This document is downloaded from DR-NTU (https://dr.ntu.edu.sg)
Nanyang Technological University, Singapore.

Multi-resolution attention convolutional neural network for crowd counting

Zhang, Youmei; Zhou, Chunluan; Chang, Faliang; Kot, Alex Chichung

2019

Zhang, Y., Zhou, C., Chang, F., & Kot, A. C. (2019). Multi-resolution attention convolutional neural network for crowd counting. Neurocomputing, 329, 144–152. doi:10.1016/j.neucom.2018.10.058

https://hdl.handle.net/10356/144965

https://doi.org/10.1016/j.neucom.2018.10.058

© 2018 Elsevier B.V. All rights reserved. This paper was published in Neurocomputing and is made available with permission of Elsevier B.V.

Downloaded on 13 Mar 2024 18:42:04 SGT

Multi-resolution Attention Convolutional Neural Network for Crowd Counting

Youmei Zhang[a], Chunluan Zhou[b], Faliang Chang[a,*], Alex C. Kot[b]

[a]*School of Control Science and Engineering, Shandong University, Jinan, China 250061.*
[b]*School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore 639798.*

Abstract

Estimating crowd counts remains a challenging task due to the problems of scale variation, non-uniform distribution and complex backgrounds. In this paper, we propose a multi-resolution attention convolution neural network (MRA-CNN) to address this challenging task. Except for the counting task, we exploit an additional density-level classification task during training and combine features learned for the two tasks, thus forming multi-scale, multi-contextual features to cope with the scale variation and non-uniform distribution. Besides, we utilize a multi-resolution attention (MRA) model to generate score maps, where head locations are with higher scores to guide the network to focus on head regions and suppress non-head regions regardless of the complex backgrounds. During the generation of score maps, atrous convolution layers are used to expand the receptive field with fewer parameters, thus getting higher-level features and providing the MRA model more comprehensive information. Experiments on ShanghaiTech, WorldExpo'10 and UCF datasets demonstrate the effectiveness of our method.

Keywords: Crowd Counting, Multi-resolution Attention (MRA) Model, Convolution Neural Network (CNN), Atrous Convolution

1. Introduction

Crowd analysis has drawn remarkable attention for its wide applications such as intelligent surveillance, public safety and urban planning [1][2]. As

*Corresponding Author: Faliang Chang. email: flchang@sdu.edu.cn.

one of the crowd analysis tasks, crowd counting plays an important role in crowd control, traffic monitoring and urban security. Thanks to the powerful image processing capability of convolutional neural networks (CNNs), recent CNN-based counting approaches have seen a great success[3][4][5]. However, due to the problems such as complex backgrounds, scale variation and non-uniform distribution, as shown in Fig. 1, crowd counting still remains a challenging task in practical applications.

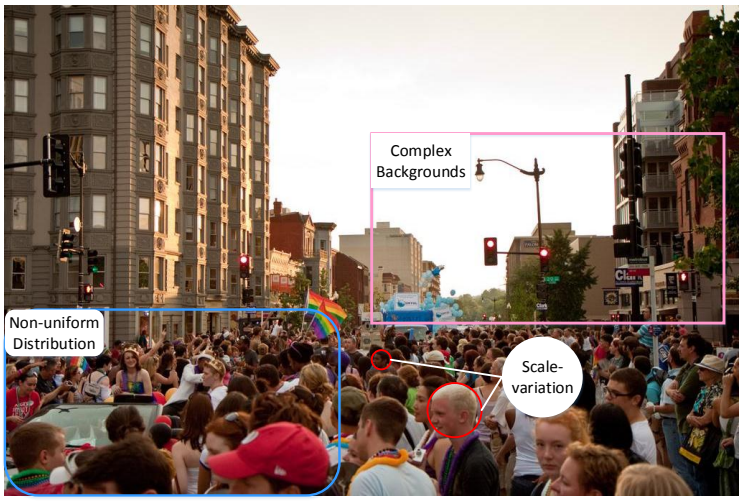

Figure 1: Challenges for crowd counting

Recent crowd counting works address [3][4] the above challenges with different CNN architectures. A Crowd-CNN model [3], which is robust to complex backgrounds, is proposed to address cross-scene crowd counting. The authors make their first attempt to use a CNN-based architecture to generate density maps for crowd counting. In [4], different receptive fields are applied in the multi-column CNN (MCNN) channels to extract multi-scale features. The MCNN is adaptive to head-scale variations caused by perspective effect or image resolution. A contextual pyramid CNN (CP-CNN) fuses both global and local context information and high-dimensional features to generate high quality density maps. The density maps generated by the CP-CNN capture the distribution of crowds. All these methods only address one or two of the above mentioned challenging problems. In addition, most of recent counting approaches conduct crowd counting by learning density maps of images (e.g. [6][7][8]), which can estimate not only the counts but also the density

and distribution of the crowds. All these approaches generate ground-truth density maps according to head locations in the images, which also conforms to human-style counting. However, the importance of head locations is not explicitly reflected in the architecture of the CNN model.

This paper aims to address the above-mentioned challenges with a MRA-CNN for crowd counting. Fig.2 illustrates the architecture of the proposed method. We take the density level classification as an auxiliary task and combine the learned features in this task with those learned in the crowd task, thus providing multi-contextual and multi-scale information. Several attention maps are applied to different feature layers to learn multi-scale score maps, which can guide the final feature layer to focus on head regions. In addition, the architecture gets a larger receptive field with fewer parameters by using atrous convolution layers.

The contributions of this work can be summarized as follows:

1) The proposed method is robust to scale variation and non-uniform distribution by learning multi-scale, multi-contextual features with multi-task learning.
2) The MRA-CNN architecture addresses the problem of complex backgrounds by estimating size-adaptive density maps and utilizing the MRA model, which guides the network to focus on head regions.
3) The atrous convolution layer is utilized to increase the receptive field with fewer parameters to get higher level features, and simultaneously provide the MRA model more comprehensive information.

