Crowd Counting using Deep Recurrent Spatial-Aware Network

Lingbo Liu[1], Hongjun Wang[1], Guanbin Li[1*], Wanli Ouyang[2], Liang Lin[1]

[1] School of Data and Computer Science, Sun Yat-sen University, Guangzhou, China

[2] School of Electrical and Information Engineering, The University of Sydney, Sydney, Australia

{liulingb,wanghq8}@mail2.sysu.edu.cn, liguanbin@mail.sysu.edu.cn,
wanli.ouyang@sydney.edu.au, linliang@ieee.org

Abstract

Crowd counting from unconstrained scene images is a crucial task in many real-world applications like urban surveillance and management, but it is greatly challenged by the camera's perspective that causes huge appearance variations in people's scales and rotations. Conventional methods address such challenges by resorting to fixed multi-scale architectures that are often unable to cover the largely varied scales while ignoring the rotation variations. In this paper, we propose a unified neural network framework, named Deep Recurrent Spatial-Aware Network, which adaptively addresses the two issues in a learnable spatial transform module with a region-wise refinement process. Specifically, our framework incorporates a Recurrent Spatial-Aware Refinement (RSAR) module iteratively conducting two components: i) a Spatial Transformer Network that dynamically locates an attentional region from the crowd density map and transforms it to the suitable scale and rotation for optimal crowd estimation; ii) a Local Refinement Network that refines the density map of the attended region with residual learning. Extensive experiments on four challenging benchmarks show the effectiveness of our approach. Specifically, comparing with the existing best-performing methods, we achieve an improvement of 12% on the largest dataset WorldExpo'10 and 22.8% on the most challenging dataset UCF_CC_50.

1 Introduction

Crowd counting, which aims at estimating the total number of people in unconstrained crowded scenes, has received increasing research interests in recent years due to its potential application in many real-world scenarios, such as video surveillance [Xiong *et al.*, 2017] and traffic monitoring [Zhang *et al.*, 2017]. Despite recent progress, crowd counting remains a challenging issue, especially in the face of extreme occlusions, changes in lighting and camera perspective.

---
*Corresponding author is Guanbin Li.

Figure 1: Examples of unconstrained crowd scenes from the ShanghaiTech dataset. There are many challenges limiting the performance of crowd counting, i.e, diversity of camera setting, the large variation of scales and rotations of people.

In recent two years, deep convolutional neural networks have been widely used in crowd counting and have made substantial progress [Sindagi and Patel, 2017a; Onoro-Rubio and López-Sastre, 2016; Sam *et al.*, 2017; Xiong *et al.*, 2017; Zhang *et al.*, 2015]. The success of most existing deep models stems from the effective modeling of the scale variation of people/heads and the varying degrees of regional crowd density. However, all of these methods, without exception, obtain the density estimation of the whole image by merging the prediction results of a number of pre-designed fixed subnetwork structures (working on the whole image or image patches). Specifically, they are either proposed to fuses the features from multiple convolutional neural networks with different receptive fields to handle the scale variation of people groups [Zhang *et al.*, 2016], or directly divide crowd scene into multiple non-overlapping patches and provide a pool of regression networks for each patch selection [Sam *et al.*, 2017]. Although these strategies can, to a certain extent, improve the adaptability of the prediction to images of diverse density regions or people of various scales, the limited enumeration of a fixed number of tailor-designed networks or several sizes of receptive fields can not well cope with all the scenarios after all. Most importantly, none of these algorithms take into account the impact of different pose and photographic angles on crowd density estimation while crafting their network struc-

tures. As shown in Figure 1, the camera viewpoints in various scenes create different perspective effects and may result in large variation of scales, in-plane and out-plane rotation of people.

To address the aforementioned concerns, we propose a Deep Recurrent Spatial-Aware Network for crowd counting. The core of our network is a Recurrent Spatial-Aware Refinement (RSAR) module, which recurrently conducting refinement on an initial crowd density map through adaptive region selection and residual learning. Specifically, the RSAR module consists of two alternately performed components: i) a Spatial Transformer Network is incorporated in each LSTM step for simultaneously region cropping and warping, which allows the network to adaptively cope with the various degrees of congestion, people scale and rotation variation in the same scene; ii) a Local Refinement Network refines the density map of the selected region. In general, the main contributions of this work are three-fold.

- We provide an adaptive mode to simultaneously handle the effect of both scale and rotation variation by introducing a spatial transform module for crowd counting. To the best of our knowledge, we are the first to address the issue of the rotation variation on this task.

