Recurrent Attentive Zooming for Joint Crowd Counting and Precise Localization

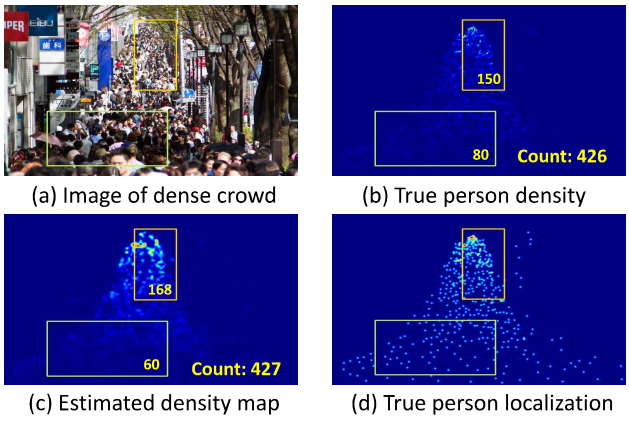
Chenchen Liu, Xinyu Weng, Yadong Mu*
Peking University
Beijing 100080, China
{liuchenchen, wengxy, myd}@pku.edu.cn

Abstract

Crowd counting is a new frontier in computer vision with far-reaching applications particularly in social safety management. A majority of existing works adopt a methodology that first estimates a person-density map and then calculates integral over this map to obtain the final count. As noticed by several prior investigations, the learned density map can significantly deviate from the true person density even though the final reported count is precise. This implies that the density map is unreliable for localizing crowd. To address this issue, this work proposes a novel framework that simultaneously solving two inherently related tasks - crowd counting and localization. The contributions are several-fold. First, our formulation is based on a crucial observation that localization tends to be inaccurate at high-density regions, and increasing the resolution is an effective albeit simple solution for improving localization. We thus propose Recurrent Attentive Zooming Network, which recurrently detects ambiguous image region and zooms it into high resolution for re-inspection. Second, the two tasks of counting and localization mutually reinforce each other. We propose an adaptive fusion scheme that effectively elevates the performance. Finally, a well-defined evaluation metric is proposed for the rarely-explored localization task. We conduct comprehensive evaluations on several crowd benchmarks, including the newly-developed large-scale UCF-QNRF dataset and demonstrate superior advantages over state-of-the-art methods.

1. Introduction

Nowadays surveillance cameras are densely mounted around many cities, which motivates the recent research enthusiasm about visual analysis for crowd scenes. This work targets the task of crowd counting in images [5, 30, 49, 15, 25, 38, 9] or videos [46], which is typically accomplished by generating and calculating integration over high-quality crowd density maps. Deep neural networks have become the de-facto standard for solving this challenging task, and a large number of such effective deep models have been developed. Key major hurdles in developing precise crowd counting algorithms include annotation errors, background clutter, camera angles, complex illumination, intra-crowd occlusions and density variation caused by crowd's varying distances to the camera etc.



(a) Image of dense crowd (b) True person density

(c) Estimated density map (d) True person localization

Figure 1. Illustration of the local inconsistency issue in density-based counting. Counting-oriented density map is clearly unsuitable for precisely localizing humans in congested scenes.

A few recent research [16] have argued that only summarizing counting numbers for congested scenes are not enough for real-life applications. Knowing accurate crowd density (rather than merely people count) is critical for spatially identifying high-risk regions and making correct decisions in applications such as crowd monitoring and traffic control. Unfortunately in a majority of existing methods, although the reported crowd count may be a good hit, the learned density map can significantly deviate from the true density. To illustrate it, we show an example in Figure 1(a)-(c). For the congested scene in sub-figure (a), a typical treatment is to generate a ground-truth density map by defining a normalized Gaussian distribution around each annotated person, the standard deviation of which can be determined either as a constant or by local crowdedness.

*Corresponding author.

It is shown in sub-figure (b). Importantly, the integral of the ground-truth density map equals to the true count. To train a deep network used for crowd counting, pixel-wise loss function is popularly utilized, which usually gauges the MSE (mean squared error) or MAE (mean absolute error) between the estimated and true densities. A typical density map generated by a learned deep model is found in sub-figure (c). As seen in sub-figures (b-c), the estimated density map does not always ensure the equality of true / estimated counts in an arbitrary sub-image, which essentially stems from the objective used during optimizing the crowd counting model.

In this paper, we propose a novel approach to simultaneous crowd counting and precise localization of people in a crowd image. We would emphasize that counting has traditionally been the central focus of current research, and precisely localizing each individual person has seldom been considered. As seen in Figure 1(b) and (d), though counting provides weak information about location of each person, accurate inference of localization through sharpening the density map is still generally infeasible. Tackling localization together has several noteworthy benefits, particularly rectifying counting errors and enabling some localization-dependent applications (such as human tracking that takes the output of localization as its initialization). In general, the results of counting alone are insufficient for general-purpose crowd analysis. An efficient and effective precise localization algorithm is particularly desired. Our technical contributions can be briefly stated as:

1) We develop a novel deep model for precise localization. Motivating our localization model we consider several desiderata: first, localization results are very sparse. Each dot location specifies a unique individual. Conventional MSE loss used in counting is thus not suitable. It is important to adopt some sparsity-encouraging loss for learning the localization model. In this work, we propose to use a normalized variant of binary cross entropy loss for this task, which is not considered in prior studies. Secondly, counting primarily only concerns the people count across entire image. It is empirically verified by us that reducing image resolution (e.g., to 1/8 of original resolution) indeed improves the counting accuracy. In contrary, using high resolution is key for precise crowd localization. In many highly congested scenes, we observe a significant localization performance gap in different regions with varying crowd density. In some image region, the high local crowd density brings serious challenge for spatially separating two nearby persons. An effective albeit simple solution is to selectively increase the local image resolution and re-feed the resized sub-image to the deep model for a second-pass processing. Guided by this intuition, we devise a network branch that learns an attention model, which hints the most likely image regions that need being zoomed and re-inspected. The procedure is recurrent until no further local zooming operation is required.