The remainder of the paper is organized as follows. Section 2 presents some recent CNN-based related works on crowd counting. In Section 3, the details of our proposed MRA-CNN architecture are introduced. Experimental results are given and discussed in Section 4. Finally, Section 5 concludes the paper.

2. Related work

CNN has demonstrated its great success in various computer vision tasks, such as classification [9][10], detection [11][12], segmentation [13], re-identification [14] and perception[15]. Researchers are motivated to explore the applications of CNN on crowd counting and have seen significant improvements. This section introduces recent CNN-based counting approaches. We broadly classify these counting methods into two categories: single task and multi-task

frameworks, and each can be further separated into counts-only and density map-based methods according to the output of the framework. Counts-only methods output the number of people directly while density map-based approaches learn the non-linear function from crowd images to their corresponding density maps, which represent the counts by the sum of pixel values.

Herein, we introduce some representative single-task, counts-only methods [16] [17]. [16] presents an end-to-end CNN architecture to map the whole image to its global counts, which makes use of sharing computations over overlapping regions. This method incorporates contextual information to tackle the problem of complex backgrounds by simultaneously learning local counts. A mixture of CNNs (MoCNN) [17] addresses the large appearance changes caused by scale and congestions with the cooperation of a gating CNN and expert CNNs. The weights learned by the gating CNN according to the appearance of the patches are multiplied to the prediction of expert CNNs, making the network robust to various appearance changes.

In some specific scenarios, e.g. shopping mall, the distributions of crowds and the counts provide managers useful information regarding the preferences of customers. Therefore, some researchers begin to explore density map-based counting methods. Boominathan et al. [18] propose a CrowdNet, which combines deep and shallow, fully convolutional networks to capture both high-level semantic information and low-level features for addressing the scene variations. The CrowdNet also tackles the varying scales and inherent difficulties in high dense crowds by augmenting the training images. [7] presents a scale-aware solution named Hydra CNN to learn the non-linear regressor to generate the density maps from a pyramid of image patches at multiple scales. Zhang et al. [4] also focus on the scale problem by designing a multi-column CNN (MCNN), which use several CNN branches with different receptive fields to extract multi-scale features. This work is further extended to address scenario variations [19], density variations [8] and complex backgrounds [5]. Marsden et al. [19] perform a multi-scale averaging step during inference to overcome the scale and perspective issues. Taking the MCNN [4] as a basic network, Sam et al. [8] design a switch to assign a best regressor(a particular CNN branch if the MCNN) to crowd scene patch. The results demonstrate that the switch relays the patch to a particular CNN column based on the density of the crowds. [5] explicitly incorporate global and local contextual information by a contextual pyramid CNN (CP-CNN), which consists of four modules: global context estimator (GCE), local

context estimator (LCE), density map estimator (DME) and a fusion-CNN (FCNN). The authors use a variant of the MCNN [4] for estimating density maps and a FCNN for fusing the features extracted from the GCE, the LCE and the DME. In our previous work [20], we make the first attempt to add a single attention model to the MCNN to guide the network to focus on head locations, thus improving the counting accuracy obviously. Two U-nets are used in [21] as both large and small generators, which estimate density maps for large image and separated patches to address scale variations.

The success of multi-task learning for various computer vision tasks [22][23] inspired researchers to combine counts estimation or density map prediction with other tasks, e.g. density level and appearance classification. [24] and [25] classify the density level and appearance respectively during predicting the counts. These approaches exploit a sub-task to extract different contextual information for crowd counting. [3] aims to address cross-scene crowd counting with a switchable learning approach and two related learning objectives: crowd density map and crowd count. Zhao et al.[26] adopt a two-phase training scheme to decompose the crowd counting into two sub-tasks: density map prediction and crowd velocity map estimation. This method is robust to variations of crowd density, velocity and direction of line-of-interest(LOI). Sindagi et al. [27] propose a cascaded framework for both density level classification and density map estimation, which is also the basic framework of the MRA-CNN. The cascaded framework learns global relevant discriminate features by incorporating a high-level prior, thus enabling it to account for large count variations. The crowd ranking network in [28] simultaneously ranks images and estimates crowd density maps. The ranking task address the lack of training samples.

3. The proposed method

As the the proposed architecture in Fig.2 shows, a stack of two convolution layers are used to firstly extract low level features, and followed with two feature extraction branches with different kernel sizes. The branch with larger kernel size (Point A to Point B in Fig.2) is used for both density level classification and density map estimation while the other branch (Point A to Point C in Fig.2) is only used for density map estimation. The density level classification task is finally achieved after several fully connected layers. Simultaneously, three atrous convolution layers and a MRA model is used to generate density maps for crowd counting based on the fused features.

