Attention Scaling for Crowd Counting

Xiaoheng Jiang[1], Li Zhang[1], Mingliang Xu[1]*,Tianzhu Zhang[2]
Pei Lv[1], Bing Zhou[1], Xin Yang[3],Yanwei Pang[4]
[1]School of Information Engineering, Zhengzhou University
[2]School of Information Science and Technology, University of Science and Technology of China
[3]School of Computer Science and Technology, Dalian University of Technology
[4]School of Electrical and Information Engineering, Tianjin University
{jiangxiaoheng, iexumingliang, ielvpei, iebzhou}@zzu.edu.cn
laridzhang@gmail.com, tzzhang@ustc.edu.cn, xinyang@dlut.edu.cn, pyw@tju.edu.cn

Abstract

Convolutional Neural Network (CNN) based methods generally take crowd counting as a regression task by outputting crowd densities. They learn the mapping between image contents and crowd density distributions. Though having achieved promising results, these data-driven counting networks are prone to overestimate or underestimate people counts of regions with different density patterns, which degrades the whole count accuracy. To overcome this problem, we propose an approach to alleviate the counting performance differences in different regions. Specifically, our approach consists of two networks named Density Attention Network (DANet) and Attention Scaling Network (ASNet). DANet provides ASNet with attention masks related to regions of different density levels. ASNet first generates density maps and scaling factors and then multiplies them by attention masks to output separate attention-based density maps. These density maps are summed to give the final density map. The attention scaling factors help attenuate the estimation errors in different regions. Furthermore, we present a novel Adaptive Pyramid Loss (APLoss) to hierarchically calculate the estimation losses of subregions, which alleviates the training bias. Extensive experiments on four challenging datasets (ShanghaiTech Part A, UCF_CC_50, UCF-QNRF, and WorldExpo'10) demonstrate the superiority of the proposed approach.

1. Introduction

The computer vision based crowd counting task is to infer the number of people presented in images or videos. It recently has drawn much attention from researchers because of its great value in a wide range of real-world applications such as video surveillance, public safety, traffic control, agriculture monitoring, and cell counting.

The solution to this problem has progressively advanced from detecting individuals to presenting crowd density distributions. The integration of density maps gives the total count. Though previous methods have achieved some success, they fail to handle highly congested crowd scenes. These scenes usually exhibit the properties of heavy occlusions, large scale variation, perspective changes, and so on. Inspired by the great success that the convolutional neural networks have made in computer vision tasks like object detection [28, 3, 27], image segmentation [4, 8], and object tracking [48, 47], it recently springs dozens of CNN based crowd counting methods [37, 22, 18, 41, 10]. These methods have tried to exploit multi-scale feature fusion [49, 29, 35], multi-task learning [34, 30, 21], and the attention mechanism [45, 44, 19] to solve the above questions. Even so, there is still much room for improvement in counting performance, especially in several challenging crowd datasets [49, 10, 11].

People in images or across scenes usually exhibit various distributions, with some regions overcrowded and other regions sparsely filled. Two main factors lead to this phenomenon. On the one hand, people scatter or gather together spontaneously in different regions of the scenes. On the other hand, people's scale varies due to the change of camera perspective. Accordingly, the people distributions in density maps present different patterns. Since CNNs depend heavily on the dataset during the training procedure, the learned data-driven counting networks are prone to be affected by different people distributions. As a result, they perform inconsistently in regions with different people distributions. It is observed that the predictions in high-density areas are likely to be higher than the ground truth, while the predictions in low-density areas are likely to be lower than

*The corresponding author is Mingliang Xu.

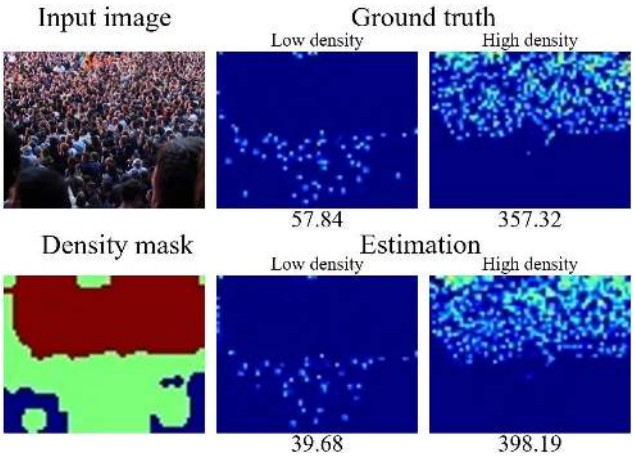

Figure 1. One example of the crowd density estimation results on the ShanghaiTech Part A dataset. The red, green, blue colors in the density mask map represent the high-density, low-density, and background regions, respectively. Compared with the ground truth, the counting network predicts a higher count in the high-density region and a lower count in the low-density region.

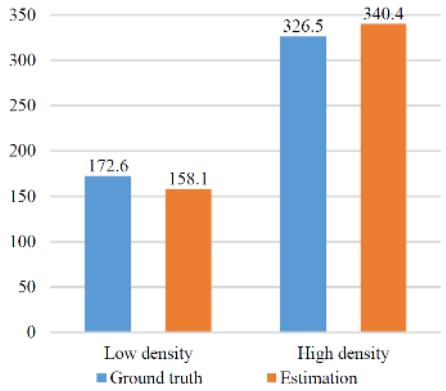

Figure 2. The comparisons between the average ground truth and the average estimations in low-density and high-density regions on the test set of the ShanghaiTech Part A dataset. The CNN based baseline counting network is prone to overestimate people count in the high-density region and underestimate people count in the low-density region.

the ground truth, as demonstrated in Figure 1 and Figure 2.

In this paper, we aim to present an approach that can handle the congested scenes with various density distributions. To this end, we construct an attention scaling convolutional neural network named ASNet. ASNet first generates scaling factors to adjust the corresponding intermediate density maps. Then ASNet outputs several attention based density maps with each only focusing on the region of one certain density level. Finally, ASNet sums these attention based density maps to give the final density map. To provide attention masks for ASNet, we present a density attention network named DANet that performs the task of pixel-wise density segmentation.

