based crowd counting networks.

On ShanghaiTech dataset, our SPANet boosts MCNN, CSRNet, SANet with relative MAE improvements of 9.5%, 8.5%, 11.3% on Part A, and 27.7%, 20.8%, 22.7% on Part B. Noted that Part A is collected from the internet while Part B is from the busy streets and has more spatial constraints. Since our SPANet can fully utilize spatial awareness, it brings more improvements on Part B. On UCF_CC_50, SPANet provides the relative MAE improvements of 22.5%, 7.6%, 10.0% for the three baselines. Noted that the improved MCNN is even comparable with other state-of-the-art methods. It clearly shows that SPANet can handle the extremely dense-crowd scenes. Similar to the above two datasets, SPANet also achieves significant improvements on UCSD and WorldExpo'10, verifying the effectiveness of our method on the sparse-crowd scenes.

4.3. Ablation Studies

Sampling positions. We first evaluate the impact of different starting positions when sampling patches for mask pooling. The results are listed in Table 3. We find that starting at the bottom is always better than the top, and the right is also better than the left. The possible reason is that it may be closely related to camera calibration. The results en-

Table 2: Comparison with the state-of-the-art methods on WorldExpo'10 [4] dataset. Only MAE is computed for each scene and then averaged to evaluate the overall performance.

Method	S1	S2	S3	S4	S5	Avg.
Zhang et al. [48]	9.8	14.1	14.3	22.2	3.7	12.9
Huang et al. [10]	4.1	21.7	11.9	11.0	3.5	10.5
Switch-CNN [32]	4.4	15.7	10.0	11.0	5.9	9.4
SaCNN [49]	2.6	13.5	10.6	12.5	3.3	8.5
CP-CNN [39]	2.9	14.7	10.5	**10.4**	5.8	8.9
MCNN [52]	3.4	20.6	12.9	13.0	8.1	11.6
CSRNet [17]	2.9	11.5	8.6	16.6	3.4	8.6
SANet [3]	2.6	13.2	9.0	13.3	**3.0**	8.2
MCNN+SPANet	3.4	14.9	15.1	12.8	4.5	10.1
CSRNet+SPANet	2.6	**11.1**	8.9	13.5	3.3	7.9
SANet+SPANet	**2.3**	12.3	**7.9**	12.9	3.2	**7.7**

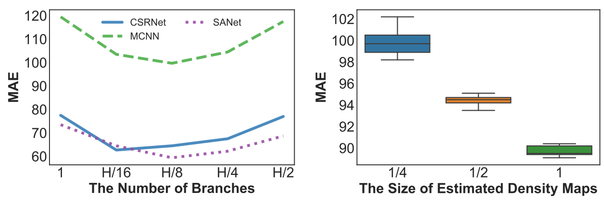

Figure 4: Ablation studies on ShanghaiTech Part A [52]. The left shows the branch number K vs. MAE, and the right illustrates the size of estimated density maps vs. MAE, performed with MCNN.

Table 3: Ablation studies of patch sampling strategy, mask pooling strategy, and losses on ShanghaiTech Part A dataset [52].

Configurations	MAE ↓	MSE ↓
Center point	101.2	153.3
Top left corner	101.5	153.7
Bottom left corner	100.7	149.2
Top right corner	100.5	149.4
Bottom right corner	99.7	146.3
Different density map	100.3	147.4
Same density map	**99.7**	**146.3**
L_2	110.2	173.2
L_r+L_{mep}	99.3	145.3
$L_2 + L_r$	107.2	164.5
$L_2 + L_r + L_{mep}$	99.7	146.3
Random	105.4	162.2
Grid Search	**98.3**	**142.5**

courage us to sample patches from the bottom right corner. Noted that the differences between these sampling schemes are quite small, which demonstrates the robustness of our method. Additionally, we also present the comparison of performing mask pooling on the same or different density maps in each branch, which is already discussed in Section 3.2 and Eq. (4). As shown in Table 3, the results of two strategies are similar. Due to the efficiency problem, we directly pool patches from the same density map.

Different losses/weights. We turn to evaluate the effect of different losses and weight schemes. As shown in Table 3, adding the ranking loss only provides slight improvement, while the significant improvement comes from the MEP loss. Besides, there is no significant difference whether L_2 is used. It demonstrates that our MEP loss can effectively learn spatial awareness to boost crowd counting. We further conduct experiments on two weight schemes: the random weight and the grid search with step 0.1. As shown in Table 3, our method is not sensitive to the weights. Even the grid search brings a very slight improvement.