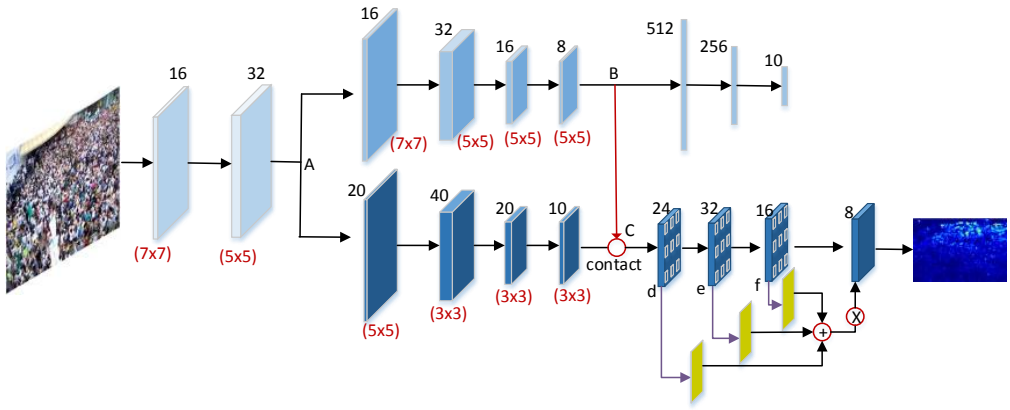

Figure 2: Architecture of the MRA-CNN

3.1. Architecture of the MRA-CNN

This section presents the detail of the MRA-CNN architecture. To introduce the structure of the network succinctly, we use simple notations to represent the parameters in different layers. 1.$Conv(o, k, p)$ and $A-Conv(o, k, p)$ for the traditional and atrous convolution layer with o outputs, kernel size k and padding p. 2.$Pool(k, s)$ represents the max pooling layer with kernel size k and stride s. 3.$AdapPool(h,w)$ stands for adaptive max pooling layer which outputs the features with size of $h * w$. 4.$ReLU$ for the activation function: rectified linear unit. 5.$FC(o)$ represents the fully connected layer with o outputs. Firstly, a stack of two convolution layers are used to extract low level features from the image (Input to Point A in Fig.2). The parameters can be represented as: $Conv(16, 7, 5) - ReLU - Conv(32, 5, 3) - ReLU$. Then the low level features are fed into 2 convolutional branches. The branch used for both 2 tasks has larger kernel sizes, which are: $Conv(16, 7, 5) - ReLU - Pool(2, 2) - Conv(32, 5, 3) - ReLU - Pool(2, 2) - Conv(16, 5, 3) - ReLU - Conv(8, 5, 3) - ReLU$. The trained feature maps are fed into the fully connected layers for density level classification and a higher CNN branch for density map estimation(Point B in Fig.2). Since the image sizes in some datasets are varied, we add an adaptive max pooling layer to get size-fixed features and fed them to a fully connected layer. The detail of the fully connected layers are: $AdapPool(32, 32) - FC(512) - FC(256) - FC(10)$. The

counting focused CNN branch can be represented as: $Conv(20,5,3)-ReLU-Pool(2,2)-Conv(40,3,1)-ReLU-Pool(2,2)-Conv(20,3,1)-ReLU-Conv(10,3,1)-ReLU$. After the feature fusion process (Point C in Fig.2), we use atrous convolution layers to further enlarge the receptive field with fewer parameters, which can be represented as: $A-Conv(24,3,3)-ReLU-A-Conv(32,3,1)-ReLU-A-Conv(16,3,1)-ReLU-Conv(8,3,1)-ReLU$. In addition, 3 score maps are generated from the atrous convolution layers by attention models. The score maps are then summed to guide the final feature layer to focus on head regions. The MRA model and atrous convolution layer will be introduced in Section3.1.1 and Section3.1.2 .

3.1.1. The MRA model

As aforementioned, the ground-truth density map is generated according to head locations. In the training and testing process, the head regions are expected to have larger values in the estimated density map. In order to guide the network to focus on the head regions, we exploit three attention models in the high-level feature extraction layers.

The attention model has been demonstrated effective for pixel-wise computer vision tasks, e.g. object classification[29], image classification[30] and image segmentation[31]. Inspired by [31] and [32], we utilize attention model to measure how much attention to pay to different regions in the feature maps. Suppose the convolution feature maps as F, then the soft attention is generated as:

$$S = \varphi(W \odot F + b) \tag{1}$$

Where φ donates a nonlinear activation, $\odot$ is convolution function. Then the attention model predicts how much attention to pay to different locations with a softmax operation applied to S:

$$M_p = \frac{e^{S_p}}{\sum_{p' \in P} e^{S_{p'}}} \tag{2}$$

M_p measures the probability of presenting head region in pixel p.

To further improve the accuracy of the score map, we totally conduct three attention measurement operation in convolution layers d, e and f, as shown in Fig.2), which is defined as the MRA model. The three score maps are summed as $\mathbb{M}$. Finally, to guide the network to focus on head regions

with the summed score map $\mathbb{M}$, we conduct element-wise product on the final feature maps:

$$F^{att} = F \otimes \mathbb{M} \tag{3}$$

To this end, the MRA model could adaptively emphasize the relevant regions where the heads are presented and assign these regions higher weights. This makes the MRA model very suitable for density map estimation. Section 4.3.2 will illustrates some representative score maps to demonstrate the effectiveness of the MRA model.

3.1.2. The Atrous convolution layer

As aforementioned in Section 3.1.1, we use a sum of three score maps which are generated from a series of convolution layers to guide the network to focus on head regions. These layers are expected to extract high-level features (which have larger receptive fields) from the fused feature maps. In addition, the three score maps are also expected to have larger gap of resolutions, which could further improve the head location prediction accuracy. Therefore, we exploit atrous convolution in the high-level feature extraction process. Atrous convolution has demonstrated its significant performance in

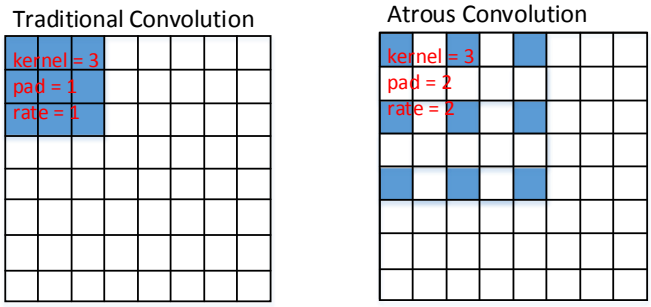

Figure 3: Illustration of traditional and atrous convolution

pixel-wise prediction tasks, such as object detection [33], semantic segmentation [34] and image segmentation[35] [36]. It allows to enlarge the receptive field without increasing the number of parameters. Fig. 3 illustrates the receptive field of traditional and atrous convolution with kernel size 3×3. As

Fig.3(b) shows, the receptive field of the atrous convolution becomes larger by inserting some "holes" in the convolution kernel.

