ADCrowdNet: An Attention-injective Deformable Convolutional Network for Crowd Understanding

Ning Liu, Yongchao Long, Changqing Zou, Qun Niu, Li Pan, Hefeng Wu
School of Data and Computer Science, Sun Yat-sen University
{longych3, niuqun}@mail2.sysu.edu.cn

Abstract

We propose an attention-injective deformable convolutional network called ADCrowdNet for crowd understanding that can address the accuracy degradation problem of highly congested noisy scenes. ADCrowdNet contains two concatenated networks. An attention-aware network called Attention Map Generator (AMG) first detects crowd regions in images and computes the congestion degree of these regions. Based on detected crowd regions and congestion priors, a multi-scale deformable network called Density Map Estimator (DME) then generates high-quality density maps. With the attention-aware training scheme and multi-scale deformable convolutional scheme, the proposed ADCrowdNet achieves the capability of being more effective to capture the crowd features and more resistant to various noises. We have evaluated our method on four popular crowd counting datasets (ShanghaiTech, UCF_CC_50, WorldEXPO'10, and UCSD) and an extra vehicle counting dataset TRANCOS, our approach overwhelmingly beats existing approaches on all of these datasets.

1. Introduction

Crowd understanding has attracted much attention recently because of its wide range of applications like public safety, congestion avoidance, and flow analysis. The current research trend for crowd understanding has developed from counting the number of people to displaying distribution of crowd through density map. Generally, generating accurate crowd density maps and performing precise crowd counting for highly congested noisy scenes is challenging due to various complexities of crowd scenes caused by background noises, occlusions, and diversified crowd distributions.

Researchers recently have leveraged deep neural networks (DNN) for accurate crowd density map generation and precise crowd counting. Although these DNNs-based methods [32, 20, 24, 14] have made significant success in solving the above issues, they still have the problem of accuracy degradation when applied in highly congested noisy scenes. As shown in Figure 1, the state-of-art approach [14], which has achieved much lower Mean Absolute Error (MAE) than the previous state-of-the-art methods, is still severely affected by background noises, occlusions, and non-uniform crowd distributions.

In this paper, we aim at an approach which is capable of dealing with highly congested noisy scenes for the crowd understanding problem. To achieve this, we designed an attention-injective deformable convolutional neural network called ADCrowdNet which is empowered by a visual attention mechanism and a multi-scale deform-able convolution scheme. The visual attention mechanism is delicately designed for alleviating the effects from various noises in the input. The multi-scale deform-able convolution scheme is specially introduced for the congested environments. The basic principle of visual attention mechanism is to use the pertinent information rather than all available information in the input image to compute the neural response. This principle of focusing on specific parts of the input has been successfully applied in various deep learning models for images classification [10], semantic segmentation [19], image deblurring [18], and visual pose estimation [5], which also suits our problem where the interest regions containing the crowd need to be recognized and highlighted out from noisy scenes. The multi-scale deformable convolution scheme takes as input the information of the dynamic sampling locations, other than evenly distributed locations, which has the capability of modeling complex geometric transformation and diverse crowd distribution. This scheme fits well the nature of the distortion caused by the perspective view of the camera and diverse crowd distributions in real world, therefore guaranteeing more accurate crowd density maps for the congested scenes.

To incorporate the visual attention mechanism and deformable convolution scheme, we leverage an architecture consisting of two neural networks as shown in Figure 2. Our training contains two stages. The first stage generates an attention map for a target image via a network called Attention Map Generator (AMG). The second stage takes the out-

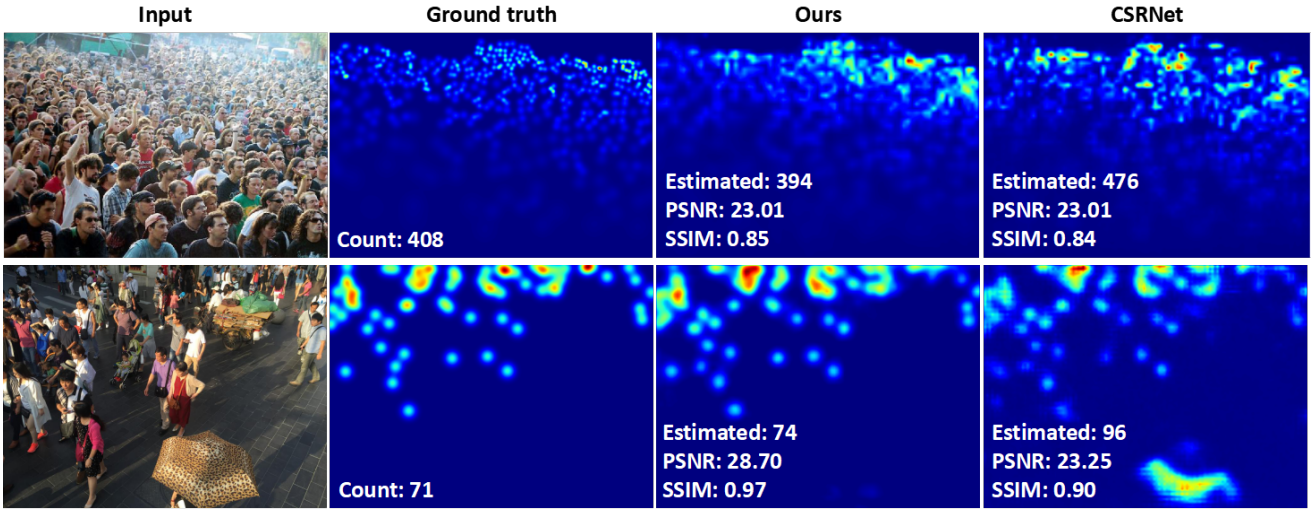

Figure 1. From left to right: a congested sample (top) and a noisy sample (bottom) from ShanghaiTech dataset [32], ground truth density map, and the generated density maps from the proposed ADCrowdNet and the state-of-the-art method [14]. ADCrowdNet outperforms the state-of-the-art method on both congested and noisy scenes.