Furthermore, we present a novel loss function named Adaptive Pyramid Loss (APLoss). APLoss first divides the density map into non-uniform pyramidal sub-regions adaptively based on local people counts and then calculates each local normalized loss. Finally, APLoss accumulates all local losses to give the final estimation loss. APLoss alleviates the training bias and improves the generalization ability of the counting network. The contributions of this paper are summarized as follows:

(1) We propose a novel attention scaling convolutional neural network (ASNet) that learns scaling factors to automatically adjust the density estimation of each corresponding sub-region, which reduces the local estimation error.

(2) We propose a density attention network (DANet) that provides ASNet with attention masks concerning regions of different density levels.

(3) We propose a novel adaptive pyramid loss (APLoss) that can ease the training bias and strengthen the generalization ability of the counting network.

(4) Compared with other sixteen newly reported state-of-the-art results, our proposed approach demonstrates its superiority on four challenging crowd datasets.

The rest of the paper is organized as follows. First, we review the previous crowd counting methods in Section 2. Then, we present the proposed approach in Section 3. After that, we demonstrate the experimental results and analysis in Section 4. Finally, we conclude this paper in Section 5.

2. Related Work

Recent years have witnessed progressive improvement in crowd counting from traditional methods [7, 10, 26, 39, 2] to CNN based methods [1, 24, 13, 14, 23, 29, 25]. In this section, we mainly review three kinds of common CNN based counting strategies.

2.1. Multi-scale Information Fusion Approaches

This kind of approach aims at exploiting multi-scale features or multi-context information to deal with the people scale variation problem. Multi-column convolutional neural network (MCNN), proposed by Zhang *et al.* [49], utilizes multi-size filters to extract features that have receptive fields of different sizes. Similarly, Sam *et al.* [29] proposed Switch-CNN that utilizes a switch classifier to choose the optimal one from the density generator pool. Sindagi *et al.* [35] proposed Contextual Pyramid CNN (CP-CNN) to capture multi-scale information by combining global and local context priors. Further, Sindagi and Patel [37] presented a multi-level bottom-top and top-bottom fusion network (MBTTBF) that is elaborately designed to combine multiple shallow and deep features. Chen *et al.* [5] proposed a Scale Pyramid Network (SPN) that parallelly utilizes dilated convolutions of different rates in a shared single-

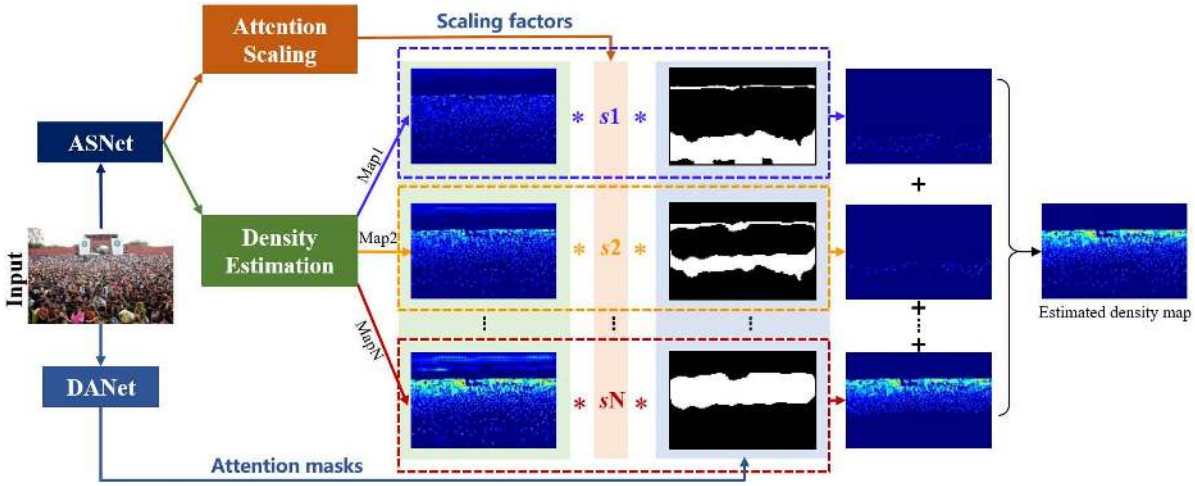

Figure 3. The architecture of the proposed approach. Density Attention Network (DANet) provides attention masks for the Attention Scaling Network (ASNet). ASNet has two branches. The Density Estimation branch generates intermediate density maps and the Attention Scaling branch generates scaling factors. ASNet multiplies intermediate density maps and scaling factors by attention masks to generate attention based density maps, which are then summed to give the final density map.

column CNN to extract multi-scale features.

2.2. Attention Guided Approaches

This kind of approach utilizes the visual attention mechanism to make the counting network intentionally focus on useful information to improve counting performance. Zhang *et al.* [45] proposed the Attentional Neural Field (ANF) that incorporates conditional random fields and non-local attention mechanisms to capture multi-scale features and long-range dependencies, strengthening the network's ability to handle large scale variation. Further, Zhang *et al.* [44] proposed a Relational Attention Network (RANet) that utilizes both local self-attention and global self-attention mechanisms to capture the interdependence information of pixels, obtaining more informative feature representations. The attention-injective deformable network (ADCrowdNet), proposed by Liu *et al.* [19], utilizes an attention map generator to provide regions and congestion degrees for the latter density map estimator. Liu *et al.* [17] proposed Recurrent Attentive Zooming Network (RAZN) that iteratively locates regions with high ambiguity and re-evaluates them in high-resolution space.