In this paper, the setting showed in Fig. 3(b) is used for the 3 high-level feature extraction layers (d,e,f in Fig. 2). By using the atrous convolution, the receptive field gap in the 3 layers becomes larger, which provide more comprehensive information for generating score maps and estimating the density maps.

3.2. Loss Function

There are two tasks in the MRA-CNN, each one corresponds to a loss function. As most of the density map generation-based counting methods did, we choose Euclidean distance as the loss function for density map estimation task, which can be formulated as:

$$L_d = \frac{1}{N}\sum_{i=1}^{N}(F(X_i,\Theta) - D_i)^2 \tag{4}$$

Where N is the number of the training samples, D is the ground-truth density map of the ith sample and F is the function that mapping the input X_i to the estimated density map with parameters Θ. Cross-entry loss is used for density level classification, which is:

$$L_c = \sum_{i=1}^{N} -\log\frac{\exp(y_{class})}{\sum_c^C \exp(y_c)} \tag{5}$$

where $y(class)$ represents the predicted probability of the input belonging to $class$ and C is the number of classes.

To make a balance of training speed for the two tasks, we set a balance weight to the cross-entry loss, and the final loss function of the MRA-CNN can be represented as:

$$L = L_d + \alpha L_c \tag{6}$$

α is set as 0.001 according to the experiments.

4. Experiments

In order to evaluate the effectiveness of the proposed method, ShanghaiTech [4], WorldExpo'10 [3] and UCF [37] datasets are employed for experimental results. We shall firstly introduce the detail of the datasets and

then present how to obtain ground-truth density map. Experimental results as well as the performance comparison with recent methods are finally provided to demonstrate the effectiveness of the MRA-CNN.

4.1. Datasets

We summarize the 3 datasets in Table 1, where "Min", "Max" and "Avg" represent the minimum, maximum and average numbers of people in examples of the dataset.

Table 1: Summary of crowd counting datasets

Dataset	No. of training samples	No. of testing samples	Resolution	Min	Max	Ave
ShanghaiTech-A[4]	300	182	Varied	33	3139	501
ShanghaiTech-B[4]	400	316	768 * 1024	9	578	123
WorldExpo'10[3]	3380	120	576 * 720	1	253	50
UCF[37]	50		Varied	94	4543	1279

ShanghaiTech dataset [4] contains both dense (part A) and sparse (part B) crowd examples, and it provides separated sets for training and testing. The resolution of the images is varied in part A and constant in part B. WorldExpo'10 [3] is the largest one focusing on cross-scene crowd counting. The density of crowds in this dataset is the lowest among these 3 datasets. It totally provides 600 images which are divided into 5 sub-sets for testing, each contains a particular scene. In addition, region of interest (ROI) maps are also provided by the WorldExpo'10 dataset, and we utilize them referring to [3]. The images in this dataset are captured in Shanghai 2010 WorldExpo and cover a large variety of scenes. The UCF dataset mainly focuses on dense crowd counting. The scenes in this dataset belong to a diverse set of events, e.g. concerts, protests, stadiums, marathons and pilgrimages. Besides, the counts vary greatly and the number of samples is limited, making the dataset more challenging. We conduct 5-fold cross-validation on this dataset, for each validation, forty of the images are used as training samples and the other ten are used for testing.

4.2. Ground-truth density maps generation

We convert the labelled head locations in the original image into ground-truth density map following the method in [4]. An adaptive Gaussian kernel

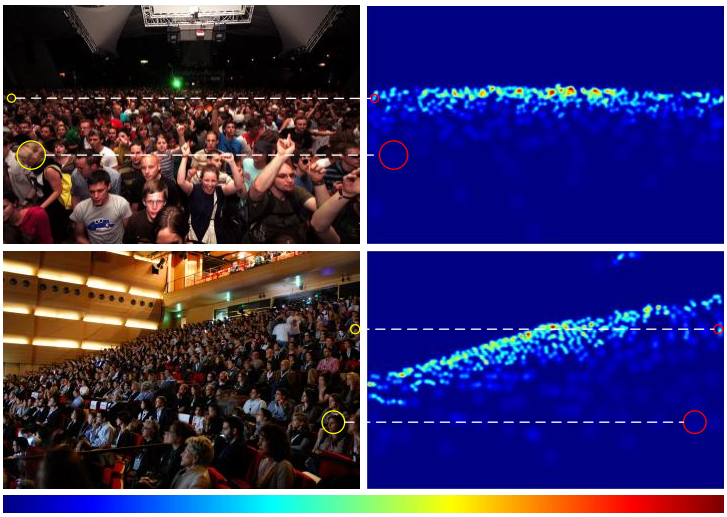

Figure 4: Representative density maps

(which is normalized to 1) computed by k-nearest neighbors (KNN) according to the average distance between the object and its 3 neighbors is covered on each head location. The generation method can be formulated as:

$$D(x) = \sum_{i=1}^{N} \delta(x - x_i) * G_{\sigma_i}(x_i) \tag{7}$$

x_i stands for the head location, G$(\cdot)$ is the Gaussian function and σ_i is the variance of the Gaussian kernel. $\delta(\cdot)$ is an impulse function. Fig. 4 shows two density maps generated from ShanghaiTech-A dataset. The color bar in the bottom of this figure illustrates the values in the density maps, where values get larger from left to right. The Gaussian kernel in the red circle of the density map is generated from the head in yellow circle of the original images, which is linked by a white dotted line. As each Gaussian kernel is normalized to 1, the one generated from larger head will also have larger size but small values in each location. However, the adaptive Gaussian kernel method is not applicable to sparse crowd examples since the distance between 2 objects varies greatly. We have 2 sparse crowd datasets: WorldExpo'10 and ShanghaiTech-B datasets. For the former one, σ_i is defined as $0.2 * P_i$ where P is the perspective map provided by this dataset. For the later one, we set σ_i as 4.