2) Our proposed deep model solves counting and localization in a mutually reinforced manner. Counting returns a density map which weakly implies the location of each person. And aggregating localization results naturally leads to a crowd count. The proposed model is multi-branched, with separate branch target counting / localization respectively. The final crowd count is computed as a consensus between the counting and localization branches. In practice, we propose to use scene-adaptive fusion weight, and the final count is obtained by averaging two counts according to the fusion weight. This way enforces strong consistency between counting and localization, and proves to elevate the accuracy of both tasks.

3) The evaluation protocol for crowd localization has not been comprehensively discussed and clearly stated in the literature. The only relevant study we found is conducted by Idrees et al. [16]. However, the evaluation protocol in [16] is problematic. It does not penalize multiple detections. Specifically, among all localized persons returned by an algorithm, multiple detections can be very close to an annotated person. According to [16], the closest one will be picked for further evaluation and others are simply ignored without receiving any penalty. This makes it difficult to compare two localization results with different counts. Moreover, crucial details (such as the way to rank all returned locations) are also unfortunately missing in [16]. Inspired by the treatments in human keypoint detection [44, 28, 27, 4, 7] and object detection [13, 12, 32, 20], we clearly define an evaluation metric for this task, with all due considerations to the complications in evaluating a crowd localization result.

The rest of the paper is organized as follows. Section 2 reviews related work. We present the proposed approach for simultaneous crowd counting and localization in Section 3. The experimental evaluations are detailed in Section 4, followed by concluding remarks in Section 5.

2. Related Work

This section briefly surveys crowd counting [24] and localization techniques. The relevant works can be roughly cast into the following categories:

Detection based methods: Detection-by-classification (or known as sliding window method) [10] is very popular in object detection owing to its conceptual simplicity and empirical high performance. These methods utilize highly discriminative local feature such as HOG feature [8] and can also detect partial objects (e.g., deformable parts based model [11]). For crowd scenes with few occlusions and low density, using well-trained detectors can lead to accurate count. The work of DecideNet [21] adopt human detectors for amending the (often over-estimated) counting in low-

density areas.

Density based methods: A thrust of early development [6, 15] deploy regression-based method to extract visual features from images and then regress the number of interested objects. More recent years have witnessed a burst of applying end-to-end deep neural networks to crowd counting [2, 29, 51]. The optimization of these deep models often boils down to learning a density map that locally approximates true crowd density. In particular, the scale variation issue has received much emphasize along this line of research. Earlier works adopt multi-scale image pyramid for boosting the accuracy in high-density areas [41, 48, 17]. Recent developments have considered spatial locality [22] cross-scale aggregation [3], and scale-adaptive [50]. The work in [37] enforces the predicted counts at different image scale as consistent as possible. [23] uses the ordinal relation between inclusive image regions for effectively augmenting the training data from unlabeled data. The idea of assembling multiple branches with different target also proves very effective, including Switch-Net [36] and Divide-and-Grow Net [35]. In [51], authors introduce a multi-column CNN for the counting purpose, defining different kernel sizes for tackling varying density. [34] includes a top-down branch to correct the initial prediction of the CNN. Other insights include the key role of large receptive field by CSRnet [19] where the authors experiment with a sequence of dilation convolutions, and the utilization of human body structure [14].

Localization in congested scenes: Ostensibly, one can naturally consider person localization by sharpening a crowd density map. However, the heavy inaccuracy of density map was extensively reported in prior studies. Ma et al. [26] develop a two-step method: first estimate density map using sliding windows, and then use integer programming to finally localize all objects of interest in an image. [33] devises a density-aware person detector, using density map as a regularizer when optimizing the detector to ensure the predicted density map more salient around true locations. Idrees el al. [16] propose a composition loss for simultaneously doing counting, density map estimation and localization in congested scenes. A so-far largest benchmark for this aim, called UCF-QNRF dataset, was also established. Another relevant work [18] predicts blobs in crowd images. Thinking beyond crowd, the localization technique can also be applied to localize objects in other domains. For example, Sirinukunwattana et al. [42] target cancer nuclei in the domain of medical imaging.

Limitations: State-of-the-art crowd counting approaches rarely emphasize on precise localization, mainly concerning the accuracy in terms of entire crowd count. This motivates our research on devising a novel solution for joint counting and person localization, as well as establishing solid evaluation metrics for benchmarking different localization algorithms.

3. The Proposed Approach

This section describes the proposed deep model that enables simultaneous crowd counting and precise localization, followed by details about learning network parameters.

3.1. Network Design

The deep network architecture can be found at Figure 2 and detailed configuration is deferred to the supplemental material due to space limit. As seen, the model is comprised of a Main Net and Recurrent Attentive Zooming (RAZ) Net. While the two nets have almost identical network layers and configurations, they do not share parameters with each other. For clarity let us first explain Main-Net and then emphasize the differences of RAZ-Net. The core of Main-Net are two branches which solve density based counting and precise localization respectively. These two are fused with data-adaptive weights. A third branch in the Main-Net is responsible for finding the optimal regions that can elevate the localization accuracy after being zoomed and re-processed by the RAZ-Net.

Counting Branch: Following the practice in [2, 19], we use truncated VGG-16 [39] as the backbone for both Main-Net and RAZ-net. The backbone essentially extracts discriminative features from the input image for further use by all branches in the network. Specifically, we decapitate the original VGG-16 by discarding all except for the first 13 convolutional layers. The output size of this truncated VGG-16 is 1/8 of the input image's resolution. Keeping more convolutional and pooling layers will obtain shrunken output, which seems not a reasonable choice for the localization branch. Such a shallow pre-trained backbone has already demonstrated strong transfer ability for crowd analysis.