2.3. Multi-task Learning Approaches

This kind of approach leverages auxiliary tasks to improve counting performance. Sindagi *et al.* [34] proposed to utilize one extra crowd count task to provide high-level priors for the density estimation task. The two jointly learned tasks enable the shared part of the network to learn more discriminative features. Shen *et al.* [30] presented the Adversarial Cross-Scale Consistency Pursuit (ACSCP) framework by exploiting the collaboration between adversarial learning and density estimation. Liu *et al.* [21] proposed to incorporate self-supervised image ranking and density estimation into a multi-task learning framework, which makes it possible to learn from an abundant unlabeled crowd dataset. Zhao *et al.* [50] proposed to formulate several heterogeneous attributes including geometric, semantic, and numeric information as auxiliary tasks to assist the counting task, which helps generate more robust features to handle scale variation and cluttered background.

3. Our Approach

The architecture of the proposed method is illustrated in Figure 3. It consists of two convolutional networks: Density Attention Network (DANet) and Attention Scaling Network (ASNet). DANet provides ASNet with attention masks concerning regions of different density levels. ASNet has two branches with Density Estimation generating intermediate density maps and Attention Scaling generating scaling factors. ASNet multiples them by attention masks to output density maps that are summed to give the final density map. In this section, we first present DANet and ASNet and then introduce the novel Adaptive Pyramid Loss (APLoss).

3.1. Density Attention Network

DANet aims to generate attention masks that represent regions of different density levels. It achieves this goal by performing a pixel-wise density segmentation task. That is, DANet classifies each pixel to one certain density level. The pixels of the same density level form the region of one attention mask.

It generally generates the ground-truth density map by utilizing a Gaussian kernel to blur each head annotation. The sum of the Gaussian kernel equals to one. Therefore,

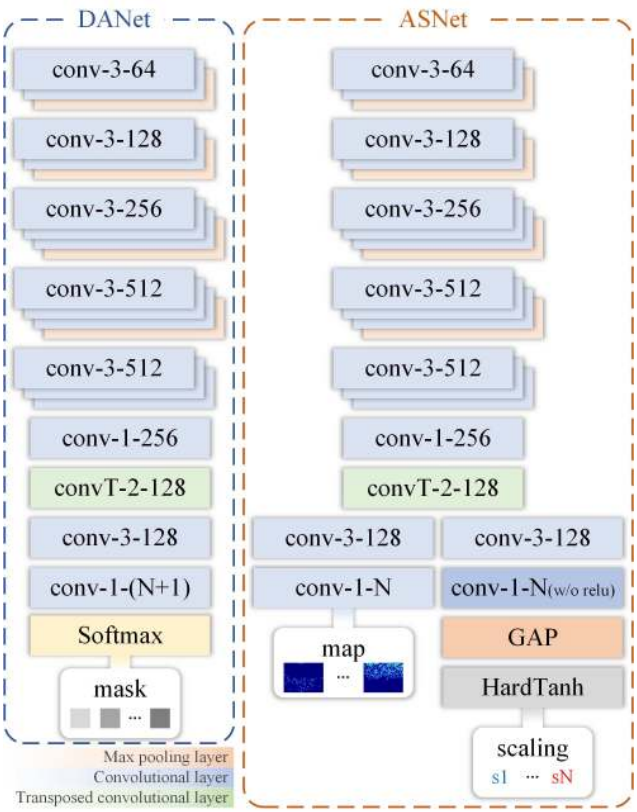

Figure 4. The configurations of the density attention network (DANet) and the attention scaling network (ASNet).

the actual value of each pixel in the density map does not represent the density level of one region. Similar to [9], we use the local count centered at one pixel to denote its density level, which makes the density level of pixels consistent with that of the local region. Specifically, we generate the pixel-wise ground-truth density level labels as follows. Firstly, we obtain all the local counts by scanning the ground-truth density maps in the training set pixel by pixel with a 64×64 sliding window. Secondly, we calculate the average value $AvgCnt_{11}$ of all non-zero local counts and find the minimum count $MinCnt$ and maximal count $MaxCnt$. Thirdly, we use $\{MinCnt, AvgCnt_{11}, MaxCnt\}$ as the threshold set to divide the density into two levels : low density and high density. Iteratively, we can calculate the average values $AvgCnt_{21}$ of all low-density counts and $AvgCnt_{22}$ of all high-density counts. And we use the new threshold set $\{MinCnt, AvgCnt_{21}, AvgCnt_{11}, AvgCnt_{22}, MaxCnt\}$ to divide the density into four levels, and so on. Fourthly, we use the obtained threshold set to label each pixel in the ground-truth density map automatically according to its corresponding local count. Given N density levels, there are $N + 1$ density labels including one extra background label.

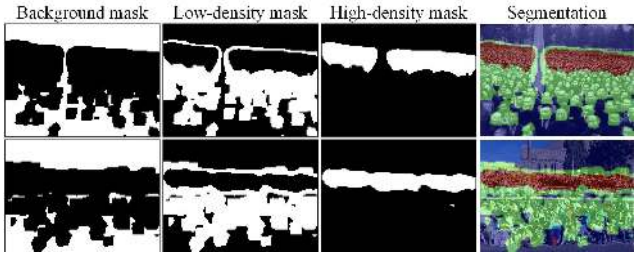

Figure 5. From the left column to the right column, they are background masks, low-density attention masks, high-level density masks, and the density level segmentation results, respectively.

Once we get the ground-truth density level labels, we train the proposed DANet to learn to classify each pixel of the input crowd image to different density levels. That is, DANet can segment the crowd image to regions of different density levels, with each region corresponding to one binary attention mask. Figure 5 shows two examples of attention masks. After obtaining the N foreground attention masks, we use one dilation operation to expand each mask. As a result, there are overlaps between adjacent attention masks. When summing the attention mask based density maps, the density values corresponding to the overlapped mask regions are averaged.

The DANet architecture is presented in the left column of Figure 4. We utilize the first 13 convolutional layers of the trained VGG-16 [33] model as the backbone. We add four new convolutional layers on top of the backbone. The first one has convolutional kernels with a size of 1×1 and has 256 output channels. The second one is a deconvolutional layer that has 2×2 kernels with a stride of 2 pixels. The third one has 3×3 kernels and 128 output channels. The fourth one has 1×1 kernels and $N + 1$ output maps. We train DANet with the two-dimensional softmax cross-entropy loss.