4.3. Experimental results

The experimental settings are as follows. 50 patches with 1/4 image sizes are cropped from each image for training. The testing images are fed into the well trained model directly for evaluation. Two commonly used standard metrics: the mean absolute error(MAE) and the mean squared error (MSE) are employed to evaluate the performance. The two metrics are defined as:

$$\begin{aligned} MAE &= \frac{1}{N}\sum_{i\in N}|y_i - y_i^{'}|, \\ MSE &= \sqrt{\frac{1}{N}\sum_{i\in N}(y_i - y_i^{'})^2} \end{aligned} \tag{8}$$

Where y_i stands for the ground-truth count and $y_i^{'}$ is the estimated count for the i-th sample.

4.3.1. Comparison to prior study on crowd counting

In order to effectively assess the performance of our algorithm, we compare it with existing state-of-the-arts algorithms. Table 2 and Table 3 present experimental results on sparse (WorldExpo'10[3] and ShanghaiTech-B[4]) and dense (ShanghaiTech-A[4] and UCF [37]) crowd examples, respectively[1]. As all of previous works did, we only give the MAE of WorldExpo'10 dataset for comparison.

For sparse crowd examples, the MRA-CNN gets the-state-of-the-art performance. Zhang et al. [3] focus on cross-scene crowd counting, their network outputs both counts and density maps. It is also the first work to use CNN for density map generation. [4] presents a MCNN with different receptive fields in each feature extraction branch, thus being robust to head size variations. Based on [4], Sam et al. [8] add a switch classifier to assign a best regressor to an image and get better performance than the MCNN. Sindagi et al. [5] take a variation of the MCNN [4] as density map estimator and combine global and local contextual information with multi-scale features. An adversarial loss is utilized generate high-quality density map, thus getting a significant improvement, especially on ShanghaiTech-B dataset. The MAE/MSE of [5]

[1] $'-'$ means that results on the dataset are not reported in that paper.

is 2.74/− and 6.3/11.2 lower than [4] the WorldExpo' 10 and ShanghaiTech-B datasets, respectively. [27] utilizes multi-task learning, which is the basic framework of the MRA-CNN. The authors combine features learned from different tasks, and the results on ShanghaiTech-B dataset is close to [5]. The AM-CNN[20] adds a single attention model on the MCNN to guide the network to focus on head locations and gets better results than the MCNN. However, the single attention model works not so good as the MRA-CNN, which use multi-resolution attention model. The MAS/MSE of the AM-CNN is 0.34/− and 3.7/5.1 higher than the MRA-CNN on the WorldExpo' 10 and ShanghaiTech-B datasets. Except for the adversarial loss, [21] use two U-nets to estimate density maps for both image and patches to address the scale variation. The MAE/MSE of this method is 1.36/− and 2.9/2.7 lower than the CP-CNN[5], which also utilize adversarial loss. The image ranking task in [28] assists the network perform well on ShanghaiTech-B dataset, with low MAE/MSE of 13.7/21.4. Apart from using multi-task learning to utilize multi-scale features, the proposed method could focus on head regions accurately by using the MRA model. In addition, the atrous convolution layer expand the receptive fields with fewer parameters, which not only gets high-level features but also expands the receptive field gap of the score maps, making the MRA model focus on head locations more accurately. All of these contributions make MRA-CNN get the state-of-the-art results on both of the 2 datasets: the MAE/MSE of the proposed method on the WorldExpo'10 and ShanghaiTech-B datasets are 7.5/− and 11.9/21.3.

Table 2: Results on Sparse Crowd examples

Dataset	WorldExpo' 10 (MAE)						ShanghaiTech-B	
Method	S1	S2	S3	S4	S5	Ave	MAE	MSE
Cross-Scene[3]	9.8	14.1	14.3	22.2	3.7	12.9	32.0	49.8
MCNN[4]	3.4	20.6	12.9	13.0	8.1	11.6	26.4	41.3
Switching-CNN[8]	4.4	15.7	10.0	11.0	5.9	9.4	21.6	33.4
CP-CNN[5]	2.9	14.7	10.5	10.4	5.8	8.86	20.1	30.1
AM-CNN[20]	2.5	13.0	9.7	**10.0**	4.0	7.84	15.6	26.4
Cascaded-MLT[27]	-	-	-	-	-	-	20.0	31.1
ACSCP[21]	-	-	-	-	-	**7.5**	17.2	27.4
Rank-Count[28]	-	-	-	-	-	-	13.7	21.4
MRA-CNN	**2.4**	**11.4**	**9.3**	10.5	**3.7**	**7.5**	**11.9**	**21.3**