Inspired by CSR-Net [19], in the counting branch we stack a number of dilated convolutional layers (dilation rate is set to 2 in all experiments) after the backbone. As verified by [19], larger receptive field is critical for crowd analysis, which inspires our use of a dilation rate of 2 in all convolutions. No size-changing layers (such as pooling or convolutional layer with $\geq$ 2 stride) are inserted. This way the final output size remains 1/8 of the input size. To generate a "true" continuous density map at each image position, we follow [51] to impose a Gaussian convolution around each annotated person. Assume $\boldsymbol{x}$ be an image pixel location and $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_N$ specify where N annotated person heads locate. The ground truth density at $\boldsymbol{x}$ is defined by

$$\widehat{\phi}(\boldsymbol{x}) = \sum_{i=1}^{N} \frac{1}{Z_i} \exp\left(-\|\boldsymbol{x} - \boldsymbol{a}_i\|^2 / \sigma_i^2\right), \text{ with } \sigma_i = \beta \bar{d}_i.$$

where $\bar{d}_i$ indicates the average spatial distance of k nearest persons (we use $k = 4$) and tends to be large at sparse areas.

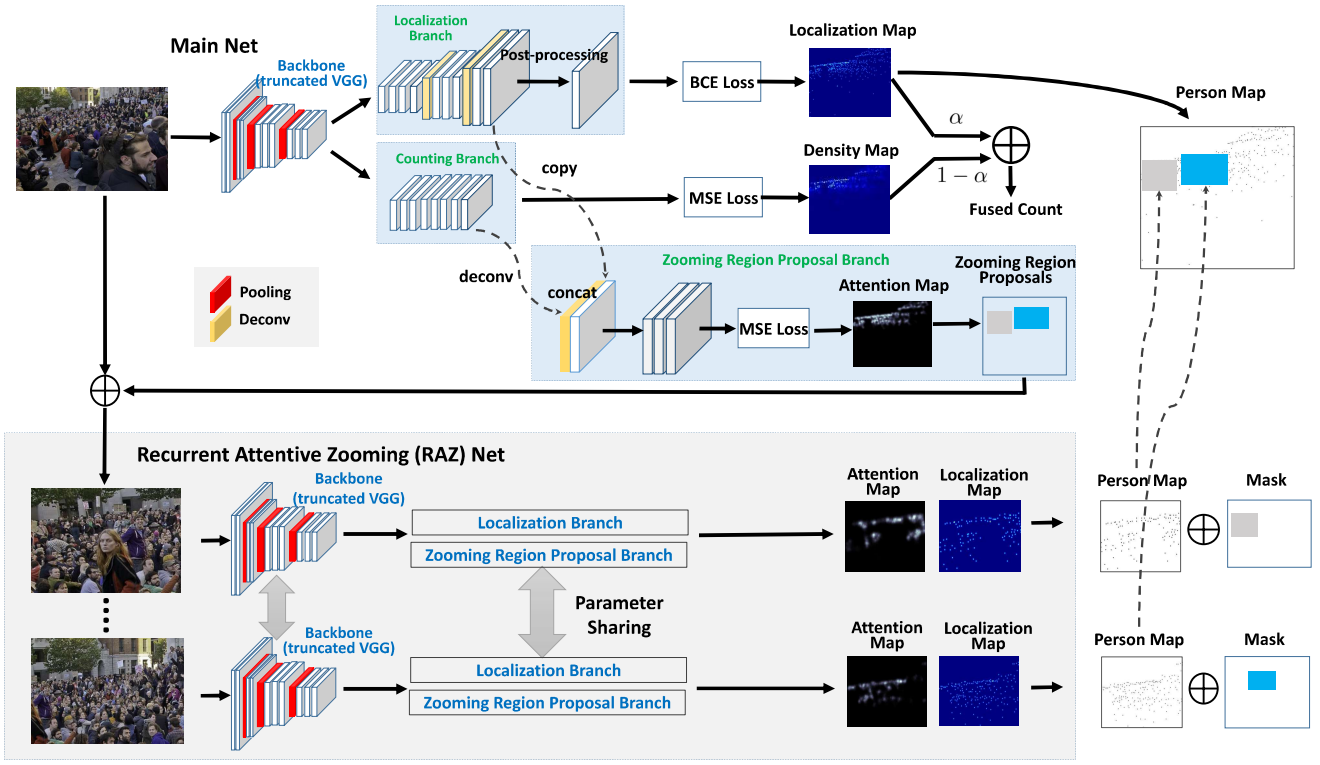


Figure 2. Illustration of our proposed network architecture for joint crowd counting and precise localization. Better viewing in color mode.

σ_i is a parameter that defines a geometry-adaptive Gaussian kernel and Z_i ensures the sum to 1. We empirically set $\beta = 0.1$ without further tuning.

Given an input image of size $m \times n$, the loss of counting branch can then be defined as the sum of pixel-wise density discrepancy:

$$\mathcal{E}_{den}(I) = 8 \cdot \sqrt{\frac{1}{mn} \sum_{i=1}^{(m/8)\times(n/8)} \left| \phi(\boldsymbol{x}_i) - \widehat{\phi}(\boldsymbol{x}_i) \right|^2}, \quad (1)$$

where $\phi(\boldsymbol{x}_i), \widehat{\phi}(\boldsymbol{x}_i)$ denote the estimated density value (output of counting branch's last convolutional layer) and the ground truth at image position $\boldsymbol{x}_i$ respectively.

Localization Branch: The results of the localization branch are generally expected to be sparse, only concerning points close to the annotations. However, in high-density areas, annotations are so dense that the annotated persons are just several pixels away from another nearest one. The spatially-varying sparsity brings severe challenges to the network design. We adopt binary cross entropy (BCE), which is sparsity-encouraging, as the loss function in the localization branch. Let us first describe the way to convert an image with labeled heads to a ground truth map. For each head position in the annotation set $\mathcal{I} = \{\boldsymbol{a}_1, \ldots, \boldsymbol{a}_N\}$, we introduce a delta function $\delta(\boldsymbol{x} - \boldsymbol{a}_i)$ at $\boldsymbol{x}$ and compute

$$\psi(\boldsymbol{x}) = \sum_{i=1}^{N} \delta(\boldsymbol{x} - \boldsymbol{a}_i) \otimes K, \quad (2)$$

where $\otimes$ denotes the convolution operator and $K = [0\ 1\ 0; 1\ 1\ 1; 0\ 1\ 0]$ is a 3×3 kernel. The label $Y(\boldsymbol{x})$ at pixel $\boldsymbol{x}$ in the localization map is set to 1 if $\psi(\boldsymbol{x}) > 0$, otherwise 0. Note that the adopted kernel K encourages a very small neighborhood of each annotated head to be classified as positive.