- We propose a novel deep recurrent spatial-aware network framework to recurrently select a region (with learnable scale and rotation parameters) from an initial density map for refinement, dependent on feature warping and residual learning.

- Extensive experiments and evaluations on several public benchmarks show that our proposed method achieves superior performance in comparison to other state-of-the-art methods. Specifically, compared with the existing best-performing methods, we achieve an improvement of 12% on the WorldExpo'10 dataset and 22.8% on the most challenging UCF_CC_50 dataset.

2 Related Work

Deep learning Methods for Crowd Counting: Inspired by the significant progress of deep learning on various computer vision tasks (such as image classification[Zhu *et al.*, 2017; Chen *et al.*, 2017] and salient object detection[Chen *et al.*, 2016; Li *et al.*, 2017]), many researchers also have attempted to adapt the deep neural network to the task of crowd counting and achieved great success. Most of the existing methods addressed the scale variation of people with multi-scale architectures. Boominathan et al. [Boominathan *et al.*, 2016] proposed to tackle the issue of scale variation using a combination of shallow and deep networks along with an extensive data augmentation by sampling patches from multi-scale image representations. HydraCNN [Onoro-Rubio and López-Sastre, 2016] proposed to learn a non-linear regression model which used a pyramid of image patches extracted at multiple scales to perform the final density prediction. A pioneering work was proposed by Zhang et al. [Zhang *et al.*, 2016], in which they utilized multi-column convolutional neural networks with different receptive fields to learn scale-robust features. Sam et al. [Sam *et al.*, 2017] proposed a Switching-CNN to map a patch from the crowd scene to one of the

three regression networks, each of which can handle the particular range of scale variation. CP-CNN [Sindagi and Patel, 2017b] proposed a Contextual Pyramid CNN to generate high-quality crowd density estimation by incorporating global and local contextual information into the multi-column networks. However, the above-mentioned methods have two significant limitations. First, as a neural network with the fixed static receptive field can only handle a limited enumeration of scale variation, these methods do not scale well to large scale change and cannot well cope with all level of scale variation in diverse scenarios. Second, they did not take the rotation variation of people into consideration, which limits the models' robustness towards camera perspective variations. To the best of our knowledge, we are the first work to simultaneously address the issue of large scale and rotation variations in an adaptively learnable mode on this task.

Deep learning Methods for Crowd Counting: Spatial transformer Network (STN) [Jaderberg *et al.*, 2015] is a subdifferentiable sampling-based module, which is designed to spatially transform its input map to an output map that corresponds to a subregion of the input map and can be hence regarded as an effective region selection mechanism. It is convenient to incorporate a spatial transformer layer to the convolutional neural network and train it with the standard back-propagation algorithm. A parameter matrix is used to determine the location of the subregion, as well as its resize scale and the rotation angle. Recently, the spatial transformer has been applied to various computer vision tasks, e.g., multi-label image recognition [Wang *et al.*, 2017] and saliency detection [Kuen *et al.*, 2016]. To the best of our knowledge, our work is the first in successfully using multiple iterations of STN within a LSTM framework for crowd counting.

3 Proposed Method

As illustrated in Figure 2, we propose a novel Deep Recurrent Spatial-Aware Network for crowd counting, which is composed of two modules, including a Global Feature Embedding (GFE) module and a Recurrent Spatial-Aware Refinement (RSAR) module. Specifically, the GFE module takes the whole image as input for global feature extraction, which is further used to estimate an initial crowd density map. And then the RSAR module is applied to iteratively locate image regions with a spatial transformer-based attention mechanism and refine the attended density map region with residual learning. For convenience in the following, we denote the crowd density map in the i-th iteration as M_i. Noted that M_0 is the initial crowd density map.