Number of branches. We measure the performance of SPANet with different branch numbers K. As illustrated in Figure 4, the performance first improves but then drops with the increasing number of K. This observation is not surprising. On one side, small K (e.g., $K = 1$) would involve large bias error due to the large difference between two patches. On the other side, large K (e.g., $K = \frac{H}{2}$, where H is the height of estimated density map) implies that the difference of two patches in each branch is very small, which cannot provide enough discrepancy for subregion generation. In experiments, K is set to $\frac{H}{8}$ for MCNN/SANet and $\frac{H}{16}$ for CSRNet, which is determined via cross validation.

Size of estimated density maps. We further validate the effect of the size of estimated density maps. We add deconvolutional layers on top of the MCNN to increase the size of the estimated density maps. Eventually, two variants of MCNN are obtained, whose estimated density maps are of $1/2$ and the same size as the input images, respectively. As shown in Figure 4, the performance is improved along with the size increase of density maps. The results indicate that predicting high-resolution density maps could bring considerable improvement.

4.4. Studies on Estimated Density Maps

We now evaluate the estimated density maps to verify whether our method can fully utilize spatial awareness. Table 4 summarizes the results. Our SPANet can significantly improve PSNR and SSIM across all baselines and datasets, which indicates that the quality of the generated density maps are significantly improved. To further verify that our method can indeed learn spatial awareness, we showcase the generated density maps of four examples from different methods in Figure 5. These four examples typically contain different crowd densities, occlusions, and scale changes. We can observe that the baseline models are always affected by the zero-mean noise, which leads to overestimation in low-density areas. In contrast, zero-mean noise is effectively suppressed in our SPANet. Besides, baseline models normally have an insufficient estimation for high-density areas, while ours can obtain a more accurate estimation for them. Noted that the ground truth itself is also generated with center points of pedestrians' heads, which inherently contains inaccurate information. It means that our method is still unable to produce the same density map to the ground truth.

4.5. Studies on Learning Curves

Finally, we study the learning curves to further evaluate our method. Figure 6 shows the training and valida-

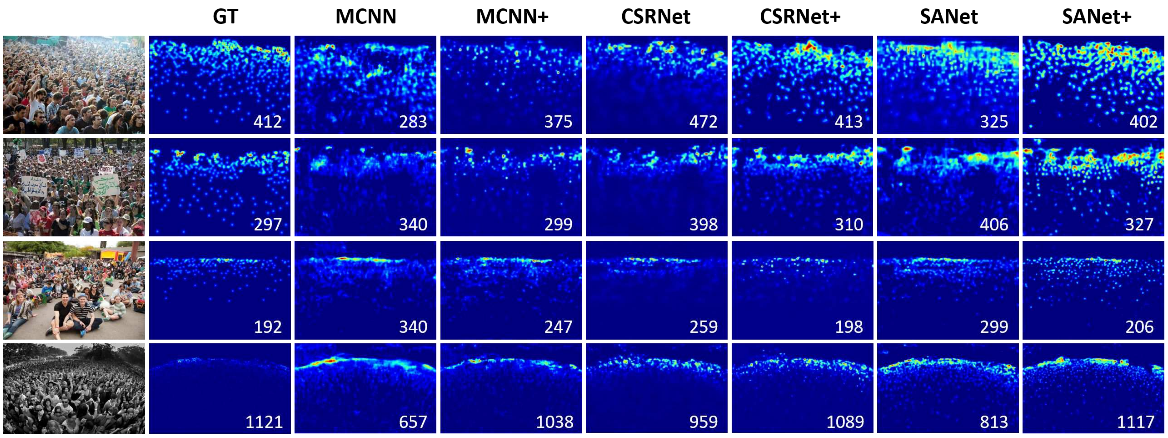

Figure 5: Comparisons of estimated density maps between baselines and our SPANet. '+' indicates combining SPANet with baselines.