The input to the localization branch is sized 1/8 of original image resolution. It goes through 3 de-convolution layers, each doubles the feature map's spatial resolution. The final output thus has the same size to the input image, and approximates $\psi(\cdot)$. Let it be $\widetilde{\psi}(\cdot)$.Given an input image of $m \times n$, the loss of localization branch can then be defined as the sum of pixel-wise discrepancy:

$$\mathcal{E}_{loc}(I) = \frac{1}{mn} \sum_{i=1}^{m \times n} \ell(\boldsymbol{x}_i), \quad (3)$$

where $\ell(\boldsymbol{x}_i) = -\gamma \cdot Y(\boldsymbol{x}_i) \cdot \log(\widetilde{\psi}(\boldsymbol{x}_i)) - (1 - Y(\boldsymbol{x}_i)) \cdot \log(1 - \widetilde{\psi}(\boldsymbol{x}_i))$. We use a large constant γ (set to 100 in all experiments) to compensate the heavy imbalance between positives / negatives in the localization map.

To compute precise person head locations from $\widetilde{\psi}(\cdot)$, we first apply an average pooling (with a size of 3×3 and stride 1) for enlarging the value of true peaks and suppressing noises. Next, all local peak responses are picked out, and non-maxima suppression (NMS) operation is locally performed for avoiding over-close detected points.

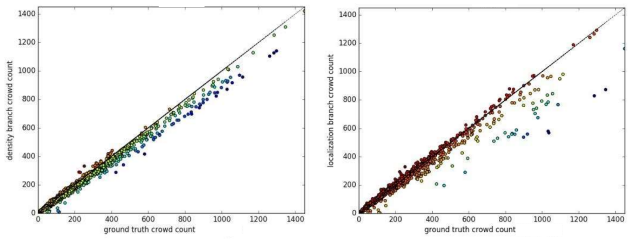


Figure 3. On the training data in ShanghaiTech-A, we plot the true counts in y-axis and the predicted counts in x-axis. Left / right sub-plots are for the counting / localization branch respectively.

In practice, the second step can be efficiently accomplished via a $k \times k$ max-pooling (we use $k = 3$) followed by pixel-wise comparison based peak identification, which ensures that any two points are at least k pixels away. This finishes the peak-finding job.

Two-Stream Fusion for Counting: Once properly fused, conventional counting and our proposed localization process can complement each other when calculating crowd count. Our adopted fusion scheme is inspired by an empirical insight in [21]: detection (which is essentially more precisely-bounded localization) reliability drops when crowd density increases, underestimating counts in those areas. To explore an optimal fusing scheme, we partition each training image into 4×4 non-overlapping sub-images, and contrast the estimated counts v.s. ground-truth counts for counting / localization branches respectively in Figure 3. Both density-based method and localization-derived counting degrade when the true density becomes dense, but the localization branch performs relatively worse, which is consistent to [21]. In other words, for high-density area, density-based estimation is more reliable and should dominate the count fusion.

To mitigate aforementioned estimation bias, we devise a fusion scheme that operates on a 4×4 image grid. For each of 16 sub-images indexed by $g = 1 \dots 16$, let C^g_{den} and C^g_{loc} be the counts returned by the counting / localization branch respectively. The final count after two-stream fusion C^g_{fused} is determined via

$$C^g_{fused} = \begin{cases} C^g_{den}, & \frac{C^g_{den}+C^g_{loc}}{2} \geq \theta_f \\ (C^g_{loc} + C^g_{den})/2, & otherwise \end{cases} \quad (4)$$

where two cases corresponds to $\alpha_g = 0, 0.5$ respectively. The final count for the entire image is computed via the sum $C_{fused} = \sum_{g=1}^{16} C^g_{fused}$. We do not choose to learn more fine-grained fusion weight α_g from data, since the crowd counting benchmarks are not sufficiently large-scale (*e.g.*, ShanghaiTech-A contains only 482 images) to combat over-fitting. In practice we empirically estimate the parameter θ_f for each grid from all training data.

Zooming Region Proposal Branch: It is difficult for the Main-Net to perfectly tackle crowds in all spatial scales. The high-density areas particularly suffer from low localization accuracy. Inspired by recent progress on attentive fine-grained recognition [47, 43, 52], we propose the idea of zooming a small number of sub-images guided by a learned attention map, which is implemented by the inclusion of a third *zooming region proposal branch* in the Main-Net, as seen in Figure 2.

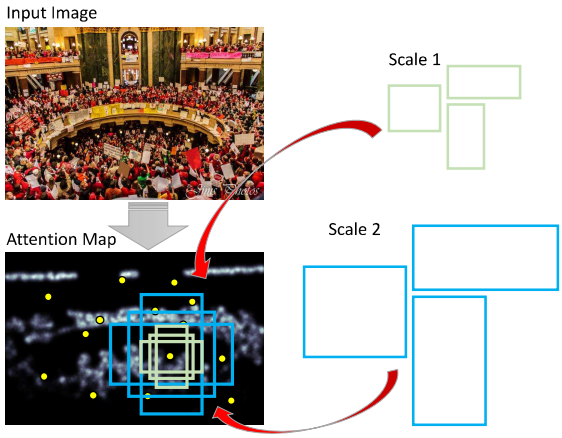


Figure 4. Illustration of multi-scale multi-ratio anchors for searching sub-images on the attention map.