3.1 Global Feature Embedding

The goal of the Global Feature Embedding module is to transform the input image into high-dimensional feature maps, which is further used to generate an initial crowd density map of the image. Inspired by previous works [Zhang *et al.*, 2016; Sindagi and Patel, 2017b], we develop our GFM module in a multi-scale mode. As shown in Figure 3(a), the GFE module is composed of three columns of CNNs, each of which has seven convolutional layers with different kernel sizes and channel numbers as well as three max-pooling
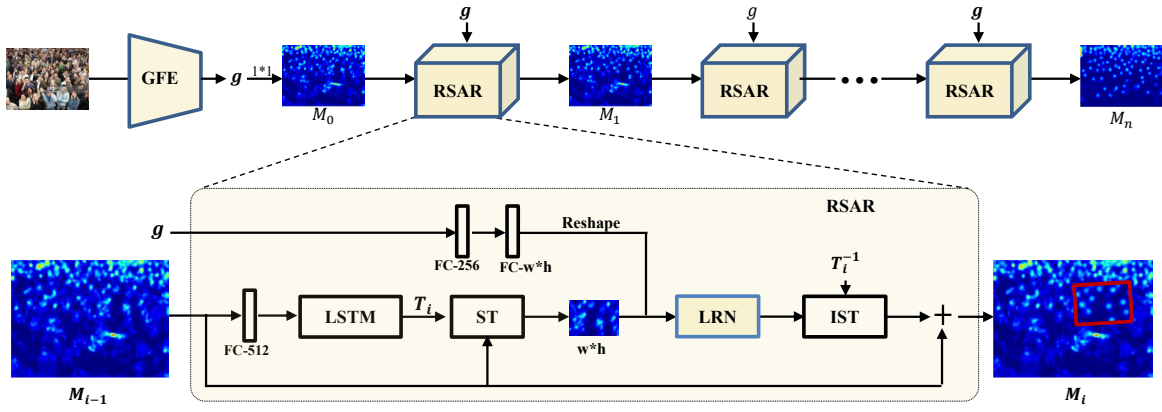
Figure 2: The architecture of the proposed Deep Recurrent Spatial-Aware Network for crowd counting. The Global Feature Embedding (GFE) module takes the whole image to extract its global feature g, which is used to estimate the initial crowd density map M_0 using a convolutional layer with a kernel size of 1×1. The Recurrent Spatial-Aware Refinement (RSAR) module iteratively selects an image region using spatial transformer network and refines the density map of the selected region incorporating global context with the Local Refinement Network (LRN). ST denotes the spatial transformer network and IST refers to the inverse spatial transformer network. FC-N stands for the fully-connected layer with N output neurons.

layers. Given an image I, we extract its global feature g by feeding it into GFM and concatenating the outputs of all the columns. After obtaining the global feature g, we generate the initial crowd density map M_0 of image I using a convolutional layer with a kernel size of 1×1, which can be expressed as:

$$g = GFM(I), M_0 = Conv(g), \qquad (1)$$

where $Conv$ denotes the convolution operation. As shown in Figure 4, the initial crowd density is too rough to estimate the number of people in the image. To further improve the quality of crowd density map, we propose a Recurrent Attentive Refinement (RSAR) module to iteratively refine the density map, which will be described in the next subsection.

3.2 Recurrent Spatial-Aware Refinement

In this section, we propose a novel Recurrent Attentive Refinement (RSAR) module to iteratively refine the crowd density map. Our proposed RSAR consists of two alternately performed components: i) a Spatial Transformer Network dynamically locates an attentional region from the crowd density map; ii) a Local Refinement Network refines the density map of the selected region with residual learning. A high-quality crowd density map with accurately estimated crowd number would be acquired after a refinement of n iterations.

Attentional Region Localization

In the i-th iteration, we first determine which image region to be refined with a spatial transformer network. We encode M_{i-1} into a 512-dimensions feature with a fully-connected layer and put it into a Long Short-Term Memory (LSTM) [Hochreiter and Schmidhuber, 1997] layer, which can be formulated as:

$$c_i, h_i = \textbf{LSTM}(c_{i-1}, h_{i-1}, FC(M_{i-1})), \qquad (2)$$

where c_i, h_i are the memory cell and hidden state of current iteration and FC is the fully-connected layer. The LSTM is utilized for capturing the past information of the density maps

and transformation actions. We define the transformation matrix T_i of the spatial transformer as:

$$T_i = \begin{bmatrix} \theta_{11}^i & \theta_{12}^i & \theta_{13}^i \\ \theta_{21}^i & \theta_{22}^i & \theta_{23}^i \end{bmatrix}, \qquad (3)$$

which allows cropping, translation, scale and rotation to be applied to the input map. We take the hidden state h_i to calculate the parameters of transformation matrix T_i with a fully-connected layers. Then we extract a region density map r_i from the whole density map M_{i-1} according to the transformation matrix T_i, which can be expressed as:

$$r_i = ST(M_{i-1}, T_i), \qquad (4)$$

where ST denotes the spatial transformer. The region map r_i will be resized to a given size $w * h$ by bilinear interpolation.

Region Density Map Refinement

Inspired by the previous works[Sindagi and Patel, 2017a; 2017b] which have validated the importance of global contextual modeling in crowd counting, we also take into consideration the global context when performing attended density region refinement. Intuitively, the global context contains the density level and density distribution of the given image. We encode the global feature g described in section 3.1 with two stacked fully-connected layers. The former one has 256 neurons and the latter one has $w * h$ neurons. We construct the global context map c_g by reshaping the output of the last FC layer to a new tensor with a size of $w * h$.