put of AMG as input and generates the crowd density map via a network called Density Map Estimator (DME). The attention map generator AMG mainly provides two types of priors for the DME network: 1) candidate crowd regions and 2) the congestion degree of crowd regions. The former prior enables the multi-scale deformable convolution scheme empowered DME network to pay more attention to those regions having people crowds, and thus improving the capacity of being resistant to various noises. The latter prior indicates each crowd region with congestion degree (i.e., how crowded each crowd region is), which provides fine-grained congestion context prior for the subsequent DME network and boosts the performance of the DME network on the scenes containing diverse crowd distribution.

By taking advantage of such innovative structure, the proposed model ADCrowdNet outperforms the state-of-the-art crowd counting solution CSRNet [14] with 3.0%, 18.8%, 3.0%, 13.9% and 5.1% lower Mean Absolute Error (MAE) on ShanghaiTech Part_A, Part_B, UCF_CC_50, WorldExpo10, UCSD datasets, respectively. Apart from crowd counting, ADCrowdNet is also general for other counting tasks. We have evaluated ADCrowdNet on a popular vehicle counting dataset named TRANCOS [9], and ADCrowdNet achieves 32.8% lower MAE than CSRNet.

2. Related Work

Counting by detection: Early approaches of crowd understanding mostly focus on the number of people in crowds [8]. The major characteristics of these approaches are the sliding window based detection scheme and hand crafted features extracted from the whole human body or particular body parts with low-level descriptors like Haar wavelets [26] and HOG [7]. Generally, approaches in these groups deliver accurate counts when their underlying assumptions are met but are not applicable in more challenging congested scenes.

Counting by regression: Counting by regression approaches differs depending on the target of regression: object count [4, 3], or object density [13]. This group of approaches avoid solving the hard detection problem. Instead, they deploy regression model to learn the mapping between image characteristics (mainly histograms of lower level or middle level features) and object count or density. These approaches that directly regress the total object count discard the information of the location of the objects and only use 1-dimensional object count for learning. As a result, a large number of training images with the supplied counts are needed in training. Lempitsky *et al*. [13] propose a method to solve counting problem by modeling the crowd density at each pixel and cast the problem as that of estimating an image density whose integral over any image region gives the count of objects within that region. Since the ideal linear mapping is hard to obtain, Pham *et al*. [17] use random forest regression to learn a non-linear mapping instead of the linear one.

Crowd understanding by CNN: Inspired by the great success in visual classification and recognition, literature also focuses on the CNN-based approaches to predict crowd density map and count the number of crowds. Walach *et al*. [27] use CNN with a layered training structure. Shang *et al*. [21] adapt an end-to-end CNN which uses the entire images as input to learn the local and global count of the images and ultimately outputs the crowd count. [1] use a dual-column network combining shallow and deep layers to generate density maps. [32] proposes a multi-column CNN to estimate density map by exacting features at differ-

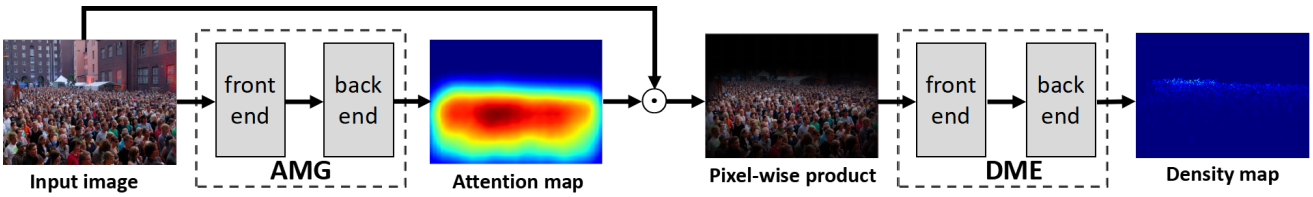

Figure 2. Architecture overview of ADCrowdNet. The well trained AMG generates the attention map of the input image. The pixel-wise product of the input image and its attention map is taken as the input to train the DME network.

ent scales. Similar idea is used in [16]. Marsden *et al*. [15] try a single-column fully convolutional network to generate density map while Sindagi *et al*. [23] present a CNN that uses high-level prior to boost accuracy.

More recently, Sindagi *et al*. [24] propose a multi-column CNN called CP-CNN that uses context at various levels to improve generate high-quality density maps. Li *et al*. [14] propose a model called CSRNet that uses dilated convolution to enlarge receptive fields and extract deeper features for boosting performance. These two approaches have achieved the state-of-the-art performances.