3.2. Attention Scaling Network

As stated in Section 1, CNN based counting networks are prone to overestimate or underestimate local counts in regions of different density levels. To correct the local density estimation, we propose the Attention Scaling Network (ASNet). As demonstrated in Figure 3, ASNet has one Attention Scaling branch (AS-branch) and one Density Estimation branch (DE-branch). DE-branch generates intermediate density maps that are to be corrected. AS-branch learns to generate scaling factors that aims at adjusting the intermediate density maps in conjunction with attention masks provided by DANet. These scaling factors help fine-tune the overall crowd count of the corresponding local regions. This can be considered as a rough estimation strategy used by human beings, which adjusts the predicted count by multiplying a factor without pixel-wise re-calculation. It is noted that we only use the foreground attention masks

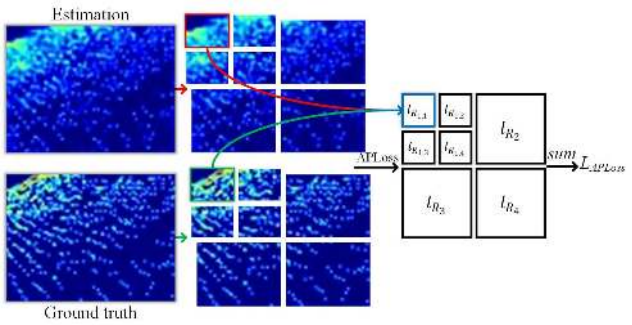

Figure 6. The demonstration of the two-level adaptive pyramid loss (APLoss).

that correspond to crowd regions. ASNet outputs N adjusted density maps by multiplying the intermediate density maps, scaling factors, and the attention masks. As a result, the adjusted density maps concentrate only on regions with the corresponding density masks. They are then summed to generate the final density map.

The configuration of the ASNet is presented in the right column of Figure 4. Similar to DANet, ASNet also uses the first 13 convolutional layers of the trained VGG-16 [33] model as the backbone. ASNet first adds one new convolutional layer and one new deconvolutional layer. On top of these layers, ASNet then splits into the AS-branch and the DE-branch. DE-branch has two new convolutional layers and outputs N intermediate density maps. AS-branch adds two new convolutional layers and the second one outputs N channels with the values un-activated. And then AS-branch utilizes the global average pooling (GAP) operation to transform the obtained N channels into N scalars that are then activated by the HardTanh function. We first set the output range of the HardTanh function to $(-1, 1)$ and then add one to make the output scaling factors be in the range of $(0, 2)$.

3.3. Adaptive Pyramid Loss

During the training stage, previous CNN based density estimation networks usually use the Euclidean distance between the whole estimated and ground-truth density maps as the loss function:

$$L(\Theta) = \frac{1}{M}\sum_{k=1}^{M}||D(X^k;\Theta) - D^k||_2^2, \tag{1}$$

where X^k is the k-th input image, D^k is its ground-truth density map, Θ is the parameters of the counting network, $D(X^k;\Theta)$ is the estimated density map, and M is the size of the training set. This loss ignores the impact of densities of different levels on the network training procedure. Since the low-density and high-density distributions are usually quite unbalanced, the corresponding estimation errors can make the trained counting network biased. This weakens the generalization ability of the counting network. Further, even in the region of the same density level (as described in Section 3.1), there are density differences in its subregions.

To deal with the above problem, we propose a novel loss named Adaptive Pyramid Loss (APLoss). APLoss is able to adaptively divide the density map into non-uniform pyramidal subregions based on the ground-truth local crowd counts. And then APLoss first calculates each local relative estimation loss and then sums them to give the final loss. Figure 6 shows a two-level APLoss calculation.

Specifically, we calculate APLoss as follows. Firstly, we divide the ground-truth density map D^k into a first-level grid of 2×2 and denote the subregion by R_{i_1} with $i_1 \in \{1, 2, 3, 4\}$. If the local crowd count of the subregion R_{i_1} is higher than a given threshold T, we divide it into a second-level sub-grid of 2×2 and denote them by R_{i_1,i_2} with $i_2 \in \{1, 2, 3, 4\}$. We iteratively divide one region into an n-th level sub-grid of 2×2 until its local crowd count is lower than T. We denote the n-th level subregion by $R_{i_1,\cdots,i_n}$ with $i_n \in \{1, 2, 3, 4\}$. After the division is completed, we can get one non-uniform pyramid grid. Secondly, we apply the obtained adaptive pyramid grid on the estimated density map $D(X^k;\Theta)$ and calculate the local loss for each sub-region:

$$l^k_{R_{i_1,\cdots,i_{n-1}}} = \begin{cases} \frac{||D_{R_{i_1,\cdots,i_{n-1}}}(X^k;\Theta) - D^k_{R_{i_1,\cdots,i_{n-1}}}||_2^2}{sum(D^k_{R_{i_1,\cdots,i_{n-1}}})+1}, \\ \qquad\qquad sum(D^k_{R_{i_1,\cdots,i_{n-1}}}) < T \\ \sum_{i_n=1}^{4} l^k_{R_{i_1,\cdots,i_n}}, \qquad otherwise \end{cases} \tag{2}$$

Finally, we aggregate all the local losses to give the final APLoss:

$$L_{APLoss} = \frac{1}{M}\sum_{k=1}^{M}\sum_{i_1=1}^{4} l^k_{R_{i_1}}. \tag{3}$$

4. Experiments

We validate the effectiveness of the proposed method on four challenging crowd datasets. The performance of the current counting networks still has a lot of room for improvement on three of the datasets including the ShanghaiTech Part A dataset [49], UCF_CC_50 dataset [10], and UCF-QNRF dataset [11]. And the WorldExpo'10 [46] dataset provides cross-scene test sets that can test the adaptive capacity of the network for different scenes.

4.1. Datasets

ShanghaiTech Part A dataset [49]. This crowd dataset contains 482 images that are randomly crawled from the Internet and are divided into the training and test sets. There

are 300 images and 182 images in the training and test sets, respectively.