After obtaining the region map r_i and global context map c_g, we can perform crowd map refinement on the attentional region with a Local Refinement Network, which learns the residual between the region map r_i and the ground truth of its corresponding region. As shown in Figure 3(b), the Local Refinement Network also consists of three columns of CNNs. It takes the concatenation of r_i and c_g as the input and calculates the residual map of the attentional region. Finally, we can get a new crowd density map M_i by adding the
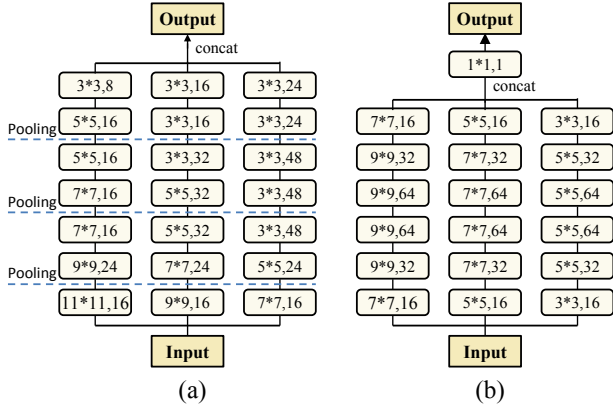
Figure 3: (a) The architecture of Global Feature Embedding module for the global feature extraction. (b) The architecture of Local Refinement Network for the region density map refinement. The block with text $k*k, N$ denotes a convolutional layer with $k*k$ kernel size and N output channels.

inverse transformed residual map to the M_{i-1}, which can be described as:

$$M_i = M_{i-1} + IST(LRN(r_i, c_g), T_i^{-1}), \quad (5)$$

$$T_i^{-1} = r * \begin{bmatrix} \theta_{22}^i & -\theta_{12}^i & \theta_{12}^i\theta_{23}^i - \theta_{13}^i\theta_{22}^i \\ -\theta_{21}^i & \theta_{11}^i & \theta_{13}^i\theta_{21}^i - \theta_{11}^i\theta_{23}^i \end{bmatrix}, \quad (6)$$

$$r = \frac{1}{\theta_{11}^i\theta_{22}^i - \theta_{12}^i\theta_{21}^i}, \quad (7)$$

where IST is the inverse spatial transformer that can transform the residual map back to the attentional region according to the inverse of T_i, denoted as T_i^{-1}.

3.3 Networks Optimization

For the existence of fully-connected layers in our model, we need to ensure that all the input images are of the same resolutions. Therefore, when optimizing the networks, we transform all training images or patches to the same resolutions and the details will be described in section 4. Suppose there are N training images after data augmentation, and each image I_i is annotated with a set of 2D points $L_i = \{l_1, ..., l_{C(i)}\}$, where $C(i)$ is the total number of the labeled pedestrian in the image. We generate ground truth density map D_i of image I_i with the specific strategy introduced by Lempitsky et al. [Lempitsky and Zisserman, 2010], which ensures that the sum of the ground truth density map $\sum_{p \in I_i} D_i(p)$ over the entire image I_i matches its crowd count with some negligible deviation.

Our networks are trained in an end-to-end manner and the overall loss function is defined as:

$$Loss = \sum_{i=1}^{N} \|M_i^0 - D_i\|^2 + \sum_{i=1}^{N} \|M_i^n - D_i\|^2, \quad (8)$$

where M_i^0 and M_i^n are the initial map and the refined density map for image I_i. We adopt the TensorFlow [Abadi et al., 2016] toolbox to implement our crowd counting network. The filter weights of all convolutional layers and fully-connected layers are initialized by truncated normal distribution with a deviation equal to 0.01. The learning rate is set to

Method	Part A		Part B	
	MAE	MSE	MAE	MSE
[Zhang et al., 2015]	181.8	277.7	32	49.8
[Zhang et al., 2016]	110.2	173.2	26.4	41.3
[Sindagi and Patel, 2017a]	101.3	152.4	20	31.1
[Sam et al., 2017]	90.4	135	21.6	33.4
[Sindagi and Patel, 2017b]	73.6	106.4	20.1	30.1
Ours	**69.3**	**96.4**	**11.1**	**18.2**

Table 1: Performance evaluation of different methods on the ShanghaiTech dataset. Our proposed method outperforms the existing state-of-the-art methods on both parts of the ShanghaiTech dataset with a margin.

10^{-4} initially and multiplied by 0.98 every 1K training iterations. The batch size is set to 1. We optimize our networks parameters with Adam optimization [Kingma and Ba, 2014] by minimizing the loss function Eq.(8).