3. Attention-injective Deformable Convolutional Network

The architecture of the proposed ADCrowdNet method is illustrated in Figure 2. It employs two concatenated networks: AMG and DME. AMG is a classification network based on fully convolutional architecture for attention map generation, while DME is a multi-scale network based on deformable convolutional layers for density map generation. Before training DME, we train the AMG module with crowd images (positive training examples) and background images (negative training examples). We then use the well-trained AMG to generate the attention map of the input image. Afterward, we train the DME module using the pixel-wise product of input images and the corresponding attention maps. In the following sections, we will detail the architectures of the AMG and DME netwroks.

3.1. Attention Map Generator

3.1.1 Attention map

Attention map is an image-sized weight map where crowd regions have higher values. In our work, attention map is a feature map from a two-category classification network AMG which classifies an input image into crowd image or background image. The idea of using feature map to find the crowd regions in the input is motivated by an object localization work [33] which points out that the feature maps of classification network contain the location information of target objects.

The pipeline of the attention map generation is shown in Figure 3. F_c and F_b are the feature maps from the last convolution layer of AMG. W_c and W_b are the spatial average of the F_c and F_b after global average pooling (i.e., GAP in Figure 3). P_c and P_b are confidence scores of the predicted two class. They are generated by softmax from W_c and W_b. The attention map is obtained by up-sampling the linear weighted fusion of the two feature maps F_c and F_b (i.e., $F_c \cdot P_c + F_b \cdot P_b$) to the same size as the input image. We also normalize the attention map such that all element values fall in the range $[0, 1]$.

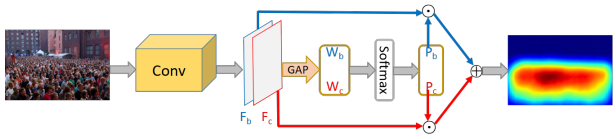

Figure 3. The pipeline of the attention map generation. Conv block denotes the AMG architecture shown in Fig. 5.

The attention map highlights the regions of crowds. In addition, it also indicates the degree of congestion in individual regions, i.e., higher congestion degree values indicate more congested crowds and lower values indicate less congested ones. Figure 4 illustrates the effect of attention maps at different density levels. The pixel-wise product between the attention map and the input image produces the input data used by the DME network.

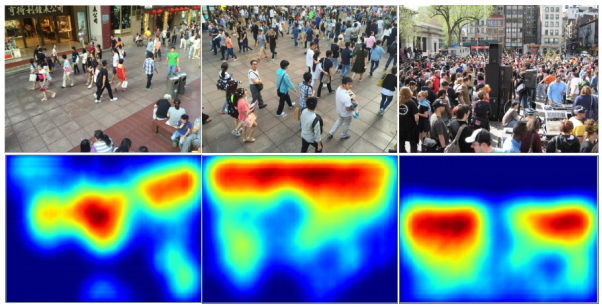

Figure 4. Attention maps generated by AMG at various crowd density levels (density level increases from left to right).

3.1.2 Architecture of attention map generator

The architecture of AMG is shown in Figure 5, we use the first 10 layers of trained VGG-16 model [22] as the front

end to extract low-level features. We build the back end by adopting multiple dilated convolution layers of different dilation rates with an architecture similar to the inception module in [25]. The multiple dilated convolution architecture is motivated from [29]. It has the capability of localizing people clusters with enlarged receptive fields. The inception module was originally proposed in [25] to process and aggregate visual information of various scales. We use this module to deal with the diversified crowd distribution in congested scenes.

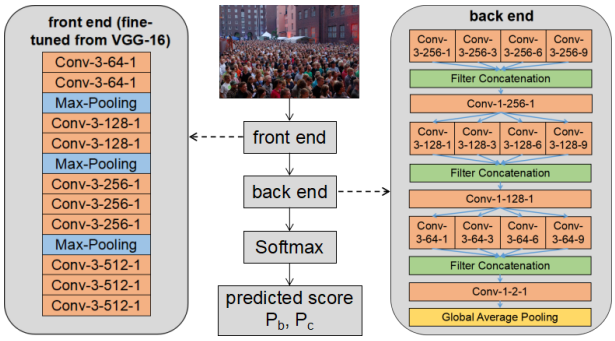

Figure 5. Architecture of AMG. All convolutional layers use padding to maintain the previous size. The convolutional layers' parameters are denoted as "Conv-(kernel size)-(number of filters)-(dilation rate)", max-pooling layers are conducted over a 2×2 pixel window, with stride 2.

3.2. Density Map Estimator

The DME network consists of two components: the front end and the back end. We remove the fully-connected layers of VGG-16 [22] and leave 10 convolutional layers to as the front end of the DME. The back end is a multi-scale deformable convolution based CNN network [6]. The architecture of DME is shown in Figure 6. The front end uses the first 10 layers of trained VGG-16 model [22] to extract low-level features. The back end uses multi-scale deformable convolutional layers with a structure similar to the inception module in [25], which enables DME to cope with various occlusion, diversified crowd distribution, and the distortion caused by perspective view.