To reduce network complexity, the input directly borrows and concatenates the last rich feature maps from the counting / localization branches that supposedly complement each other. Note that the feature map from the counting branch has 1/8 of the original image resolution. To enable a valid concatenation, it need to first go through an up-sampling layer to become 8 times larger. Ideally, the ground-truth attention map for this branch is expected to reflect spatial crowdness, and high attention is given to high-density areas. To this end, we adopt k-d tree [1] for efficiently finding 3-nearest heads for pixel i in the image and calculate the averaged spatial distance as $\bar{d}_i$. The true attention value for i is then computed via a non-linear transform $v_i = \exp(-\bar{d}_i^2/2\sigma_z^2)$, where we empirically set $\sigma_z = 3$ to de-emphasize low-density pixels.

To solicit near-optimal sub-images to be zoomed and re-checked, we adopt an exhaustive search scheme in order escalate the difficult combinatorial optimization problem. The core data structure in our proposed method are a set of multi-scale, multi-ratio *anchors*, inspired by its analogues in Faster R-CNN [32]. The process is illustrated in Figure 4. Several hundred interest points $p_1, p_2, \dots, p_m$ (yellow dots in Figure 4) are first randomly sampled from the attention map. Around each interest point, we apply all pre-defined anchors, each of which defines a sub-image centered at this specific point. All sub-images are then ranked in descending order according to their mean attention values. Those with a mean value below θ_z is abandoned and

others will undertake a standard non-maximum suppression (NMS) [11, 12, 32]. In particular, if two anchor-derived sub-images have a high intersection-over-union (IoU) [11], say 0.5 in our implementation, the one with lower mean value will be removed. At most k_z sub-images are chosen eventually. In all practice, we use $m = 200$, $\theta_z = 0.01$ and $k_z = 10$. The anchors we used have three height-to-width ratios (1:1, 1:2, 2:1) and two spatial scales (one of them is twice larger than the other). In the smaller scale, the anchor with 1:1 ratio is sized by 1/3 of shorter edge of the image, and we ensure anchors at the same spatial scale yet with other ratios occupy the same image area.

Recurrent Attentive Zooming (RAZ) Net: For simplicity, the chosen sub-images are zoomed uniformly by 2x. They are then fed into the RAZ-Net as shown in Figure 2. Different from the Main-Net, counting branch is removed since it is not the focus of RAZ-Net. Accordingly, the input to the attention branch now becomes only the feature map from the localization branch. All branches have the same network configurations to the Main-Net and copy the parameters from the Main-Net for initialization. The data used for training the RAZ-Net are fully comprised of sub-images solicited by the Main-Net. We omit the details of learning RAZ-Net since it is almost identical to that of the Main-Net. Importantly, RAZ-Net can be recurrently trained and used. For example, the attention map generated by the first RAZ-Net can infer a number of sub-images which are still over-dense and need additional zooming. After collecting those sub-images, we can further learn a second RAZ-Net with the same network architecture yet different data distribution / network parameters. However, to avoid over-fitting to the increasing smaller set of sub-images, we copy the parameters of the first RAZ-Net to all other recurrent ones without more fine-tuning.

3.2. Implementation Details

Joint learning the Main-Net and RAZ-Net is infeasible since the latter needs to wait for sub-images collected by the former net. We therefore train the two nets sequentially. Ostensibly one can learn the Main-Net in an end-to-end manner. However, we find this strategy often leads to heavy effort before convergence and very sensitive to initialization. We instead adopt a scheme of sequentially performing counting branch $\rightarrow$ localization branch $\rightarrow$ zooming region proposal branch.

The proposed algorithm is implemented in PyTorch and experimented on a private cluster with about 30 Titan X GPUs. Standard momentum is utilized with a parameter 0.95. The learning rate is initially set to 10^{-4} and halved after the validation performance is stuck in some plateau for many iterations.

Dataset	Images	Count	Resolution
ShanghaiTech_A [51]	482	501	589×868
ShanghaiTech_B [51]	716	123	768×1024
WorldExpo [49]	3980	56	576×720
UCF_QNRF [16]	1535	815	2013×2902

Table 1. Summary of benchmarks used for evaluations.

4. Evaluations

4.1. Dataset Description

We demonstrate our proposed approach in four different public benchmarks for crowd analysis. The key information is sketched briefly in Table 1. Particularly, the newly-introduced **UCF_QNRF** [16] is known as the dataset most qualified and challenging for experimenting with crowd localization, owing to its high-resolution images and tremendous annotations (over 1.25M annotated heads).

4.2. Evaluation Protocol

Let us first introduce the evaluation metrics. For the crowd counting task, we use the widely-adopted Mean Absolute Error (MAE) and (Root) Mean Squared Error (MSE). If the predicted count for image i is C_i and the ground truth count is $\widehat{C}_i$, the MAE and RMSE can be computed as (n is the image number):

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|\widehat{C}_i - C_i\right|, MSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left\|\widehat{C}_i - C_i\right\|^2}.$$

As argued in prior section, precise crowd localization is a relatively unexplored task, and its evaluation metric has not been firmly established in the literature. The only relevant work in [16] proposes 1-1 matching / ranking based scores for precise localization. Unfortunately, a close investigation will find that the metric in [16] leads to optimistic estimation. The key issue essentially stems from not penalizing over-detection: if multiple points closely match with a true person head, the nearest one will be kept and others are simply ignored without receiving any penalty. This fails it to serve as a widely-acknowledged metric for fair comparison.