4 Experiments

In this section, we first compare our method with recent state-of-the-art methods of the crowd counting task on four public challenging datasets. We further conduct extensive ablation studies to demonstrate the effectiveness of each component of our model.

4.1 Evaluation Metric

Following the existing work [Zhang et al., 2016; Sindagi and Patel, 2017b], we adopt the mean absolute error (MAE) and mean squared error (MSE) as metrics to evaluate the accuracy of crowd counting estimation, which are defined as:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}\|p_i - \hat{p}_i\|, MSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\|p_i - \hat{p}_i\|^2}, \quad (9)$$

where N is the total number of testing images, p_i and $\hat{p}_i$ are the ground truth count and estimated count of the i-th image respectively. As discussed in section 3.3, $\hat{p}_i$ can be calculated by summing over the estimated crowd density map.

4.2 Evaluations and Comparisons

ShanghaiTech [Zhang et al., 2016]. This dataset contains 1,198 images of unconstrained scenes with a total of 330,165 annotated people. And it is divided into two parts: Part A with 482 images crawled from the Internet, and Part B with 716 images taken from the busy shopping streets. For each part, we resize the images to its maximal resolution when training and testing.

We compare our method with five recently published state-of-the-art methods on this dataset, i.e., Zhang et al. [Zhang et al., 2015], MCNN [Zhang et al., 2016], Cascaded-MTL [Sindagi and Patel, 2017a], Switching-CNN [Sam et al., 2017], CP-CNN [Sindagi and Patel, 2017b]. As shown in Table 1, our proposed method outperforms other competing methods on both parts of the ShanghaiTech dataset. Specifically, our method achieves a significant improvement of 49.7% in MAE and 39.5% in MSE over the existing best-performing algorithm CP-CNN on Part B.

UCF_CC_50 [Idrees et al., 2013]. As an extremely challenging benchmark, this dataset contains 50 annotated images

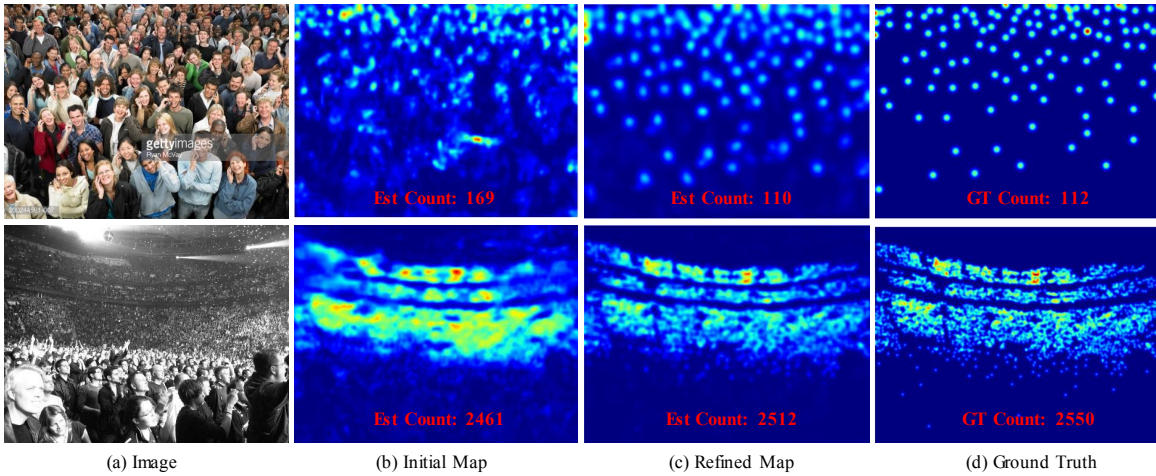| (a) Image | (b) Initial Map | (c) Refined Map | (d) Ground Truth |

Figure 4: Visualization of the crowd density maps generated by our method on ShanghaiTech dataset. We can see that more accurate crowd density map and estimated count are obtained after refinement.

Method	MAE	MSE
[Zhang *et al.*, 2015]	467	498.5
[Zhang *et al.*, 2016]	377.6	509.1
[Walach and Wolf, 2016]	364.4	341.4
[Sindagi and Patel, 2017a]	322.8	341.4
[Sam *et al.*, 2017]	318.1	439.2
[Sindagi and Patel, 2017b]	295.8	320.9
[Xiong *et al.*, 2017]	284.5	297.1
Ours	**219.2**	**250.2**

Table 2: Performance evaluation of different methods on the UCF_CC_50 dataset.