UCF_CCF_50 dataset [10]. This dataset shows a lot of challenges. It randomly collects only 50 images from the Internet. The number of people in these images varies largely with a wide range from 94 to 4,543. There are a total of 63,974 head annotations and the average number per image is 1280. Besides, this dataset has diverse scenes with varying perspective distortions.

UCF-QNRF dataset [11]. This dataset contains 1,535 images with a total of 1,251,642 head annotations. The images are divided into the training set with 1,201 images and the test set with 334 images, respectively. This dataset has much more annotated heads than currently available crowd datasets and is suitable for deep CNN based methods.

WorldExpo'10 dataset [46]. This dataset contains 1,132 annotated video sequences that are captured by 108 surveillance cameras from Shanghai 2010 WorldExpo event. There are a total of 199,923 annotated pedestrians from 3,980 frames. The dataset is divided into the training set with frames from 103 scenes and the test set with 600 frames from another 5 scenes.

4.2. Settings

Data. For the DANet, we augment the training data by cropping nine image patches at random locations in one image. Each image patch is one-fourth of the size of the original image. For the ASNet, we crop fixed-size image patches of 128×128 pixels at random locations in one image. Also, we flip each image patch horizontally to double the training set. Further, random color jitter is used in each epoch during the training. In particular, because the image resolution of the UCF-QNRF dataset [11] is too large, we resized its longer side to 1024 pixels and kept the aspect ratio constant.

Ground Truth Generation. We generate the ground-truth density maps by using a normalized Gaussian kernel to blur each head annotation, thus summing the density map equals the crowd count. In our experiments, we use a fixed spread Gaussian to generate density maps.

Training. For both the DANet and the ASNet, the first 13 convolutional layers are initialized from a pre-trained VGG-16 [33] model and the rest layers are randomly initialized by a Gaussian distribution with the mean of 0 and the standard deviation of 0.01. The Adam algorithm [15] is used to optimize the model. Both the DANet and the ASNet are trained in an end-to-end manner. The cross-entropy is adopted as the loss function for the DANet. We firstly train the DANet to generate attention masks and set the size of the training batch to 1. Then we train the ASNet and set the size of the training batch to 8.

4.3. Evaluation Metrics

We adopt the Mean Absolute Error (MAE) and the Mean Squared Error (MSE) to evaluate our method. The MAE and MSE are defined as follows:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|C_i - \hat{C}_i|, \tag{4}$$

$$MSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}||C_i - \hat{C}_i||^2}, \tag{5}$$

where N is the number of the test images, C_i and $\hat{C}_i$ are the ground-truth and estimated counts of the *i*-th image, respectively.

4.4. Evaluation and Analysis

In this section, we carry out experiments on the four datasets. We first conduct an ablation study to analyze the attention scaling and APLoss on the ShanghaiTech Part A dataset. And then we present the experimental results on the four datasets in detail. Finally, we present a qualitative analysis on the ShanghaiTech Part A dataset [49].

4.4.1 Ablation Study

Attention scaling. Our goal is to find out the impact of the attention scaling on crowd counting performance. Since the density can be divided into different levels, we aim to find the relatively optimal density levels. As described in Section 3.2, we can divide the density via the threshold set. In our experiments, we test two, four, and eight density levels. Accordingly, DANet provides two, four, and eight attention masks, without the background mask. The ablation results are presented in Table 1. We first train a baseline counting network by just utilizing the backbone and the DE-branch

	w/ scale		w/o scaling	
Masks	MAE	MSE	MAE	MSE
0	-	-	68.31	109.74
2	60.16	98.61	62.70	104.41
4	61.44	106.70	63.37	106.92
8	61.78	102.94	63.64	107.74

Table 1. Attention scaling ablation of our ASNet on ShanghaiTech Part A dataset, with 0, 2, 4, and 8 attention masks.

MSE Loss		2-level APLoss		3-level APLoss	
MAE	MSE	MAE	MSE	MAE	MSE
60.16	98.61	**57.78**	**90.13**	58.99	95.97

Table 2. APLoss ablation of our ASNet on ShanghaiTech Part A dataset.

	SHTech Part A			UCF_CC_50			UCF-QNRF			WorldExpo10							
Method	MAE	MSE	R.	MAE	MSE	R.	MAE	MSE	R.	S1	S2	S3	S4	S5	avg.	R.	avg. R.
HA-CCN [36]	62.9	94.9	9	256.2	348.4	14	118.1	180.4	11	-	-	-	-	-	-	-	11.3
SPN [5]	61.7	99.5	7	259.2	335.9	15	-	-	-	-	-	-	-	-	-	-	11
TEDnet [12]	64.2	109.1	13	249.4	354.5	13	113	188	10	2.3	10.1	11.3	13.8	2.6	8.0	6	10.5
ADCrowdNet [19]	63.2	98.9	12	266.4	358.0	16	-	-	-	1.6	13.2	8.7	10.6	2.6	7.3	3	10.3
ASD [40]	65.6	98.0	17	196.2	270.9	3	-	-	-	-	-	-	-	-	-	-	10
CFF [32]	65.2	109.4	16	-	-	-	93.8	146.5	3	-	-	-	-	-	-	-	9.5
SFCN [38]	64.8	107.5	15	214.2	318.2	6	102.0	171.4	6	-	-	-	-	-	-	-	9
PACNN [31]+ [16]	62.4	102.0	11	241.7	320.7	11	-	-	-	2.3	12.5	9.1	11.2	3.8	7.8	5	9
SPN+L2SM [42]	64.2	98.4	14	188.4	315.3	2	104.7	173.6	8	-	-	-	-	-	-	-	8
CAN [20]	62.3	100.0	10	212.2	243.7	5	107	183	9	2.9	12.0	10.0	7.9	4.3	7.4	4	7
PGCNet [43]	57.0	86.0	1	244.6	361.2	12	-	-	-	2.5	12.7	8.4	13.7	3.2	8.1	7	6.7
SPANet+SANet [6]	59.4	92.5	4	232.6	311.7	9	-	-	-	-	-	-	-	-	-	-	6.5
MBTTB-SCFB [37]	60.2	94.1	5	233.1	300.9	10	97.5	165.2	4	-	-	-	-	-	-	-	6.3
BL [22]	62.8	101.8	8	229.3	308.2	8	88.7	154.8	1	-	-	-	-	-	-	-	5.7
DSSINet [18]	60.63	96.04	6	216.9	302.4	7	99.1	159.2	5	1.57	9.51	9.46	10.35	2.49	6.67	2	5
S-DCNet [41]	58.3	95.0	3	204.2	301.3	4	104.4	176.1	7	-	-	-	-	-	-	-	4.7
Ours	57.78	90.13	2	174.84	251.63	1	91.59	159.71	2	2.22	10.11	8.89	7.14	4.84	6.64	1	1.5