ShanghaiTech-A and UCF datasets are both dense crowd datasets, and the proposed method performs better than most of the-state-of-the-arts algorithms on ShanghaiTech-A. The MoCNN [17] is a counts-only method. Although using a single-task style algorithm, it learns the appearance of the crowds and assign the appearance-weights to different expert CNNs during crowd counting, thus getting better performance than its previous algorithms [3] [4] on UCF datasets. The AM-CNN[20] adds a single attention model on the MCNN[4] and gets great improvements, with 22.9/40.5 and 98.1/131.3 lower MAE/MSE than the MCNN on ShanghaiTech-A and UCF datasets. [7] [27] and [5] utilize multi-contextual information based on multi-scale features and get better performances than previous algorithms [3] [17]. Apart from using multi-contextual information, [5] makes a combination of adversarial loss and pixel-level Euclidean loss for higher-quality density maps, and performing better than [7] and [27]. Besides the adversarial loss, a combination of large and small density map generators for image and patches makes the ACSCP (Adversarial Cross-Scale Consistency Pursuit) network[21] get 4.8/17.3 lower MAE/MSE on the UCF dataset than [5]. Thanks to the auxiliary image ranking task, the method in [28] performs best on ShanghaiTech-A dataset, with the MAE of 72.0. ShanghaiTech-A dataset provides a large number of dense crowd examples, the proposed method performs better than most of the existing methods on this dataset. The performance of the MRA-CNN is a little worse (MAE/MSE is 2.2/5.9 higher) than [28], however still competitive. For UCF dataset, which is a dataset with limited samples, the proposed method performs best among these methods.

4.3.2. Multi-resolution Attention vs. None Multi-resolution Attention

One of the important ideas of the proposed method is the ability of the MRA model. Therefore, it is necessary to compare the performances of the method with and without MRA model. We remove the MRA model from the MRA-CNN while reserve the atrous convolution layers and test it on ShanghaiTech dataset since it contains both sparse and dense crowd samples. In addition, we use a single attention model in the last feature extraction layer to compare with the MRA-CNN and thus demonstrating the necessary of using MRA model. Table 4 displays the comparison results. By using a single attention model, the performances on these two datasets get better, with the MAE/MSE 4.5/2.5 and 0.6/0.2 lower than that without attention model. However, only a single attention model could not utilize more comprehensive information. When use the MRA model, the network could focus on head

Table 3: Results on dense crowd examples

Dataset	ShanghaiTech-A		UCF	
Method	MAE	MSE	MAE	MSE
Cross-Scene[3]	181.8	277.7	467.0	498.5
MCNN[4]	110.2	173.2	377.6	509.1
MoCNN[17]	-	-	361.7	493.3
Hydra-CNN[7]	-	-	333.7	425.2
Cascaded-MLT[27]	101.3	152.4	322.8	397.9
Switching-CNN[8]	90.4	135.0	318.1	439.2
CP-CNN[5]	73.6	106.4	295.8	421.3
AM-CNN[20]	87.3	132.7	279.5	377.8
ACSCP[21]	75.7	**102.7**	291	404
Rank-Count[28]	**72.0**	106.6	279.6	388.9
MRA-CNN	74.2	112.5	**240.8**	**352.6**

regions more accurately. The performances of the proposed method are further improved by using the MRA model, with the MAE/MSE 11.3/20.0 and 1.5/1.6 lower than that of with only one attention model on the ShanghaiTech part A and part B datasets, respectively.

Table 4: Multi-resolution Attention vs. None Multi-resolution Attention

Dataset	ShanghaiTech-A		ShanghaiTech-B	
Method	MAE	MSE	MAE	MSE
Method w/o MRA model	90.0	130.5	14.0	23.1
Method with single attention model	85.5	132.5	13.4	22.9
Method with MRA model	**74.2**	**112.5**	**11.9**	**21.3**

To visualize the ability of the MRA model, we display the score maps generated from different atrous convolution layers on Fig. 5 and Fig. 6. The first row illustrates test images. Rows $2-4$ are score maps generated from the last 3 atrous convolution layers. To illustrate the score maps clearly, we resize these score maps and overlay them on the original images. The transparency is set as 0.7 to display both the score map and the original image clearly. Figures in the last two rows are estimated and ground-truth density maps, respectively. As this figure shows, the score maps get clearer