Method	MAE	MSE
[Chan *et al.*, 2008]	3.72	20.1
[Chen *et al.*, 2012]	3.59	19.0
[Chen *et al.*, 2013]	3.43	17.7
[Pham *et al.*, 2015]	2.50	10.0
[Zhang *et al.*, 2016]	2.24	8.5
Ours	**1.72**	**2.1**

Table 3: Performance evaluation on the MALL dataset.

of diverse scenes collected from the Internet. Except for different resolutions, aspect ratios and perspective distortions, the images of this dataset also suffer from a wide range of person counts, varying from 94 to 4,543. Following the standard protocol discussed in [Idrees *et al.*, 2013], we split the dataset into five subsets and perform a five-fold cross-validation. When training, we randomly crop some regions with a range of $[0.5, 0.9]$ from the original images and resize them to $1024 * 768$. The testing images are directly resized to the same resolution.

We perform comparison against several state-of-the-art methods based on deep-learning, including Zhang et al. [Zhang *et al.*, 2015], MCNN [Zhang *et al.*, 2016], Walach et al. [Walach and Wolf, 2016], Cascaded-MTL [Sindagi and Patel, 2017a], Switching-CNN [Sam *et al.*, 2017], CP-CNN [Sindagi and Patel, 2017b] and ConvLSTM-nt [Xiong *et al.*, 2017]. As shown in Table 2, our method achieves superior performance in comparison to other competing methods on UCF_CC_50 dataset. Specifically, our method outperforms the existing best-performing method ConvLSTM-nt by 22.8% over the MAE metric.

MALL [Chen *et al.*, 2012]. This dataset was captured by a publicly accessible surveillance camera in a shopping mall with more challenging lighting conditions and glass surface reflections. The video sequence consists of 2,000 frames with a resolution of $320 * 240$. Following the same setting as [Chen *et al.*, 2012], we use the first 800 frames for training

and the remaining 1,200 frames for evaluation. The images are kept the original resolution when training and testing.

We compare our method with Gaussian Process Regression [Chan *et al.*, 2008], Ridge Regression [Chen *et al.*, 2012], Cumulative Attribute Regression [Chen *et al.*, 2013], Count forest [Pham *et al.*, 2015] and ConvLSTM [Zhang *et al.*, 2016]. The results of all methods are summarized in Table 3. Our still-image based method outperforms the existing best method ConvLSTM, which utilized extra temporal dependencies from surveillance video to estimate the crowd count.

WorldExpo'10 [Zhang *et al.*, 2015]. This dataset contains 1,132 video sequences captured by 108 surveillance cameras during the Shanghai WorldExpo in 2010. The training set consists of 3,380 annotated frames from 103 scenes, while the testing images are extracted from other five different scenes with 120 frames per scene. The images are kept the original resolution for training and testing. When testing, the Region of Interest (ROI) maps for each scene are provided and we only consider the crowd count under the ROI.

We compare our method with seven recently published state-of-the-art methods, including Chen et al. [Chen *et al.*, 2013], Zhang et al. [Zhang *et al.*, 2015], MCNN [Zhang *et al.*, 2016], Shang et al. [Shang *et al.*, 2016], Switching-CNN [Sam *et al.*, 2017], CP-CNN [Sindagi and Patel, 2017b] and ConvLSTM [Xiong *et al.*, 2017]. As demonstrated in Table 4, our method achieves state-of-the-art performance with respect to the average MAE of five scenes. Specifically, our model gets the lowest MAE in Scene 2 and Scene 5, which are the two most challenging scenes in the testing set.

Method	Scene1	Scene2	Scene3	Scene4	Scene5	Average
[Chen *et al.*, 2013]	**2.1**	55.9	**9.6**	11.3	**3.4**	16.5
[Zhang *et al.*, 2015]	9.8	**14.1**	14.3	22.2	3.7	12.9
[Zhang *et al.*, 2016]	3.4	20.6	12.9	13	8.1	11.6
[Shang *et al.*, 2016]	7.8	15.4	14.9	11.8	5.8	11.7
[Sam *et al.*, 2017]	4.4	15.7	**10**	11	5.9	9.4
[Sindagi and Patel, 2017b]	2.9	14.7	10.5	**10.4**	5.8	**8.86**
[Xiong *et al.*, 2017]	7.1	15.2	15.2	13.9	**3.5**	10.9
Ours	**2.6**	**11.8**	10.3	**10.4**	3.7	**7.76**

Table 4: Mean Absolute Error of different methods on the WorldExpo'10 dataset. Our method achieves superior performance with respect to the average MAE of five scenes. The best results and the second best results are highlighted in red and blue, respectively. Best viewed in color.