Table 3. Comparisons of our ASNet with sixteen state-of-the-art methods on four datasets. The average ranking (denoted by avg. R.) is obtained by using the sum of all rankings that one method gains to divide the number of datasets it utilizes.

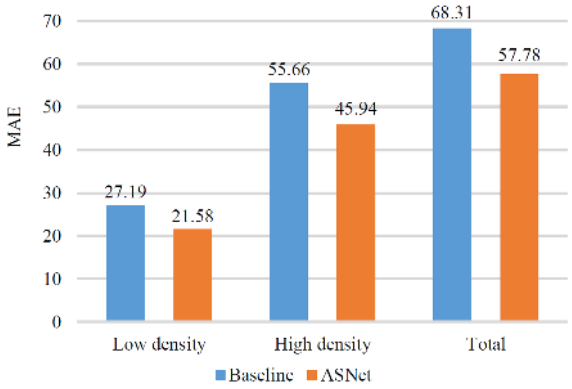

Figure 7. Comparison of ASNet with the baseline on ShanghaiTech Part A dataset. ASNet evidently reduces the estimation errors in both low-density and high-density regions.

of the proposed ASNet. It achieves an MAE of 68.31 and an MSE of 109.74. ASNet achieves the MAEs of 60.16, 61.44, and 61.78 when it uses 2, 4, and 8 attention masks, respectively. It means that the attention scaling mechanism does improve counting performance. Further, we carry out extra experiments by removing the scaling factors but still using the attention masks provided by DANet. The results are presented in the right column of Table 1. It achieves the MAEs of 62.70, 63.37, and 63.64 when it only uses 2, 4, and 8 attention masks without scaling factors, respectively.

It is seen that ASNet with 2 attention masks achieves the best result. More masks can bring more detailed density distribution information for the density prediction. However, the increase in attention masks may make it harder to fuse the whole density distribution.

APLoss. Since the training images for ASNet are of the same size of 128×128 pixels, we use the $AvgCnt_{11}$ as the threshold T. We test 2-level and 3-level APLoss, respectively. The experimental results are presented in Table 2. It is seen that the APLoss further improves the counting performance of ASNet. ASNet with 2-level APLoss achieves an MAE of 57.78, outperforming ASNet with MSE loss and ASNet with 3-level APLoss. The 2-level APLoss has a better generalization ability than the 3-level APLoss.

Besides, we carry out one statistical analysis to show that our method indeed reduces the estimation errors in regions of different density levels. The results are presented in Figure 7. Compared with the baseline network, our ASNet reduces the MAEs of low-density and high-density regions from 27.19 to 21.58 and from 55.66 to 45.94, respectively.

4.4.2 Results on Four Datasets

In this section, we evaluate our approach against sixteen currently reported methods [36, 5, 12, 19, 40, 32, 38, 31, 42, 20, 43, 6, 37, 22, 18, 41, 10] on ShanghaiTech Part A dataset [49], UCF_CCF_50 dataset [10], UCF-QNRF Dataset [11] and WorldExpo'10 dataset [46]. For simplicity, we denote ShanghaiTech Part A Dataset by SHTech Part A in our experiments. During the test, each whole image in the test sets of the four datasets is sent directly into our ASNet model. There are a few things that need to be made clear first. We perform a 5-fold cross-validation on the UCF_CCF_50 Dataset [10] by following the standard protocol adopted in [10]. On the WorldExpo'10 dataset [46],

we prune the last convolutional layer by setting the features out of ROI regions to zero, which is consistent with the previous work [46]. In addition, we only use the MAE metric to evaluate our approach. We first calculate the MAE for each test scene and then averages all the MAEs to evaluate the performance of ASNet across different test scenes.

Experimental results are presented in Tabel 3. All our results are achieved by using the ASNet model trained with the 2-level APLoss. (1) On the ShanghaiTech Part A dataset, our method achieves the second-best result with an MAE of 57.78, which is only 1.3% higher than that of the best method PGCNet [43]. However, on the UCF_CC_50 dataset, our method achieves the lowest MAE of 174.84 that is 28.5% lower than that of PGCNet. On the WorldExpo'10 dataset, our method also achieves the lowest MAE of 6.64 that is 9.1% lower than that of PGCNet. It should be noted that PGCNet uses additional perspective information to boost the accuracy of the prediction on the ShanghaiTech Part A dataset. (2) On the UCF_CC_50 dataset, our method reduces the MAE by 7.2% compared with the second-best method SPN+L2SM [42]. (3) On the UCF-QNRF dataset, our method achieves the second-best result with an MAE of 91.59, which is only 3.3% higher than that of the best BL [43] method. However, the MAE of our method is 8.0% and 23.8% lower than the BL method on the ShanghaiTech Part A dataset and the UCF_CC_50 dataset, respectively. The BL method uses the VGG-19 [33] model as the backbone which has deeper convolutional than our VGG-16 [33] backbone. (4) On the WorldExpo'10 dataset, our method surpasses all the other methods.

It noted that in Tabel 3 that some methods only achieve good performance on one dataset and relatively poor performance on the rest of these datasets. To make a comprehensive evaluation of the performance of all these methods on the four datasets, we introduce a simple evaluation metric named average ranking (denoted by avg. R. in Tabel 3). We obtain the average ranking value by using the sum of all ranks that one method gains to divide the number of datasets it utilizes. It is demonstrated that our method achieves the first average ranking which means it is able to excel in dealing with various complex crowd scenes.