Inspired by the evaluation metric used in the keypoint detection task of MS-COCO [20], we propose to assess a localization algorithm as follows: 1) all predicted head points are ranked according to their confidences returned via the model (the localization branch in our approach); 2) from the top-ranked points to the least confident ones, we sequentially classify each point as either *true positive* or *false positive*. A point under investigation will first be matched to a nearest true person. Only when this true person has not be matched by some higher-ranked points and their affinity is above some pre-defined threshold, it

Method	ShanghaiTech_A		ShanghaiTech_B		WorldExpo						UCF_QNRF	
	MAE	MSE	MAE	MSE	S1	S2	S3	S4	S5	avg.	MAE	MSE
Zhang *et al.* [49]	181.8	277.7	32.0	49.8	9.8	14.1	14.3	22.2	3.7	12.9	-	-
MCNN [51]	110.2	173.2	26.4	41.3	3.4	20.6	12.9	13.0	8.1	11.6	277	426
Cascaded-MTL [40]	101.3	152.4	20.0	31.1	-	-	-	-	-	-	252	514
Huang *et al.* [14]	-	-	20.2	35.6	4.1	21.7	11.9	11.0	3.5	10.5	-	-
Switch-CNN [36]	90.4	135.0	21.6	33.4	4.4	15.7	10.0	11.0	5.9	9.4	228	445
CP-CNN [41]	73.6	106.4	20.1	30.1	2.9	14.7	10.5	10.4	5.8	8.9	-	-
CSRNet [19]	68.2	115.0	10.6	16.0	2.9	11.5	8.6	16.6	3.4	8.6	-	-
SANet [3]	67.0	104.5	8.4	13.6	2.6	13.2	9.0	13.3	3.0	8.2	-	-
Idrees [16]	-	-	-	-	-	-	-	-	-	-	132	191
RAZ_density	67.5	109.5	9.6	15.4	2.5	11.3	9.5	13.4	4.7	8.3	126	208
RAZ_localization	75.2	133.0	13.5	25.4	2.0	13.9	9.1	44.5	3.2	14.6	135	246
RAZ_localization+	71.6	120.1	9.9	15.6	2.0	13.9	9.1	32.3	3.2	12.5	118	198
RAZ_average	66.5	111.5	8.6	14.2	2.0	12.3	8.8	28.2	3.2	10.9	117	195
RAZ_fusion	65.1	106.7	8.4	14.1	2.0	11.8	9.0	13.6	3.3	8.0	116	195

Table 2. Experimental results for crowd counting on all four benchmarks. '-' denotes either no reported results or the model does not converge. See main text for more explanation.

is marked as a true positive, and the matched true person will be marked as *matched* accordingly. With this ranked point list with binary true / false classification, standard Average Precision (AP) and Average Recall (AR) score can be efficiently computed.

4.3. Results Analysis

Crowd Counting: We first reiterate the significance of this work compared with the state-of-the-art for the counting task. Table 2 presents all quantitative results. We have included several strong competing algorithms (such as CSR-Net and SANet) when this work is done. We denote our own algorithmic variants using the prefix "RAZ". Among them, *RAZ_density*, *RAZ_localization* are results of counting / localization branch in the Main-Net respectively. *RAZ_localization+* is an enhanced version of *RAZ_localization* since it amends the count of localized person heads by feedbacking the zoom-and-localize results in the RAZ-Net. Both *RAZ_average* and *RAZ_fusion* are the results obtained by fusing counting / localization branches. Differently, *RAZ_average* naively averages the two and *RAZ_fusion* adopts our proposed fusing scheme.

Some facts can be observed in Table 2. First, RAZ-Net significantly improves the accuracies of the localization branch (see the gap between *RAZ_localization* and *RAZ_localization+*). Secondly, fusing the counts returned by our two branches is almost always beneficial, even using a blind uniform fusion. Since our proposed fusion scheme empirically considers the source of prediction errors, it further elevates the performance in most cases. This provides strong evidence that the two tasks of counting / localization are closely inter-related and arguably complementary. Last, though we did not utilize sophisticated tricks such as multi-scale kernels as in [3], our proposed approach still predict the most accurate crowd counts in a majority of cases.

Person Localization: Table 3 presents the localization results evaluated in our proposed metric. Around each annotated person head, we impose an un-normalized Gaussian function parameterized by σ. A predicted point will be regarded as a true positive only if the Gaussian function (when taking this point as an input) returns an output greater than 0.50 when evaluating the performance at AP.50, and 0.75 at AP.75. In Table 3, we report performances at three different accuracy levels, which establishes a reasonable baseline for further comparison with other algorithms. In addition, some representative localization results are shown in Figure 5. It can be seen that even for very dense crowds, the proposed method still generates precise localization.

Effect of Resolution: We also investigate a key problem in crowd analysis: how the resolution of feature map affects the performance? For tasks of localizing human joints / heads, Xiao et al. [45] have conducted a very comprehensive study and proved the key role of high resolution, which explains our choice of using full-sized feature map for the localization branch. There exist prior works that consider the resolution issue in crowd counting. For example, [31] presents a two branch architecture, where the first branch generates a low resolution density map, and the second branch incorporates the low resolution prediction and feature maps from the first branch to generate a high resolution density map. However, no strong evidence about the benefit of high-resolution map is given therein. To clarify the effect of resolution, we separately train the counting branch using the same backbone on ShanghaiTech_A. We tune the number of de-convolution layers such that the feature map resolutions increases as $1/8 \rightarrow 1/4 \rightarrow 1/2$ of input size. Interesting, the MAE scores on the test set vary as $67.5 \rightarrow 70.5 \rightarrow 81.4$, which justifies our design of the counting branch.

Recurrence Depth: It is interesting to know whether recurrently zooming the attended sub-images helps. To this end, we conduct more experiments on three benchmarks