Method	Part A		Part B	
	MAE	MSE	MAE	MSE
Basic	83.1	120.4	20.1	31.8
T	75.9	109.2	16.5	21.9
T + S	72.9	103.1	14.6	21.4
T + S + R	**69.3**	**96.4**	**11.6**	**19.5**

Table 5: Comparison of the performance of our model with different constraints of the spatial transformer on ShanghaiTech dataset. T, S, and R correspond to translation, scale, and rotation respectively.

Method	Part A		Part B	
	MAE	MSE	MAE	MSE
W/O Global Context	74.44	100.1	15.7	24.9
W/ Global Context	**69.3**	**96.4**	**11.6**	**19.5**

Table 6: Effectiveness verification of global context on the ShanghaiTech dataset.

Steps	Part A		Part B	
	MAE	MSE	MAE	MSE
$n = 0$	83.1	120.4	20.1	31.8
$n = 10$	82.1	111.3	18.8	30.4
$n = 20$	73.2	113.1	11.2	19.3
$n = 30$	**69.3**	**96.4**	**11.1**	**18.2**
$n = 40$	74.12	110.33	14.27	20.78

Table 7: ShanghaiTech dataset experimental results on the variants of our model using different refinement steps. Our method has the best performance when the density map is refined by $n = 30$ steps.

4.3 Ablation Study

In this section, we perform extensive ablation studies on ShanghaiTech dataset and demonstrate the effects of several components in our framework.

Effectiveness of Spatial Transformation: The transformation matrix defined in Eq.(3) allows translation, rotation and scale to be applied to the input map. To validate the effectiveness of each of them, we train our model with three different constraints of spatial transformer: i) spatial transformation only with translation by directly setting θ_{11}, θ_{22} to 1 and θ_{12}, θ_{21} to 0, denoted as T; ii) spatial transformation with translation and scale but without rotation by directly setting θ_{12}, θ_{21} to 0, denoted as $T + S$; iii) spatial transformation with translation, rotation and scale, denoted as $T + S + R$. The performance of these variants are summarized in Table 7 and the performance of initial density map is denoted as "Basic". We can see that directly locating a region without scale and rotation operation for refinement can reduce the MAE from 83.1 to 75.9 on ShanghaiTech Part A. When dynamically resizing the local regions, the MAE on Part A can be reduced to 73.2. When simultaneously resizing and rotating the local regions, we can get the best performance on ShanghaiTech dataset. The experiment has demonstrated the effectiveness of the spatial transformation for crowd density map refinement.

Effectiveness of Global Context: In order to validate the effectiveness of the global context, we train a variant of our model that refines the local region map without the global context. As shown in Table 6, the performance would drop over all metrics when removing the global context. Actually, the global context contains the information of density level and density distribution, which is beneficial for the crowd counting inference.

Effectiveness of Recurrent Refinement: To validate the superiority of the recurrent refinement mechanism of our method, we train our model with different refinement iterations. As shown in Table 7, the performance gradually increases with more iterations and it slightly drops until 40 steps. Our method gets the best performance with 30 iterations, therefore we set the refinement iteration n to 30 in our model for all datasets. Figure 4 shows the visual comparison with the refined maps and the initial maps and we can see that our method can generate more accurate crowd density maps after multi-step refinements.

5 Conclusion

In this paper, we introduce a novel Deep Recurrent Spatial-Aware Network for crowd counting which simultaneously models the variations of crowd density as well as the pose changes in a unified learnable module. It can be regarded as a general framework for crowd map refinement. Extensive experiments on four challenging benchmarks show that our proposed method achieves superior performance in comparison to the existing state-of-the-art methods. In our future research, we plan to delve into the research of incorporating our model in other existing crowd flow prediction framework.

Acknowledgments

References

[Abadi *et al.*, 2016] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, et al. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*, 2016.

[Boominathan *et al.*, 2016] Lokesh Boominathan, Srinivas SS Kruthiventi, and R Venkatesh Babu. Crowdnet: A deep convolutional network for dense crowd counting. In *ACM MM*, pages 640–644. ACM, 2016.

[Chan *et al.*, 2008] Antoni B Chan, Zhang-Sheng John Liang, and Nuno Vasconcelos. Privacy preserving crowd monitoring: Counting people without people models or tracking. In *CVPR*, pages 1–7. IEEE, 2008.

[Chen *et al.*, 2012] Ke Chen, Chen Change Loy, Shaogang Gong, and Tony Xiang. Feature mining for localised crowd counting. In *BMVC*, volume 1, page 3, 2012.