4.4.3 Qualitative Analysis

We further carry out a qualitative analysis to investigate the performance of our ASNet. We present some qualitative comparisons between the baseline network and our ASNet in Figure 8. The main difference between them is that ASNet introduces attention mechanisms. It is observed from the visualization that our ASNet is much more robust on scenes with cluttered backgrounds like trees. We use red rectangles to mark the regions with cluttered backgrounds. There is obvious evidence from red rectangles of the first

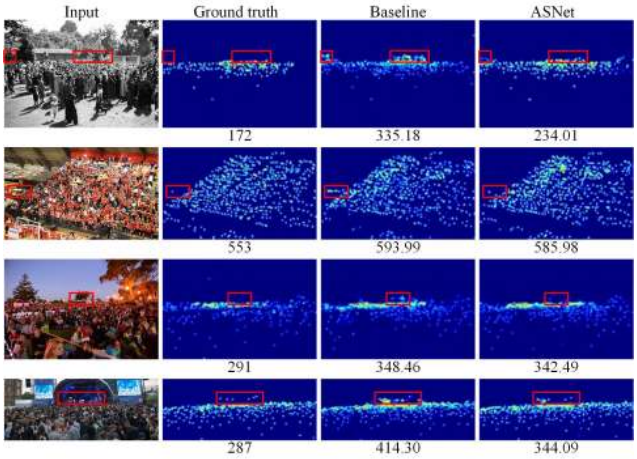

Figure 8. Visually qualitative analysis on the ShanghaiTech Part A dataset. ASNet is more robust to cluttered backgrounds than the baseline network.

and third rows where trees are close to the high-density regions. The baseline network causes much more estimation errors than the ASNet. Further, the second and the fourth rows show evidence in red rectangles where there are people in cluttered background. Our ASNet has a much more accurate density estimation. This demonstrates that the attention scaling mechanism not only attenuates the estimation error in regions of different density levels but also plays an important role in reducing the estimation error in cluttered background.

5. Conclusion

In this paper, we have presented a novel attention scaling based counting network that exploits attention masks and scaling factors to correct density estimations in regions of different density levels. To this end, We present one density attention network (DANet) to provide attention masks for the attention scaling network (ASNet). ASNet is responsible for generating scaling factors and outputting attention based density maps that only focus on their corresponding attention regions. These local density estimations together form the final density map. Besides, we introduce a novel adaptive pyramid loss (APLoss) that hierarchically calculates local estimation loss, strengthening the generalization ability of the counting network. Extensive experiments on four challenging datasets demonstrate the superiority of the proposed approach over current sixteen state-of-the-art methods.

Acknowledgments. This work was supported in part by the National Natural Science Foundation of China (Grant nos. 61802351, 61822701, 61872324, 61772474), in part by China Postdoctoral Science Foundation (Grant no. 2018M632802), and in part by Key R&D and Promotion Projects in Henan Province (Grant no. 192102310258).

References

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

[2] Jiale Cao, Yanwei Pang, and Xuelong Li. Pedestrian detection inspired by appearance constancy and shape symmetry. *IEEE Transactions on Image Processing*, 25(23):5538–5551, 2016. 2

[3] Jiale Cao, Yanwei Pang, and Xuelong Li. Learning multilayer channel features for pedestrian detection. *IEEE Transactions on Image Processing*, 26(7):3210–3220, 2017. 1

[4] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European Conference on Computer Vision*, pages 801–818, 2018. 1

[5] Xinya Chen, Yanrui Bin, Nong Sang, and Changxin Gao. Scale pyramid network for crowd counting. In *IEEE Winter Conference on Applications of Computer Vision*, pages 1941–1950, 2019. 2, 7

[6] Zhi-Qi Cheng, Jun-Xiu Li, Qi Dai, Xiao Wu, and Alexander G. Hauptmann. Learning spatial awareness to improve crowd counting. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6151–6160, 2019. 7

[7] Piotr Dollar, Christian Wojek, Bernt Schiele, and Pietro Perona. Pedestrian detection: an evaluation of the state of the art. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(4):743–761, 2012. 2

[8] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2961–2969, 2017. 1

[9] Mohammad Hossain, Mehrdad Hosseinzadeh, Omit Chanda, and Yang Wang. Crowd counting using scale-aware attention networks. In *IEEE Winter Conference on Applications of Computer Vision*, pages 1280–1288, 2019. 4

[10] Haroon Idrees, Imran Saleemi, Cody Seibert, and Mubarak Shah. Multi-source multi-scale counting in extremely dense crowd images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2547–2554, 2013. 1, 2, 5, 6, 7

[11] Haroon Idrees, Muhmmad Tayyab, Kishan Athrey, Dong Zhang, Somaya Al-Maadeed, Nasir Rajpoot, and Mubarak Shah. Composition loss for counting, density map estimation and localization in dense crowds. In *Proceedings of the European Conference on Computer Vision*, pages 532–546, 2018. 1, 5, 6, 7

[12] Xiaolong Jiang, Zehao Xiao, Baochang Zhang, Xiantong Zhen, Xianbin Cao, David Doermann, and Ling Shao. Crowd counting and density estimation by trellis encoder-decoder networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6133–6142, 2019. 7

[13] Xiaoheng Jiang, Li Zhang, Pei Lv, Yibo Guo, Ruijie Zhu, Yafei Li, Yanwei Pang, Xi Li, Bing Zhou, and Mingliang Xu. Learning multi-level density maps for crowd counting. *IEEE Transactions on Neural Networks and Learning Systems*, DOI:10.1109/TNNLS.2019.2933, 2019. 2

[14] Di Kang, Zheng Ma, and Antoni B Chan. Beyond counting: comparisons of density maps for crowd analysis tasks-counting, detection, and tracking. *IEEE Transactions on Circuits and Systems for Video Technology*, 29(5):1408–1422, 2018. 2