The deformable convolution scheme was originally proposed in [6]. Beneficial from the adaptive (deformable) sampling location selection scheme, deformable convolution has shown its effectiveness on various tasks, such as object detection, in the wild environment. The deformable convolution treats the offsets of sampling locations as learning parameters. Rather than uniform sampling, the sampling locations in the deformable convolution can be adjusted and optimized via training (see Figure 7 for the learned sampling points by the deformable convolution on an example form ShanghaiTech Part_A dataset [32]). Compared to the uniform sampling scheme, this kind of dynamic

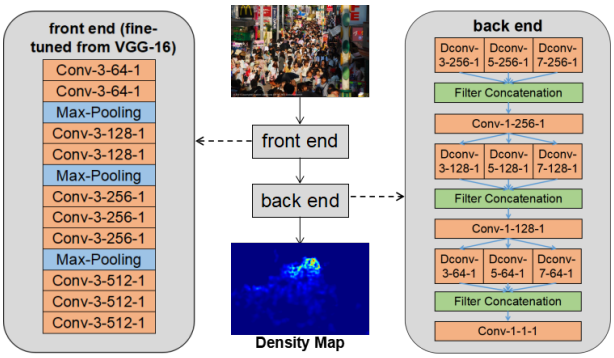

Figure 6. Architecture of DME. The convolutional layers' parameters are denoted as "Conv-(kernel size)-(number of filters)-(stride)", max-pooling layers are conducted over a 2×2 pixel window, with stride 2. The deformable convolutional layers' parameters are denoted as "Dconv-(kernel size)-(number of filters)-(stride)".

sampling scheme is more suitable for the crowd understanding problem of congested noisy scenes. We will show the comparative advantages of the deformable convolution in our experimental section.

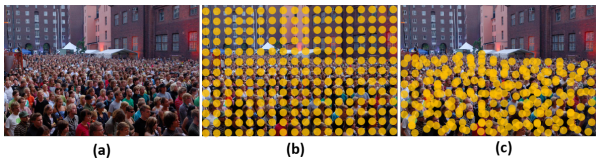

Figure 7. Illustration of the deformed sampling locations. (a) input image. (b) original sampling locations. (c) deformed sampling locations after learning.

4. Experiments

4.1. Datasets and Settings

We evaluate ADCrowdNet on four challenging datasets for crowd counting: ShanghaiTech dataset [32], the UCF_CC_50 dataset [11], the WorldExpo'10 dataset [30], and the UCSD dataset [2].

- **ShanghaiTech dataset**. The ShanghaiTech dataset [32] contains 1,198 images with a total of 330,165 people. It is divided into two parts: Part_A and Part_B. Part_A contains 482 pictures of congested scenes, in which 300 images are used as training dataset and 182 images are used as testing dataset; Part_B contains 716 images of sparse scene, 400 images of which are used as training dataset and 316 are used as testing dataset.

- **UCF_CC_50 dataset**. [11] contains 50 images downloaded from the Internet. The number of persons per image ranges from 94 to 4543 with an average of 1280 individuals. It is a very challenging dataset with two

problems: the limited number of the images and the large span in person count between images. We used 5-fold-cross-validation setting described in [11].

- **The WorldExpo'10 dataset**. The WorldExpo'10 dataset [30] contains 3980 frames from 1132 video sequences captured by 108 different surveillance cameras. Among 3980 images, 3380 images are used as training dataset and the remaining 600 images from 5 different scenes are used as testing dataset. Region-of-Interest (ROI) regions are provided in this dataset.

- **UCSD dataset**. The UCSD dataset [2] contains 2000 images in sparse scene. The dataset also provides ROI region information. We created the ground truth in the same way as we did for the WorldExpo'10 dataset. Since the size of each image is too small to support the generation of high-quality density maps, we therefore enlarge each image to 952×632 size by bilinear interpolation. Among the 2000 images, 800 images were used as training dataset, and the rest were used as testing dataset. Region-of-Interest (ROI) regions are also provided in this dataset.

We show a representative example for each crowd counting dataset in Figure 8. These four crowd counting datasets have their own characteristics. In general, the scenes in ShanghaiTech Part_A dataset are congested and noisy. Examples in ShanghaiTech Part_B are noisy but not highly congested. The UCF_CC_50 dataset consists of extremely congested scenes which have hardly any background noises. Both WorldExpo'10 dataset and UCSD dataset provide example with sparse crowd scenes in the form of ROI regions. Scenes in the ROI regions of the WorldExpo'10 dataset are generally noisier than the only one scene in the UCSD dataset.

Following [24, 14], we use the mean absolute error (MAE) and the mean square error (MSE) for quantitative evaluation of the estimated density maps. PSNR (Peak Signal-to-Noise Ratio) and SSIM [28] are used to measure the quality of the generated density map. For fair comparison, we follow the measurement procedure in [14] and resize the density map and ground truth to the size of the original input image by linear interpolation.

4.2. Training

4.2.1 AMG Training

Training data for the binary classification network AMG consists of two groups of samples: positive and negative samples. The positive samples are from the training sets of the four crowd counting datasets. The negative samples are 650 background images downloaded from the Internet. These negative samples are shared by the training of each individual dataset. These 650 negative samples contain various outdoor scenes where people appear, such as streets, squares, etc., ensuring that the biggest difference between positive sample and negative samples is whether the image contains people. Adam [12] is selected as the optimization method with the learning rate at 1e-5 and Standard cross-entropy loss is used as the loss function.