[Chen *et al.*, 2013] Ke Chen, Shaogang Gong, Tao Xiang, and Chen Change Loy. Cumulative attribute space for age and crowd density estimation. In *CVPR*, pages 2467–2474, 2013.

[Chen *et al.*, 2016] Tianshui Chen, Liang Lin, Lingbo Liu, Xiaonan Luo, and Xuelong Li. Disc: Deep image saliency computing via progressive representation learning. *IEEE TNNLS*, 27(6):1135–1149, 2016.

[Chen *et al.*, 2017] Tianshui Chen, Zhouxia Wang, Guanbin Li, and Liang Lin. Recurrent attentional reinforcement learning for multi-label image recognition. *arXiv preprint arXiv:1712.07465*, 2017.

[Hochreiter and Schmidhuber, 1997] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[Idrees *et al.*, 2013] Haroon Idrees, Imran Saleemi, Cody Seibert, and Mubarak Shah. Multi-source multi-scale counting in extremely dense crowd images. In *CVPR*, pages 2547–2554, 2013.

[Jaderberg *et al.*, 2015] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. In *NIPS*, pages 2017–2025, 2015.

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

[Kuen *et al.*, 2016] Jason Kuen, Zhenhua Wang, and Gang Wang. Recurrent attentional networks for saliency detection. *arXiv preprint arXiv:1604.03227*, 2016.

[Lempitsky and Zisserman, 2010] Victor Lempitsky and Andrew Zisserman. Learning to count objects in images. In *NIPS*, 2010.

[Li *et al.*, 2017] Guanbin Li, Yuan Xie, Liang Lin, and Yizhou Yu. Instance-level salient object segmentation. In *CVPR*, pages 247–256. IEEE, 2017.

[Onoro-Rubio and López-Sastre, 2016] Daniel Onoro-Rubio and Roberto J López-Sastre. Towards perspective-free object counting with deep learning. In *ECCV*, pages 615–629. Springer, 2016.

[Pham *et al.*, 2015] 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*, pages 3253–3261, 2015.

[Sam *et al.*, 2017] Deepak Babu Sam, Shiv Surya, and R Venkatesh Babu. Switching convolutional neural network for crowd counting. In *CVPR*, volume 1, page 6, 2017.

[Shang *et al.*, 2016] Chong Shang, Haizhou Ai, and Bo Bai. End-to-end crowd counting via joint learning local and global count. In *ICIP*, pages 1215–1219. IEEE, 2016.

[Sindagi and Patel, 2017a] Vishwanath A Sindagi and Vishal M Patel. Cnn-based cascaded multi-task learning of high-level prior and density estimation for crowd counting. In *AVSS*, pages 1–6. IEEE, 2017.

[Sindagi and Patel, 2017b] Vishwanath A Sindagi and Vishal M Patel. Generating high-quality crowd density maps using contextual pyramid cnns. In *ICCV*, pages 1879–1888. IEEE, 2017.

[Walach and Wolf, 2016] Elad Walach and Lior Wolf. Learning to count with cnn boosting. In *ECCV*, pages 660–676. Springer, 2016.

[Wang *et al.*, 2017] Zhouxia Wang, Tianshui Chen, Guanbin Li, Ruijia Xu, and Liang Lin. Multi-label image recognition by recurrently discovering attentional regions. In *ICCV*, 2017.

[Xiong *et al.*, 2017] Feng Xiong, Xingjian Shi, and Dit-Yan Yeung. Spatiotemporal modeling for crowd counting in videos. In *ICCV*, pages 5161–5169. IEEE, 2017.

[Zhang *et al.*, 2015] Cong Zhang, Hongsheng Li, Xiaogang Wang, and Xiaokang Yang. Cross-scene crowd counting via deep convolutional neural networks. In *CVPR*, pages 833–841, 2015.

[Zhang *et al.*, 2016] Yingying Zhang, Desen Zhou, Siqin Chen, Shenghua Gao, and Yi Ma. Single-image crowd counting via multi-column convolutional neural network. In *CVPR*, pages 589–597, 2016.

[Zhang *et al.*, 2017] Shanghang Zhang, Guanhang Wu, Joao P Costeira, and José MF Moura. Fcn-rlstm: Deep spatio-temporal neural networks for vehicle counting in city cameras. In *ICCV*, pages 3687–3696. IEEE, 2017.

[Zhu *et al.*, 2017] Feng Zhu, Hongsheng Li, Wanli Ouyang, Nenghai Yu, and Xiaogang Wang. Learning spatial regularization with image-level supervisions for multi-label image classification. In *CVPR*, pages 5513–5522, 2017.