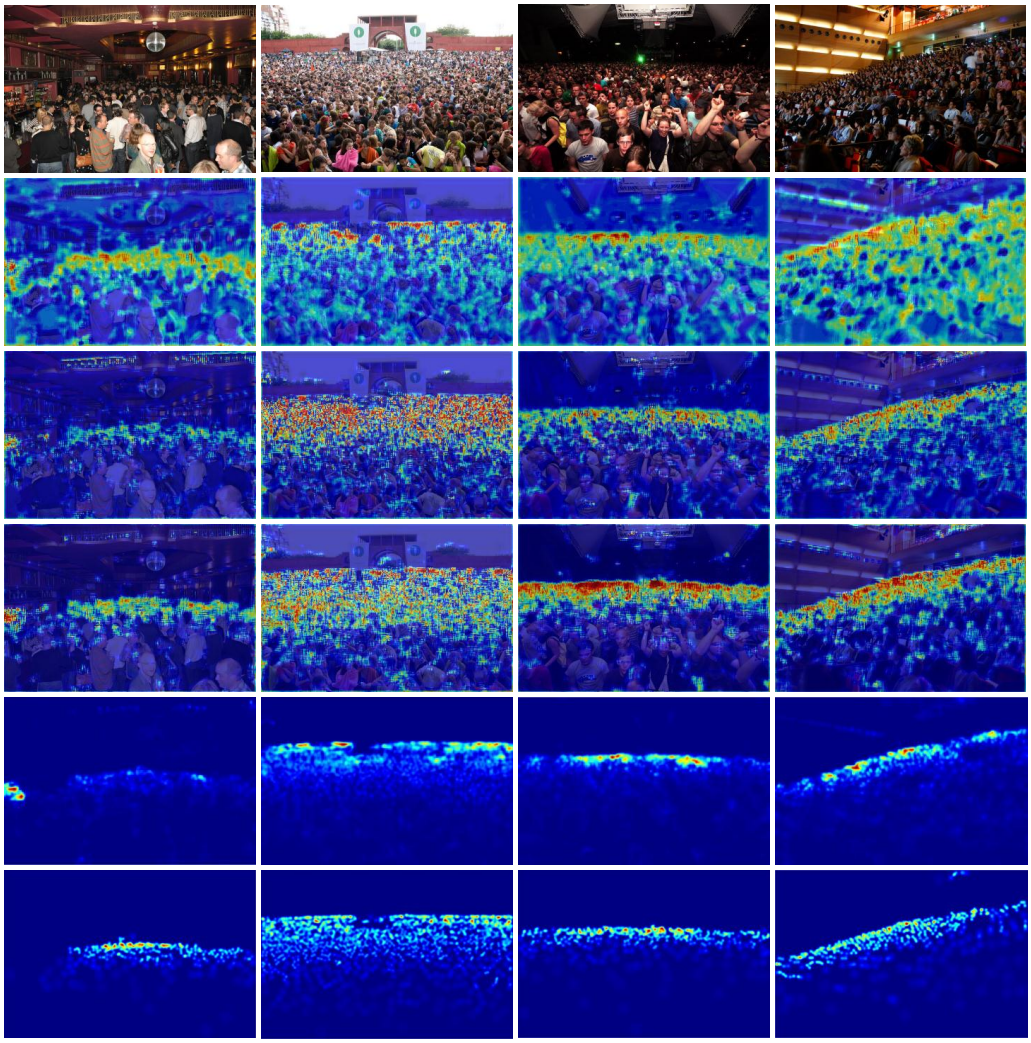

Figure 5: Representative samples of ShanghaiTech-A dataset

as the atrous convolution layer gets deeper. The score maps of the first attention model contain much backgrounds, especially the structured ones. The score maps of the second attention model filter more backgrounds than the first one. The third attention model generate clearer score maps on Shanghai-A dataset while get similar score maps with the second attention model on ShanghaiTech-B. Someone may argue that why not only use the

Figure 6: Representative samples of ShanghaiTech-B dataset

third attention model, which generates the most clear score map. On one hand, although the first two attention models get worse score maps than the

third one, they still emphasize head locations comparatively. The sum of the three attention maps can further emphasize the head locations. On the other hand, the loss generated by the first two attention models can guide the corresponding atrous convolution layer to learn better features, which also resulting in better features on the third atrous layer. The experimental results in Table 4 also demonstrate the necessary of the MRA model.

4.3.3. Atrous Convolution Layer vs. Traditional Convolution Layer

For another contribution of the proposed method: atrous convolution layer, we also conduct the comparison based on ShanghaiTech dataset. In this experiment, we replace the atrous convolution layer with traditional convolution layer and reserve the MRA model. As aforementioned, the atrous convolution gets larger receptive fields without increasing the number of parameters. By using atrous convolution in the last feature extraction layers, the network can generate higher-level features and provide the attention models more comprehensive information. Table 5 shows the comparison results.

Table 5: Atrous convolution layer vs. Traditional Convolution Layer

Dataset	ShanghaiTech-A		ShanghaiTech-B	
Method	MAE	MSE	MAE	MSE
Method w/o Atrous convolution layer	87.0	131.5	13.6	23.8
Method with Atrous convolution layer	**74.2**	**112.5**	**11.9**	**21.3**

As the results show, the performance (MAE/MSE) of the proposed method are 12.8/19.0 and 1.7/2.5 lower on ShanghaiTech-A (dense) and ShanghaiTech-B (sparse) than traditional CNN, respectively, demonstrating the effectiveness of the atrous convolution layer.

5. Conclusion

In this paper, we presented a MRA-CNN crowd counting algorithm, where an additional density-level classification task is utilized to learn multi-scale, multi-contextual information. In order to well exploit the head locations, a MRA model is used to generate score maps from different feature extraction layers and guide the network to emphasize head regions. In addition, we use atrous convolution layer to extract higher-level features with fewer

parameters as well as provide the MRA model more comprehensive information. Various experimental results and comparisons have demonstrated the superiority of the proposed method over the existing ones.

Acknowledgment

This work was supported in part by the National Natural Science Foundation of China under Grant61673244, Grant61273277 and Grant61703240) and was carried out at the Rapid-Rich Object Search (ROSE) Lab at the Nanyang Technological University, Singapore. The ROSE Lab is supported by the Infocomm Media Development Authority, Singapore.

References

[1] Y. Yuan, J. Wan, Q. Wang, Congested scene classification via efficient unsupervised feature learning and density estimation, Pattern Recognition 56 (2016) 159–169.

[2] C. Zhang, K. Kang, H. Li, X. Wang, R. Xie, X. Yang, Data-driven crowd understanding: A baseline for largr-scale crowd dataset, IEEE Transactions on Multimedia 18 (6) (2016) 1048–1061.

[3] C. Zhang, H. Li, X. Wang, X. Yang, Cross-scene crowd counting via deep convolutional neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 833–841.

[4] Y. Zhang, D. Zhou, S. Chen, S. Gao, Y. Ma, Single-image crowd counting via multi-column convolutional neural network, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 589–597.

[5] V. A. Sindagi, V. M. Patel, Generating high-quality crowd density maps using contextual pyramid cnns, in: 2017 IEEE International Conference on Computer Vision (ICCV), IEEE, 2017, pp. 1879–1888.

[6] V.-Q. Pham, T. Kozakaya, O. Yamaguchi, R. Okada, Count forest: Covoting uncertain number of targets using random forest for crowd density estimation, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 3253–3261.

[7] D. Onoro-Rubio, R. J. López-Sastre, Towards perspective-free object counting with deep learning, in: European Conference on Computer Vision, Springer, 2016, pp. 615–629.

[8] D. B. Sam, S. Surya, R. V. Babu, Switching convolutional neural network for crowd counting, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Vol. 1, 2017, p. 6.