4.2.2 DME training

We simply crop 9 patches from each image where each patch is 1/4 of the original image size. The first four patches contain four quarters of the image without overlapping. The other five patches are randomly cropped from the image. After that, we mirror the patches so that we double the training dataset. We generate the ground truth for DME training following the procedure in [14]. We select Adam [12] as the optimization method with the learning rate at 1e-5. As previous works [32, 20, 14], we use the euclidean distance to measure the difference between the generating density map and ground truth and define the loss function as

$$L(\Theta) = \frac{1}{2N}\sum_{i=1}^{N} ||F(X_i;\Theta) - F_i||_2^2 \qquad (1),$$

where N is the batch size, $F(X_i;\Theta)$ is the estimated density map generated by DME with the parameterΘ, X_i is the input image, and F_i is the ground truth of X_i.

4.3. Results and Analyses

In this section, we first study several alternative network design of ADCrowdNet. After that, we evaluate the overall performance of ADCrowdNet and compare it with previous state-of-the-art methods.

4.3.1 Alternative study

Single-DME. Our first study is to investigate the influence of the AMG network, we compared two network designs on all the four datasets. The first one named AMG-DME has the architecture shown in Figure 2. The other one named DME uses the only DME network. Our quantitative experimental results in Table 1 show that AMG-DME is significantly superior than DME on the those datasets which are characteristic of noisy scenes: ShanghaiTech Part_A, Part_B and WorldExpo'10. In Figure 9, we illustrate two representative samples from the testing set of ShanghaiTech Part_A. On the top example which contains a congested noisy scene, estimated people number of AMG-DME is 198 that is much closer to the groud truth 171 than that estimated by DME. From the density map in the 3rd column of Figure 9, we can see the trees in the distance have been recognized as people by the single DME model. However, AMG-DME does not suffer this problem due to the help from the AMG

Figure 8. Representative examples from four crowd counting datasets.

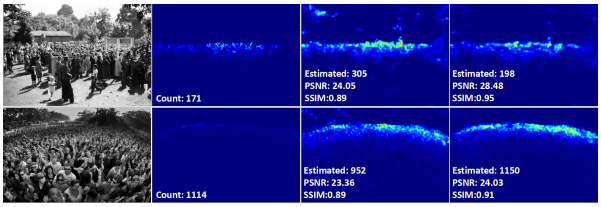

Figure 9. DME vs. AMG-DME. From left to right: Representative samples from the ShanghaiTech Part_A dataset, ground truth density map, density map generated from the architectures of single DME and AMG-DME.

network. On the middle-row example containing a noisy and more congested scene, the performances of AMG-DME and DME agree with those on the top example. The comparison results indicate that AMG-DME is more effective than DME on those noisy examples.

On the UCF_CC_50 dataset, AMG-DME has approximate performance (slightly higher AME but lower MSE) with DME. It may due to the fact most of examples in the UCF_CC_50 dataset have a large regions of congested crowds while rarely have background noises. On the UCSD dataset where scenes are neither congested nor noisy, both MSE and MAE of AMG-DME is slightly higher than DME. This might because the examples in the UCSD dataset have already provide the accurate information of ROI regions. The attention map generated by the AMG network may destroy the ROI regions, which degrades the performance of the DME network since some ROI regions may be erased from its input.

	DME		AMG-DME		AMG-bAttn-DME		AMG-attn-DME	
Dataset	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE
ShanghaiTech Part_A [32]	68.5	107.5	66.1	102.1	**63.2**	**98.9**	70.9	115.2
ShanghaiTech Part_B [32]	9.3	16.9	**7.6**	13.9	8.2	15.7	7.7	**12.9**
UCF_CC_50 [11]	**257.1**	363.5	257.9	**357.7**	266.4	358.0	273.6	362.0
The WorldExpo'10 [30]	8.5	-	7.4	-	**7.7**	-	**7.3**	-
The UCSD [2]	**0.98**	**1.25**	1.10	1.42	1.39	1.68	1.09	1.35

Table 1. Results of different variants of ADCrowdNet on four crowd counting datasets.

AMG-bAttn-DME. Since the AMG network has show its strength in coping with noise background of scenes, our second study is to explore if a hard binary attention mask is more effective than the soft attention employed by AMG-DME. We therefore set up an variant of AMG-DME

Figure 10. Illustration of the ROI regions extracted by different attention thresholds. The pixels which have the value of attention lower than t are changed to black.

	Part_A		Part_B	
Threshold	MAE	MSE	MAE	MSE
t = 0.2	68.0	104.1	9.2	17.8
t = 0.1	63.2	98.9	8.2	15.7
t = 0.0	63.2	100.6	8.6	15.0

Table 2. Performance of AMG-bAttn-DME under different binarization thresholds on the ShanghaiTech dataset.