Method	σ	Results of Localization Branch in Main Net						Results of Main Net + RAZ Net					
		AP.50	AP.75	mAP	AR.50	AR.75	mAR	AP.50	AP.75	mAP	AR.50	AR.75	mAR
ShanghaiTech_A	40	0.739	0.702	0.687	0.836	0.811	0.803	0.745	0.699	0.691	0.847	0.820	0.812
	20	0.670	0.597	0.576	0.791	0.746	0.729	0.667	0.601	0.584	0.799	0.753	0.741
	5	0.308	0.120	0.147	0.529	0.326	0.331	0.360	0.205	0.197	0.409	0.579	0.422
ShanghaiTech_B	40	0.742	0.695	0.692	0.833	0.804	0.809	0.753	0.716	0.710	0.857	0.834	0.831
	20	0.673	0.622	0.598	0.793	0.761	0.746	0.687	0.649	0.634	0.820	0.794	0.784
	5	0.356	0.156	0.181	0.575	0.375	0.374	0.456	0.281	0.280	0.661	0.516	0.492
WorldExpo	40	0.756	0.707	0.695	0.878	0.850	0.840	0.762	0.704	0.692	0.872	0.840	0.829
	20	0.674	0.593	0.555	0.828	0.771	0.741	0.669	0.580	0.546	0.815	0.753	0.724
	5	0.195	0.060	0.086	0.428	0.230	0.252	0.187	0.057	0.083	0.409	0.216	0.239
UCF_QNRF	40	0.558	0.458	0.449	0.692	0.624	0.604	0.573	0.481	0.462	0.719	0.652	0.636
	20	0.390	0.252	0.250	0.570	0.452	0.438	0.414	0.287	0.284	0.602	0.497	0.483
	5	0.047	0.017	0.021	0.178	0.092	0.100	0.079	0.031	0.037	0.242	0.143	0.148

Table 3. Summary of localization performance on four crowd benchmarks in terms of standard AP/AR used in MS-COCO [20]. mAP is computed via averaging from AP.50 to AP.95, with a stride of .05. More explanation about the metric is in the supplemental material.

Figure 5. Representative localization results on some testing images. The red / green blobs denote true and predicted person heads respectively. Better viewing if enlarging the images.

Method	ShanghaiTech_A		ShanghaiTech_B		UCF_QNRF	
	MAE	MSE	MAE	MSE	MAE	MSE
density	67.5	109.5	9.6	15.4	126	208
loc_1	75.2	133.0	13.5	25.4	135	246
loc_2	71.6	120.1	9.9	15.6	118	198
loc_3	70.5	120.8	9.4	13.3	120	200
loc_4	71.4	124.2	9.6	14.6	120	201
average_2	66.5	111.5	8.6	14.2	117	195
average_3	65.8	111.5	8.5	13.5	118	196
average_4	66.4	113.4	8.6	13.7	119	197
fusion_1	66.1	108.4	9.5	15.2	123	203
fusion_2	65.1	106.7	8.4	14.1	116	195
fusion_3	64.7	106.6	8.5	13.5	118	195
fusion_4	65.3	108.8	8.5	13.4	119	198

Table 4. Investigation of recurrent depth. See text for more details

and the results are shown in Table 4. *density* denotes the performance merely based on the counting branch. For all methods, we use the suffix to imply the number of recurrently using the RAZ-Net. *loc_**, *average_**, *fusion_** represent the localization branch and two variants of fusing counting / localization branches respectively. We clearly observe a non-trivial improvement when recurrently using RAZ-Net twice. However, the effect quickly decays with more recurrence.

5. Conclusion

This paper investigated jointly estimating counts and precise localization in congested scenes. Our main insight is that sparsity-encouraging loss function and high resolution are key for the rarely-explored task of precise localization. We also show that density-based counting and localization can collaboratively boost the accuracy of prediction. An evaluation metric is well-defined for comparing different localization algorithms. Our future work includes extending this approach to other domains, such as cells or bacteria from microscopic images.

Acknowledgement: This work is supported by Beijing Municipal Commission of Science and Technology under grant No. 181100008918005, NSF China under grant no. 61772037 and NJUST Key Laboratory of Intelligent Perception and Systems for High-Dimensional Information under grant no. JYB201701.

References

[1] Jon Louis Bentley. Multidimensional binary search trees used for associative searching. *Commun. ACM*, 18(9):509–517, 1975. 5

[2] Lokesh Boominathan, Srinivas S. S. Kruthiventi, and R. Venkatesh Babu. Crowdnet: A deep convolutional network for dense crowd counting. In *ACM Multimedia*, 2016. 3

[3] Xinkun Cao, Zhipeng Wang, Yanyun Zhao, and Fei Su. Scale aggregation network for accurate and efficient crowd counting. In *ECCV*, 2018. 3, 7

[4] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, 2017. 2

[5] Antoni B. Chan, Zhang-Sheng John Liang, and Nuno Vasconcelos. Privacy preserving crowd monitoring: Counting people without people models or tracking. In *CVPR*, 2008. 1

[6] Antoni B. Chan and Nuno Vasconcelos. Bayesian poisson regression for crowd counting. In *ICCV*, 2009. 3

[7] Yilun Chen, Zhicheng Wang, Yuxiang Peng, Zhiqiang Zhang, Gang Yu, and Jian Sun. Cascaded pyramid network for multi-person pose estimation. *CoRR*, abs/1711.07319, 2017. 2

[8] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005. 2

[9] Xinghao Ding, Zhirui Lin, Fujin He, Yu Wang, and Yue Huang. A deeply-recursive convolutional network for crowd counting. In *ICASSP*, 2018. 1

[10] Piotr Dollár, Christian Wojek, Bernt Schiele, and Pietro Perona. Pedestrian detection: A benchmark. In *CVPR*, 2009. 2

[11] Pedro F. Felzenszwalb, Ross B. Girshick, David A. McAllester, and Deva Ramanan. Object detection with discriminatively trained part-based models. *IEEE Trans. Pattern Anal. Mach. Intell.*, 32(9):1627–1645, 2010. 2, 6

[12] Ross B. Girshick. Fast R-CNN. In *ICCV*, 2015. 2, 6

[13] Ross B. Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, 2014. 2

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

[15] Haroon Idrees, Imran Saleemi, Cody Seibert, and Mubarak Shah. Multi-source multi-scale counting in extremely dense crowd images. In *CVPR*, 2013. 1, 3

[16] Haroon Idrees, Muhmmad Tayyab, Kishan Athrey, Dong Zhang, Somaya Al-Máadeed, Nasir M. Rajpoot, and Mubarak Shah. Composition loss for counting, density map estimation and localization in dense crowds. In *ECCV*, 2018. 1, 2, 3, 6, 7

[17] Di Kang and Antoni B. Chan. Crowd counting by adaptively fusing predictions from an image pyramid. In *BMVC*, 2018. 3