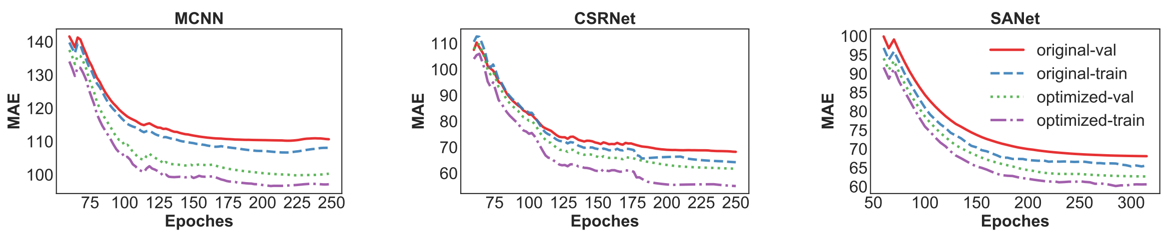

Figure 6: Learning Curves. Mean absolute error (MAE) on training and validation sets, vs. the number of training epochs of MCNN [52], CSRNet [17] and SANet [3] on ShanghaiTech Part A dataset [52].

Table 4: Density map quality comparison. Values on the left of '|' are from original baselines, while values on the right of '|' are results when integrating with the proposed SPANet.

	MCNN		CSRNet		SANet	
Dataset	**PSNR ↑**	**SSIM ↑**	**PSNR ↑**	**SSIM ↑**	**PSNR ↑**	**SSIM ↑**
ShanghaiTech-A [52]	21.42 | 22.18	0.52 | 0.66	23.79 | 24.88	0.76 | 0.85	23.36 | 25.33	0.78 | 0.85
ShanghaiTech-B [52]	23.43 | 26.19	0.78 | 0.85	27.02 | 29.50	0.89 | 0.92	27.44 | 29.17	0.89 | 0.91
UCF_CC_50 [11]	14.44 | 18.25	0.37 | 0.51	18.76 | 20.17	0.52 | 0.78	18.35 | 20.01	0.51 | 0.76
UCSD [48]	17.43 | 18.52	0.75 | 0.83	20.02 | 21.80	0.86 | 0.89	21.33 | 22.20	0.84 | 0.90
WorldExpo'10 [4]	23.53 | 25.97	0.76 | 0.85	26.94 | 29.05	0.92 | 0.93	26.22 | 28.54	0.90 | 0.92

tion mean absolute error (MAE) at every epoch on ShanghaiTech Part A dataset. For better viewing, we smooth the learning curves by exponential moving average (EMA) with a smoothing factor $\alpha = 0.1$. Compared with original results, baselines integrated with our SPANet exhibit lower MAE on both training and testing set. Since the performance on the training and testing set generally denotes the fitting and generalization degree, this result demonstrates the promising capability on both sides. In addition, it also means that our method can significantly improve the stability during model training.

5. Conclusion

In this paper we present a novel deep architecture called SPatial Awareness Network (SPANet) for crowd counting, which is able to capture the spatial variations by finding the pixel-level subregion with high discrepancy to the ground truth. It could be integrated into all CNN-based methods and is end-to-end trainable. Experiments on four datasets and three various networks fully demonstrate that it can significantly improve all baselines and outperforms the state-of-the-art methods. It provides the elegant views of effectively using spatial awareness to improve crowd counting. In future work we will study how to preserve spatial awareness as much as possible in the ground truth generation.

Acknowledgements

This research was supported in part through the financial assistance award 60NANB17D156 from U.S. Department of Commerce, National Institute of Standards and Technology and by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior/Interior Business Center (DOI/IBC) contract number D17PC00340, National Natural Science Foundation of China (61772436), Foundation for Department of Transportation of Henan Province, China (2019J-2-2), Sichuan Science and Technology Innovation Seedling Fund (2017RZ0015), China Scholarship Council (201707000083) and Cultivation Program for the Excellent Doctoral Dissertation of Southwest Jiaotong University (D-YB 201707).

References

[1] Lokesh Boominathan, Srinivas S. S. Kruthiventi, and R. Venkatesh Babu. Crowdnet: A deep convolutional network for dense crowd counting. In *Proceedings of ACM International Conference on Multimedia*, pages 640–644, 2016. 2

[2] Gabriel J Brostow and Roberto Cipolla. Unsupervised bayesian detection of independent motion in crowds. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, pages 594–601, 2006. 2

[3] Xinkun Cao, Zhipeng Wang, Yanyun Zhao, and Fei Su. Scale aggregation network for accurate and efficient crowd counting. In *Proceedings of European Conference on Computer Vision*, pages 757–773, 2018. 1, 2, 3, 5, 6, 7, 8

[4] Antoni B Chan, Zhang-Sheng John Liang, and Nuno Vasconcelos. Privacy preserving crowd monitoring: Counting people without people models or tracking. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–7, 2008. 1, 2, 6, 7, 8