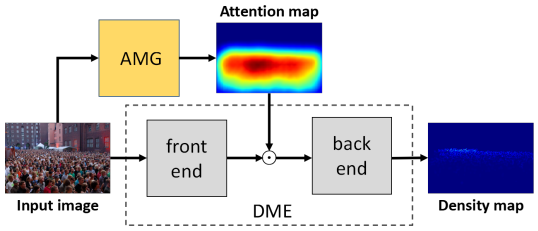

Figure 11. Architecture of AMG-attn-DME in both training and testing phases.

called AMG-bAttn-DME in Table 1. AMG-bAttn-DME has the same architecture as AMG-DME while differing with AMG-DME on the attention map (i.e., the attention maps of AMG-bAttn-DME contain either 0 or 1, other than a floating point within $[0, 1]$ in the attention maps of AMG-DME). We first conducted the experiments on the ShanghaiTech dataset to find out the optimal binarization threshold for AMG-bAttn-DME. We set three different threshold attention values,$\{0.2, 0.1, 0.0\}$, for the binarization of attention maps. The ROI regions are gradually enlarged with the decreasing the threshold values as shown in Figure 10. The

results shown in Table 2 indicates AMG-bAttn-DME with attention threshold of 0.1 achieved the best performance. We then evaluated AMG-bAttn-DME with this optimal attention threshold on the rest three datasets and reported the results in Table 1. It is observed that AMG-bAttn-DME is superior than AMG-DME only on ShanghaiTech Part_A while AMG-DME outperforms AMG-bAttn-DME on all other datasets. It may due to the AMG network can learn more accurate attention maps on ShanghaiTech Part_A and the binarization process does not destroy too much information of the crowd regions.

AMG-attn-DME. Complement to the above experiments, we stretched the design choice exploration to studying an alternative way of injecting the learned attention from the AMG network to the DME network. In our proposed architecture, the DME network directly takes the crowd images as input. An alternative architecture is to weigh intermediate the feature map of a certain layer of the DME network with the attention map from the AMG network. In our implementation, we inject the attention map into the output of the front end of the DME network as shown in Figure 11. Following the same training procedures as those in Table 1, this alternative architecture, named AMG-attn-DME, performs slightly worse than AMG-DME on the datasets with congested noisy scenes like ShanghaiTech Part_A and ShanghaiTech Part_B. This may be due to some non-crowd pixels in the attention map from the AMG network having an attention value of zero, which, during the injection, would make convolution features at those corresponding locations vanish, reducing the feature information learned by previous convolutional lays from the input. On the UCF_CC_50 dataset and UCSD dataset, AMG-attn-DME is worse than the the only DME network as AMG-bAttn-DME and AMG-DME. This is because the scenes of these two datasets have less noisy background, AMG-attn-DME may reduce the information of the ROI regions through the injected attention map. On the UCSD and WorldExpo'10 datasets, AMG-attn-DME achieved higher effectiveness. It may because the convolution feature vanishing problem has been alleviated by the black regions around the ROI regions in the input.

4.3.2 Quantitative results

In this section, we study the overall performance of ADCrowdNet and compare it with existing methods on each individual crowd counting dataset.

Comparison on MAE and MSE. We first compare the variants of the proposed ADCrowdNet network with the state-of-art work CSRNet [14] along with several previous methods including CP-CNN [24], MCNN [32], Cascaded-MTL [23], Switching-CNN [20] on the ShanghaiTech dataset and the UCF_CC_50 dataset. These two datasets are characteristic of congested and/or noisy scenes. The comparison results were summarized in Table 3. On the ShanghaiTech dataset, two of our approach variants ADCrowdNet(AMG-DME) and ADCrowdNet(AMG-bAttn-DME) achieved better performances than existing approaches. The only DME network achieved the performance generally close to the state-of-the-art approach CSRNet [14].

	Part_A		**Part_B**		**UCF_CC_50**	
Method	MAE	MSE	MAE	MSE	MAE	MSE
MCNN [32]	110.2	173.2	26.4	41.3	377.6	509.1
Cascaded-MTL [23]	101.3	152.4	20.0	31.1	322.8	397.9
Switching-CNN [20]	90.4	135.0	21.6	33.4	318.1	439.2
CP-CNN [24]	73.6	106.4	20.1	30.1	295.8	**320.9**
CSRNet [14]	68.2	115.0	10.6	16.0	266.1	397.5
ADCrowdNet(DME)	68.5	107.5	9.3	16.9	**257.1**	363.5
ADCrowdNet(AMG-DME)	66.1	102.1	**7.6**	13.9	257.9	357.7
ADCrowdNet(AMG-bAttn-DME)	**63.2**	**98.9**	8.2	15.7	266.4	358.0
ADCrowdNet(AMG-attn-DME)	70.9	115.2	7.7	**12.9**	273.6	362.0

Table 3. Estimation errors on ShanghaiTech and UCF_CC_50.

On the two relatively less challenging datasets WorldExpo'10 and UCSD, we compared ADCrowdNet with recent state-of-art recent approaches including Switching-CNN [20], MCNN [32], and CSRNet [14]. The comparison results are shown in Table 4. Our method achieved the best accuracy in scenes 1, 4, 5 as well as the best average accuracy on the WorldExpo'10 dataset . On the UCSD dataset, our DME model achieved the best accuracy on terms of both MAE and MSE.