[18] Issam H. Laradji, Negar Rostamzadeh, Pedro O. Pinheiro, David Vázquez, and Mark W. Schmidt. Where are the blobs: Counting by localization with point supervision. In *ECCV*, 2018. 3

[19] Yuhong Li, Xiaofan Zhang, and Deming Chen. Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes. In *CVPR*, 2018. 3, 7

[20] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In *ECCV*, 2014. 2, 6, 8

[21] Jiang Liu, Chenqiang Gao, Deyu Meng, and Alexander G. Hauptmann. Decidenet: Counting varying density crowds through attention guided detection and density estimation. *CoRR*, abs/1712.06679, 2017. 2, 5

[22] Lingbo Liu, Hongjun Wang, Guanbin Li, Wanli Ouyang, and Liang Lin. Crowd counting using deep recurrent spatial-aware network. In *IJCAI*, 2018. 3

[23] Xialei Liu, Joost van de Weijer, and Andrew D. Bagdanov. Leveraging unlabeled data for crowd counting by learning to rank. In *CVPR*, 2018. 3

[24] Chen Change Loy, Ke Chen, Shaogang Gong, and Tao Xiang. Crowd counting and profiling: Methodology and evaluation. *Modeling, Simulation and Visual Analysis of Crowds*, pages 347–382, 2013. 2

[25] Zheng Ma and Antoni B. Chan. Crossing the line: Crowd counting by integer programming with local features. In *CVPR*, 2013. 1

[26] Zheng Ma, Lei Yu, and Antoni B. Chan. Small instance detection by integer programming on object density maps. In *CVPR*, pages 3689–3697, 2015. 3

[27] Alejandro Newell, Zhiao Huang, and Jia Deng. Associative embedding: End-to-end learning for joint detection and grouping. In *NIPS*, 2017. 2

[28] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *ECCV*, 2016. 2

[29] Daniel Oñoro-Rubio and Roberto Javier López-Sastre. Towards perspective-free object counting with deep learning. In *ECCV*, 2016. 3

[30] Viet-Quoc Pham, Tatsuo Kozakaya, Osamu Yamaguchi, and Ryuzo Okada. COUNT forest: Co-voting uncertain number of targets using random forest for crowd density estimation. In *ICCV*, 2015. 1

[31] Viresh Ranjan, Hieu Le, and Minh Hoai. Iterative crowd counting. In *ECCV*, 2018. 7

[32] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster R-CNN: towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(6):1137–1149, 2017. 2, 5, 6

[33] Mikel Rodriguez, Ivan Laptev, Josef Sivic, and Jean-Yves Audibert. Density-aware person detection and tracking in crowds. In *ICCV*, 2011. 3

[34] Deepak Babu Sam and R. Venkatesh Babu. Top-down feedback for crowd counting convolutional neural network. In *AAAI*, 2018. 3

[35] Deepak Babu Sam, Neeraj N. Sajjan, and R. Venkatesh Babu. Divide and grow: Capturing huge diversity in crowd images with incrementally growing CNN. In *CVPR*. 3

[36] Deepak Babu Sam, Shiv Surya, and R. Venkatesh Babu. Switching convolutional neural network for crowd counting. In *CVPR*, 2017. 3, 7

[37] Zan Shen, Yi Xu, Bingbing Ni, Minsi Wang, Jianguo Hu, and Xiaokang Yang. Crowd counting via adversarial cross-scale consistency pursuit. In *CVPR*, 2018. 3

[38] Z. Shi, L. Zhang, Y. Liu, X. Cao, Y. Ye, M-M. Cheng, and G. Zheng. Crowd counting with deep negative correlation learning. In *CVPR*, 2018. 1

[39] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014. 3

[40] Vishwanath A. Sindagi and Vishal M. Patel. Cnn-based cascaded multi-task learning of high-level prior and density estimation for crowd counting. In *14th IEEE International Conference on Advanced Video and Signal Based Surveillance, AVSS 2017, Lecce, Italy, August 29 - September 1, 2017*, pages 1–6, 2017. 7

[41] Vishwanath A. Sindagi and Vishal M. Patel. Generating high-quality crowd density maps using contextual pyramid cnns. In *ICCV*, 2017. 3, 7

[42] Korsuk Sirinukunwattana, Shan e Ahmed Raza, Yee-Wah Tsang, David R. J. Snead, Ian A. Cree, and Nasir M. Rajpoot. Locality sensitive deep learning for detection and classification of nuclei in routine colon cancer histology images. *IEEE Trans. Med. Imaging*, 35(5):1196–1206, 2016. 3

[43] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, 2017. 5

[44] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. Convolutional pose machines. In *CVPR*, 2016. 2

[45] Bin Xiao, Haiping Wu, and Yichen Wei. Simple baselines for human pose estimation and tracking. In *ECCV*, 2018. 7

[46] Feng Xiong, Xingjian Shi, and Dit-Yan Yeung. Spatiotemporal modeling for crowd counting in videos. In *ICCV*, 2017. 1

[47] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C. Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, 2015. 5

[48] Lingke Zeng, Xiangmin Xu, Bolun Cai, Suo Qiu, and Tong Zhang. Multi-scale convolutional neural networks for crowd counting. In *ICIP*, 2017. 3

[49] Cong Zhang, Hongsheng Li, Xiaogang Wang, and Xiaokang Yang. Cross-scene crowd counting via deep convolutional neural networks. In *CVPR*, 2015. 1, 6, 7

[50] Lu Zhang, Miaojing Shi, and Qiaobo Chen. Crowd counting via scale-adaptive convolutional neural network. In *WACV*, 2018. 3

[51] Yingying Zhang, Desen Zhou, Siqin Chen, Shenghua Gao, and Yi Ma. Single-image crowd counting via multi-column convolutional neural network. In *CVPR*, 2016. 3, 6, 7

[52] Heliang Zheng, Jianlong Fu, Tao Mei, and Jiebo Luo. Learning multi-attention convolutional neural network for fine-grained image recognition. In *ICCV*, 2017. 5