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

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

[17] Chenchen Liu, Xinyu Weng, and Yadong Mu. Recurrent attentive zooming for joint crowd counting and precise localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1217–1226, 2019. 3

[18] Lingbo Liu, Zhilin Qiu, Guanbin Li, Shufan Liu, Wanli Ouyang, and Liang Lin. Crowd counting with deep structured scale integration network. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1774–1783, 2019. 1, 7

[19] Ning Liu, Yongchao Long, Changqing Zou, Qun Niu, Li Pan, and Hefeng Wu. Adcrowdnet: An attention-injective deformable convolutional network for crowd understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3225–3234, 2019. 1, 3, 7

[20] Weizhe Liu, Mathieu Salzmann, and Pascal Fua. Context-aware crowd counting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5099–5108, 2019. 7

[21] Xialei Liu, Joost van de Weijer, and Andrew D Bagdanov. Leveraging unlabeled data for crowd counting by learning to rank. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7661–7669, 2018. 1, 3

[22] Zhiheng Ma, Xing Wei, Xiaopeng Hong, and Yihong Gong. Bayesian loss for crowd count estimation with point supervision. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6141–6150, 2019. 1, 7

[23] Mark Marsden, Kevin McGuinness, Suzanne Little, and Noel E O'Connor. Resnetcrowd: a residual deep learning architecture for crowd counting, violent behaviour detection and crowd density level classification. In *Proceedings of the IEEE Conference on Advanced Video and Signal Based Surveillance*, pages 1–7, 2017. 2

[24] Daniel Onoro-Rubio and Roberto J López-Sastre. Towards perspective-free object counting with deep learning. In *Proceedings of the European Conference on Computer Vision*, pages 615–629, 2016. 2

[25] Yanwei Pang, Jiale Cao, and Xuelong Li. Learning sampling distributions for efficient object detection. *IEEE Transactions on Cybernetics*, 47(1):117–129, 2017. 2

[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 the IEEE International Conference on Computer Vision*, pages 3253–3261, 2015. 2

[27] Joseph Redmon, Santosh Kumar Divvala, Ross B. Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 779–788, 2016. 1

[28] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Proceedings of the Neural Information Processing Systems*, pages 91–99, 2015. 1

[29] Deepak Babu Sam, Shiv Surya, and R Venkatesh Babu. Switching convolutional neural network for crowd counting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, page 6, 2017. 1, 2

[30] Zan Shen, Yi Xu, Bingbing Ni, Minsi Wang, Jianguo Hu, and Xiaokang Yang. Crowd counting via adversarial cross-scale consistency pursuit. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5245–5254, 2018. 1, 3

[31] Miaojing Shi, Zhaohui Yang, Chao Xu, and Qijun Chen. Revisiting perspective information for efficient crowd counting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7279–7288, 2019. 7

[32] Zenglin Shi, Pascal Mettes, and Cees G. M. Snoek. Counting with focus for free. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4199–4208, 2019. 7

[33] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 4, 5, 6, 8

[34] 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 the IEEE Conference on Advanced Video and Signal Based Surveillance*, pages 1–6, 2017. 1, 3

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

[36] Vishwanath A Sindagi and Vishal M Patel. Ha-ccn: Hierarchical attention-based crowd counting network. *IEEE Transactions on Image Processing*, 29:323–335, 2019. 7

[37] Vishwanath A. Sindagi and Vishal M. Patel. Multi-level bottom-top and top-bottom feature fusion for crowd counting. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1002–1012, 2019. 1, 2, 7

[38] Qi Wang, Junyu Gao, Wei Lin, and Yuan Yuan. Learning from synthetic data for crowd counting in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8198–8207, 2019. 7

[39] Yi Wang and Yuexian Zou. Fast visual object counting via example-based density estimation. In *Proceedings of the IEEE International Conference on Image Processing*, pages 3653–3657. IEEE, 2016. 2

[40] Xingjiao Wu, Yingbin Zheng, Hao Ye, Wenxin Hu, Jing Yang, and Liang He. Adaptive scenario discovery for crowd counting. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 2382–2386, 2019. 7

[41] Haipeng Xiong, Hao Lu, Chengxin Liu, Liang Liu, Zhiguo Cao, and Chunhua Shen. From open set to closed set: Counting objects by spatial divide-and-conquer. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 8361–837, 2019. 1, 7

[42] Chenfeng Xu, Kai Qiu, Jianlong Fu, Song Bai, Yongchao Xu, and Xiang Bai. Learn to scale: Generating multipolar normalized density maps for crowd counting. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 8381–8389, 2019. 7, 8

[43] Zhaoyi Yan, Yuchen Yuan, Wangmeng Zuo, Xiao Tan, Yezhen Wang, Shilei Wen, and Errui Ding. Perspective-guided convolution networks for crowd counting. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 952–961, 2019. 7, 8

[44] Anran Zhang, Jiayi Shen, Zehao Xiao, Fan Zhu, Xiantong Zhen, Xianbin Cao, and Ling Shao. Relational attention network for crowd counting. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6787–6796, 2019. 1, 3

[45] Anran Zhang, Lei Yue, Jiayi Shen, Fan Zhu, Xiantong Zhen, Xianbin Cao, and Ling Shao. Attentional neural fields for crowd counting. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5713–5722, 2019. 1, 3

[46] Cong Zhang, Hongsheng Li, Xiaogang Wang, and Xiaokang Yang. Cross-scene crowd counting via deep convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 833–841, 2015. 5, 6, 7, 8

[47] Tianzhu Zhang, Changsheng Xu, and Ming-Hsuan Yang. Learning multi-task correlation particle filters for visual tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):365–378, 2019. 1

[48] Tianzhu Zhang, Changsheng Xu, and Ming-Hsuan Yang. Robust structural sparse tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):473–486, 2019. 1

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

[50] Muming Zhao, Jian Zhang, Chongyang Zhang, and Wenjun Zhang. Leveraging heterogeneous auxiliary tasks to assist crowd counting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12736–12745, 2019. 3