	The WorldExpo'10						**UCSD**	
Method	Sce.1	Sce.2	Sce.3	Sce.4	Sce.5	Ave.	MAE	MSE
MCNN [32]	3.4	20.6	12.9	13.0	8.1	11.6	1.07	1.35
Switching-CNN [20]	4.4	15.7	10.0	11.0	5.9	9.4	1.62	2.10
CSRNet [14]	2.9	**11.5**	**8.6**	16.6	3.4	8.6	1.16	1.47
ADCrowdNet(DME)	**1.6**	15.8	11.0	10.9	3.2	8.5	**0.98**	**1.25**
ADCrowdNet(AMG-DME)	**1.6**	13.8	10.7	8.0	3.2	7.4	1.10	1.42
ADCrowdNet(AMG-bAttn-DME)	1.7	14.4	11.5	**7.9**	3.0	7.7	1.39	1.68
ADCrowdNet(AMG-attn-DME)	**1.6**	13.2	8.7	10.6	**2.6**	**7.3**	1.09	1.35

Table 4. Estimation error comparison on the WorldExpo'10 and UCSD. Note that only MAE is provided on WorldExpo'10 as previous approaches.

Comparison on PSNR and SSIM. To study the quality of the density maps generated by ADCrowdNet, another experiment was conducted on all the five datasets for both ADCrowdNet and the state-of-the-art method CSRNet [14]. The comparison results are shown in Table 5. Our method outperforms CSRNet on all the five datasets. On UCF_CC_50 dataset, our method improves 7.03% on PSNR and 55.76% on SSIM. On USCD dataset, our method improves 31.81% on PSNR and 8.13% on SSIM.

Evaluation on vehicle counting dataset. We conducted experiments on the TRANCOS [9] dataset for vehicle counting to evaluate the generalization capability of the proposed approach. The positive samples for training are from the training set of TRANCOS [9]. The negative samples use 250 background images downloaded from the Internet, including various road scenes without vehicle. As

	CSRNet [14]		ADCrowdNet	
Dataset	PSNR	SSIM	PSNR	SSIM
ShanghaiTech Part_A [32]	23.79	0.76	**24.48**	**0.88**
ShanghaiTech Part_B [32]	27.02	0.89	**29.35**	**0.97**
UCF_CC_50 [11]	18.76	0.52	**20.08**	**0.81**
The WorldExpo'10 [30]	26.94	0.92	**29.12**	**0.95**
The UCSD [2]	20.02	0.86	**26.39**	**0.93**
TRANCOS [9]	27.10	0.93	**29.56**	**0.97**

Table 5. CSRNet vs. ADCrowdNet (AMG-DME) on PSNR and SSIM.

previous work CSRNet [14], we use the Grid Average Mean Absolute Error (GAME) to measure the counting accuracy. The comparison results are shown in Table 6. It clearly shows that the ADCrowdNet approach achieved the best performance at all levels of GAMEs.

Method	GAME0	GAME1	GAME2	GAME3
Hydra-3s[16]	10.99	13.75	16.69	19.32
FCN-HA [31]	4.21	-	-	-
CSRNet [14]	3.56	5.49	8.75	15.04
ADCrowdNet(DME)	2.65	4.49	7.09	14.29
ADCrowdNet(AMG-DME)	**2.39**	4.23	6.89	14.82
ADCrowdNet(AMG-bAttn-DME)	2.69	4.61	7.13	14.14
ADCrowdNet(AMG-attn-DME)	2.44	**4.14**	**6.78**	**13.58**

Table 6. Evaluation on TRANCOS by Grid Average Mean Absolute Error (GAME).

4.3.3 Qualitative results

In this section, we further investigate the general performance of the proposed ADCrowdNet by qualitative results. We mainly compared ADCrowdNet with the state-of-art approach CSRNet [14] which have demonstrated the best performance on the datasets including the ShanghaiTech, UCF_CC_50, the WorlExpo'10, and UCSD datasets. In general, CSRNet has a front-end and back-end architecture as the DME network of the proposed ADCrowdNet. It is empowered by a dilated convolution design in the back-end of its architecture. Apart from the additional AMG netwok, ADCrowdNet differs from CSRNet by two additional features in its DME network: 1) the multiple-scale convolution scheme different from the single scale scheme of CSRNet, and 2) the deformable sampling scheme different from the evenly fixed-offset sampling in the dilated convolution of CSRNet.

Figure 12 shows some qualitative comparisons between the proposed ADCrowdNet (the variant AMG-DME is used) and the state-of-art approach CSRNet [14]. Through visualization, it is observed that CSRNet is much less effective on those examples with various noises than ADCrowdNet. We can see the evidence from the noise regions marked by red boxes of the 1st column where noises exist in the background regions, as well as the marked regions of the 3rd column where noises can be found in the crowd regions. This may be due to CSRNet directly takes the crowd image as input while the DME network of ADCrowdNet takes as the input the crowd information highlighted by its AMG network. On the example of the 2nd column where there is not much noise but a significantly non-uniform crowd distribution, ADCrowdNet also clearly outperforms CSRNet. This indicates that the multi-scale deformable convolution scheme in ADCrowdNet is more effective than the single-scale fixed-offset dilated convolution scheme in CSRNet.

On the rightmost example of Figure 12 which have highly occluded crowd regions (see the regions within the two green dotted bordered rectangle), ADCrowdNet only recognized part of the severely occluded crowd regions. It may because the AMG network of ADCrowdNet cannot highlight out the whole occluded crowd regions for the DME network. Nevertheless, ADCrowdNet still achieved better performance in terms of all the measurement parameters: estimated number, PSNR and SSIM.