[5] Antoni B Chan and Nuno Vasconcelos. Counting people with low-level features and bayesian regression. *IEEE Transactions on Image Processing*, 21(4):2160–2177, 2012. 2

[6] Zhi-Qi Cheng, Jun-Xiu Li, Qi Dai, Xiao Wu, Jun-Yan He, and Alexander Hauptmann. Improving the learning of multi-column convolutional neural network for crowd counting. In *Proceedings of the 26th ACM International Conference on Multimedia*, 2019. 1

[7] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, pages 886–893, 2005. 2

[8] Piotr Dollár, Boris Babenko, Serge Belongie, Pietro Perona, and Zhuowen Tu. Multiple component learning for object detection. In *Proceedings of European Conference on Computer Vision*, pages 211–224, 2008. 2

[9] Siyu Huang, Xi Li, Zhiqi Cheng, Zhongfei Zhang, and Alexander G. Hauptmann. Stacked pooling: Improving crowd counting by boosting scale invariance. *CoRR*, abs/1808.07456, 2018. 1, 3

[10] Siyu Huang, Xi Li, Zhongfei Zhang, Fei Wu, Shenghua Gao, Rongrong Ji, and Junwei Han. Body structure aware deep crowd counting. *IEEE Transactions on Image Processing*, 27(3):1049–1059, 2018. 6, 7

[11] Haroon Idrees, Imran Saleemi, Cody Seibert, and Mubarak Shah. Multi-source multi-scale counting in extremely dense crowd images. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 2547–2554, 2013. 1, 2, 6, 8

[12] Haroon Idrees, Khurram Soomro, and Mubarak Shah. Detecting humans in dense crowds using locally-consistent scale prior and global occlusion reasoning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(10):1986–1998, 2015. 2

[13] Di Kang and Antoni B. Chan. Crowd counting by adaptively fusing predictions from an image pyramid. In *Proceedings of British Machine Vision Conference*, page 89, 2018. 1, 2

[14] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[15] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 4681–4690, 2017. 2

[16] Victor S. Lempitsky and Andrew Zisserman. Learning to count objects in images. In *Proceedings of Conference on Neural Information Processing Systems*, pages 1324–1332, 2010. 1, 2, 3

[17] Yuhong Li, Xiaofan Zhang, and Deming Chen. Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 1091–1100, 2018. 1, 3, 5, 6, 7, 8

[18] 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. 2

[19] Jiang Liu, Chenqiang Gao, Deyu Meng, and Alexander G. Hauptmann. Decidenet: Counting varying density crowds through attention guided detection and density estimation. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 5197–5206, 2018. 1, 2

[20] Ning Liu, Yongchao Long, Changqing Zou, Qun Niu, Li Pan, and Hefeng Wu. Adcrowdnet: An attention-injective deformable convolutional network for crowd understanding. *CoRR*, abs/1811.11968, 2018. 1, 3

[21] Weizhe Liu, Krzysztof Lis, Mathieu Salzmann, and Pascal Fua. Geometric and physical constraints for head plane crowd density estimation in videos. *CoRR*, abs/1803.08805, 2018. 3

[22] Weizhe Liu, Mathieu Salzmann, and Pascal Fua. Context-aware crowd counting. *CoRR*, abs/1811.10452, 2018. 2

[23] Xialei Liu, Joost van de Weijer, and Andrew D. Bagdanov. Leveraging unlabeled data for crowd counting by learning to rank. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 7661–7669, 2018. 2, 4

[24] Frank J Massey Jr. The kolmogorov-smirnov test for goodness of fit. *Journal of the American statistical Association*, 46(253):68–78, 1951. 3

[25] Daniel Oñoro-Rubio and Roberto Javier López-Sastre. Towards perspective-free object counting with deep learning. In *Proceedings of European Conference on Computer Vision*, pages 615–629, 2016. 6

[26] 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 *Proceedings of International Conference on Computer Vision*, pages 3253–3261, 2015. 2

[27] Viresh Ranjan, Hieu Le, and Minh Hoai. Iterative crowd counting. In *Proceedings of European Conference on Computer Vision*, pages 278–293, 2018. 2, 3, 6

[28] Carlo S Regazzoni and Alessandra Tesei. Distributed data fusion for real-time crowding estimation. *Signal Processing*, 53(1):47–63, 1996. 2

[29] David Ryan, Simon Denman, Clinton Fookes, and Sridha Sridharan. Crowd counting using multiple local features. In *Digital Image Computing: Techniques and Applications*, pages 81–88, 2009. 2