[9] S. Yu, S. Jia, C. Xu, Convolutional neural networks for hyperspectral image classification, Neurocomputing 219 (2017) 88–98.

[10] W. Zhang, W. Zhang, K. Liu, J. Gu, A feature descriptor based on local normalized difference for real-world texture classification, IEEE Transactions on Multimedia 20 (4) (2018) 880–888.

[11] X. Sun, P. Wu, S. C. Hoi, Face detection using deep learning: An improved faster rcnn approach, Neurocomputing 299 (2018) 42–50.

[12] W. Zhang, X. Yu, X. He, Learning bidirectional temporal cues for video-based person re-identification, IEEE Transactions on Circuits and Systems for Video Technology.

[13] D. Fourure, R. Emonet, E. Fromont, D. Muselet, N. Neverova, A. Trémeau, C. Wolf, Multi-task, multi-domain learning: application to semantic segmentation and pose regression, Neurocomputing 251 (2017) 68–80.

[14] W. Zhang, B. Ma, K. Liu, R. Huang, Video-based pedestrian re-identification by adaptive spatio-temporal appearance model, IEEE transactions on image processing 26 (4) (2017) 2042–2054.

[15] W. Zhang, Q. Chen, W. Zhang, X. He, Long-range terrain perception using convolutional neural networks, Neurocomputing 275 (2018) 781–787.

[16] C. Shang, H. Ai, B. Bai, End-to-end crowd counting via joint learning local and global count, in: Image Processing (ICIP), 2016 IEEE International Conference on, IEEE, 2016, pp. 1215–1219.

[17] S. Kumagai, K. Hotta, T. Kurita, Mixture of counting cnns: Adaptive integration of cnns specialized to specific appearance for crowd counting, arXiv preprint arXiv:1703.09393.

[18] L. Boominathan, S. S. Kruthiventi, R. V. Babu, Crowdnet: A deep convolutional network for dense crowd counting, in: Proceedings of the 2016 ACM on Multimedia Conference, ACM, 2016, pp. 640–644.

[19] M. Marsden, K. McGuiness, S. Little, N. E. O'Connor, Fully convolutional crowd counting on highly congested scenes, arXiv preprint arXiv:1612.00220.

[20] Y. Zhang, C. Zhou, F. Chang, A. Kot, Attention to head loctions for crowd counting, arXiv preprint arXiv:1806.10287.

[21] Z. Shen, Y. Xu, B. Ni, M. Wang, J. Hu, X. Yang, Crowd counting via adversarial cross-scale consistency pursuit, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 5245–5254.

[22] L. Xu, J. Li, W. Lin, Y. Zhang, L. Ma, Y. Fang, Y. Yan, Multi-task rank learning for image quality assessment, IEEE Transactions on Circuits and Systems for Video Technology 27 (9) (2017) 1833–1843.

[23] Y. Yang, W. Hu, W. Zhang, T. Zhang, Y. Xie, Discriminative reverse sparse tracking via weighted multitask learning, IEEE Transactions on Circuits and Systems for Video Technology 27 (5) (2017) 1031–1042.

[24] Y. Hu, H. Chang, F. Nian, Y. Wang, T. Li, Dense crowd counting from still images with convolutional neural networks, Journal of Visual Communication and Image Representation 38 (2016) 530–539.

[25] Y. Zhang, F. Chang, M. Wang, F. Zhang, C. Han, Auxiliary learning for crowd counting via count-net, Neurocomputing 273 (2018) 190–198.

[26] Z. Zhao, H. Li, R. Zhao, X. Wang, Crossing-line crowd counting with two-phase deep neural networks, in: European Conference on Computer Vision, Springer, 2016, pp. 712–726.

[27] V. A. Sindagi, V. M. Patel, Cnn-based cascaded multi-task learning of high-level prior and density estimation for crowd counting, in: Advanced Video and Signal Based Surveillance (AVSS), 2017 14th IEEE International Conference on, IEEE, 2017, pp. 1–6.

[28] X. Liu, J. van de Weijer, A. D. Bagdanov, Leveraging unlabeled data for crowd counting by learning to rank, arXiv preprint arXiv:1803.03095.

[29] B. Zhao, X. Wu, J. Feng, Q. Peng, S. Yan, Diversified visual attention networks for fine-grained object classification, IEEE Transactions on Multimedia 19 (6) (2017) 1245–1256.

[30] T. Xiao, Y. Xu, K. Yang, J. Zhang, Y. Peng, Z. Zhang, The application of two-level attention models in deep convolutional neural network for fine-grained image classification, in: Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on, IEEE, 2015, pp. 842–850.

[31] L.-C. Chen, Y. Yang, J. Wang, W. Xu, A. L. Yuille, Attention to scale: Scale-aware semantic image segmentation, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 3640–3649.

[32] X. Chu, W. Yang, W. Ouyang, C. Ma, A. L. Yuille, X. Wang, Multi-context attention for human pose estimation, arXiv preprint arXiv:1702.07432 1 (2).

[33] J. Dai, Y. Li, K. He, J. Sun, R-fcn: Object detection via region-based fully convolutional networks, in: Advances in neural information processing systems, 2016, pp. 379–387.

[34] F. Yu, V. Koltun, Multi-scale context aggregation by dilated convolutions, arXiv preprint arXiv:1511.07122.

[35] J. Dai, K. He, Y. Li, S. Ren, J. Sun, Instance-sensitive fully convolutional networks, in: European Conference on Computer Vision, Springer, 2016, pp. 534–549.

[36] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, A. L. Yuille, Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs, IEEE transactions on pattern analysis and machine intelligence 40 (4) (2018) 834–848.

[37] H. Idrees, I. Saleemi, C. Seibert, M. Shah, Multi-source multi-scale counting in extremely dense crowd images, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 2547–2554.