[30] Deepak Babu Sam and R. Venkatesh Babu. Top-down feedback for crowd counting convolutional neural network. In *Proceedings of Conference on Artificial Intelligence*, pages 7323–7330, 2018. 3

[31] 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 *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 3618–3626, 2018. 3, 6

[32] Deepak Babu Sam, Shiv Surya, and R. Venkatesh Babu. Switching convolutional neural network for crowd counting. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 4031–4039, 2017. 2, 6, 7

[33] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 618–626, 2017. 2

[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 IEEE Conference on Computer Vision and Pattern Recognition*, pages 5245–5254, 2018. 6

[35] Miaojing Shi, Zhaohui Yang, Chao Xu, and Qijun Chen. Perspective-aware CNN for crowd counting. *CoRR*, abs/1807.01989, 2018. 3

[36] 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 IEEE Conference on Computer Vision and Pattern Recognition*, pages 5382–5390, 2018. 6

[37] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 6

[38] Vishwanath A. Sindagi and Vishal M. Patel. Cnn-based cascaded multi-task learning of high-level prior and density estimation for crowd counting. In *Proceedings of International Conference on Advanced Video and Signal Based Surveillance*, pages 1–6, 2017. 1, 2, 6

[39] Vishwanath A. Sindagi and Vishal M. Patel. Generating high-quality crowd density maps using contextual pyramid cnns. In *Proceedings of International Conference on Computer Vision*, pages 1879–1888, 2017. 1, 2, 3, 6, 7

[40] Yukun Tian, Yimei Lei, Junping Zhang, and James Z. Wang. Padnet: Pan-density crowd counting. *CoRR*, abs/1811.02805, 2018. 2

[41] Paul Viola, Michael J Jones, and Daniel Snow. Detecting pedestrians using patterns of motion and appearance. *International Journal of Computer Vision*, 63(2):153–161, 2005. 2

[42] Elad Walach and Lior Wolf. Learning to count with CNN boosting. In *Proceedings of European Conference on Computer Vision*, pages 660–676, 2016. 1, 2

[43] Meng Wang and Xiaogang Wang. Automatic adaptation of a generic pedestrian detector to a specific traffic scene. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 3401–3408, 2011. 2

[44] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. 6

[45] Ze Wang, Zehao Xiao, Kai Xie, Qiang Qiu, Xiantong Zhen, and Xianbin Cao. In defense of single-column networks for crowd counting. In *Proceedings of British Machine Vision Conference*, page 78, 2018. 1, 2, 3

[46] Xingjiao Wu, Yingbin Zheng, Hao Ye, Wenxin Hu, Jing Yang, and Liang He. Adaptive scenario discovery for crowd counting. *CoRR*, abs/1812.02393, 2018. 2

[47] Lingke Zeng, Xiangmin Xu, Bolun Cai, Suo Qiu, and Tong Zhang. Multi-scale convolutional neural networks for crowd counting. In *Proceedings of International Conference on Image Processing*, pages 465–469, 2017. 1, 2

[48] Cong Zhang, Hongsheng Li, Xiaogang Wang, and Xiaokang Yang. Cross-scene crowd counting via deep convolutional neural networks. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 833–841, 2015. 1, 6, 7, 8

[49] Lu Zhang, Miaojing Shi, and Qiaobo Chen. Crowd counting via scale-adaptive convolutional neural network. In *Proceedings of Winter Conference on Applications of Computer Vision*, pages 1113–1121, 2018. 3, 6, 7

[50] Xiaolin Zhang, Yunchao Wei, Jiashi Feng, Yi Yang, and Thomas S Huang. Adversarial complementary learning for weakly supervised object localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1325–1334, 2018. 2, 4, 5

[51] Youmei Zhang, Chunluan Zhou, Faliang Chang, and Alex C. Kot. Attention to head locations for crowd counting. *CoRR*, abs/1806.10287, 2018. 3

[52] Yingying Zhang, Desen Zhou, Siqin Chen, Shenghua Gao, and Yi Ma. Single-image crowd counting via multi-column convolutional neural network. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 589–597, 2016. 1, 2, 3, 5, 6, 7, 8

[53] Tao Zhao, Ram Nevatia, and Bo Wu. Segmentation and tracking of multiple humans in crowded environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(7):1198–1211, 2008. 2

[54] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2921–2929, 2016. 2, 4