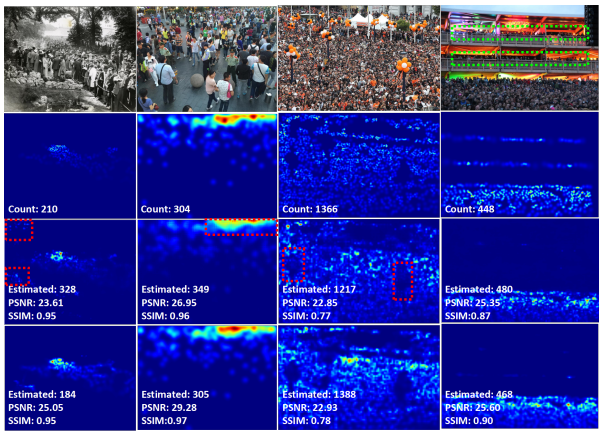

Figure 12. From top to bottom: representative samples from the testing set of the ShanghaiTech dataset, ground truth density maps, estimated density maps generated by the state-of-art approach CSRNet [14] and ADCrowdNet (AMG-DME) respectively.

5. Conclusion

We propose a convolutional neural network based architecture named ADCrowdNet for crowd understanding of congested noisy scenes. Benefiting from the multi-scale deformable convolutional layers and attention-aware training scheme, ADCrowdNet generally achieved more accurate crowd counting and density map estimation than existing methods by suppressing the problems caused by noises, occlusions, and diversified crowd distributions commonly presented in highly congested noisy environments. On four popular crowd counting datasets (ShanghaiTech, UCF_CC_50, WorldEXPO'10, UCSD) and an extra vehicle counting dataset TRANCOS, ADCrowdNet achieved significant improvements over recent state-of-the-art approaches.

References

[1] L. Boominathan, S. S. Kruthiventi, and R. V. Babu. Crowdnet: A deep convolutional network for dense crowd counting. In *Proc. ACM MM*, pages 640–644, 2016.

[2] A. B. Chan, Z.-S. J. Liang, and N. Vasconcelos. Privacy preserving crowd monitoring: Counting people without people models or tracking. In *Proc. IEEE CVPR*, pages 1–7, 2008.

[3] K. Chen, S. Gong, T. Xiang, and C. C. Loy. Cumulative attribute space for age and crowd density estimation. In *Proc. IEEE CVPR*, pages 2467–2474, 2013.

[4] K. Chen, C. C. Loy, S. Gong, and T. Xiang. Feature mining for localised crowd counting. In *Proc. BMVC*, pages 1–11, 2012.

[5] X. Chu, W. Yang, W. Ouyang, C. Ma, A. L. Yuille, and X. Wang. Multi-context attention for human pose estimation. In *Proc. IEEE CVPR*, pages 1831–1840, 2018.

[6] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei. Deformable convolutional networks. *CoRR, abs/1703.06211*, 1(2):3, 2017.

[7] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Proc. IEEE CVPR*, pages 886–893, 2005.

[8] P. Dollár, C. Wojek, B. Schiele, and P. Perona. Pedestrian detection: An evaluation of the state of the art. *IEEE Trans. Pattern Anal. Mach. Intell.*, 34(4):743–761, 2012.

[9] R. Guerrero-Gómez-Olmedo, B. Torre-Jiménez, R. López-Sastre, S. Maldonado-Bascón, and D. Onoro-Rubio. Extremely overlapping vehicle counting. In *Proc. Springer IbPRIA*, pages 423–431, 2015.

[10] J. Hu, L. Shen, and G. Sun. Squeeze-and-excitation networks. In *Proc. IEEE CVPR*, 2018.

[11] H. Idrees, I. Saleemi, C. Seibert, and M. Shah. Multi-source multi-scale counting in extremely dense crowd images. In *Proc. IEEE CVPR*, pages 2547–2554, 2013.

[12] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *Proc. ICLR*, 2015.

[13] V. Lempitsky and A. Zisserman. Learning to count objects in images. In *Proc. NIPS*, pages 1324–1332, 2010.

[14] Y. Li, X. Zhang, and D. Chen. Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes. In *Proc. IEEE CVPR*, pages 1091–1100, 2018.

[15] M. Marsden, K. McGuinness, S. Little, and N. E. O'Connor. Fully convolutional crowd counting on highly congested scenes. *arXiv preprint arXiv:1612.00220*, 2016.

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

[17] V.-Q. Pham, T. Kozakaya, O. Yamaguchi, and R. Okada. Count forest: Co-voting uncertain number of targets using random forest for crowd density estimation. In *Proc. IEEE ICCV*, pages 3253–3261, 2015.

[18] R. Qian, R. T. Tan, W. Yang, J. Su, and J. Liu. Attentive generative adversarial network for raindrop removal from a single image. In *Proc. IEEE CVPR*, pages 2482–2491, 2018.

[19] M. Ren and R. S. Zemel. End-to-end instance segmentation with recurrent attention. In *Proc. IEEE CVPR*, pages 21–26, 2017.

[20] D. B. Sam, S. Surya, and R. V. Babu. Switching convolutional neural network for crowd counting. In *Proc. IEEE CVPR*, pages 4031–4039, 2017.

[21] C. Shang, Bo, H. Ai, and Bai. End-to-end crowd counting via joint learning local and global count. In *Proc. IEEE ICIP*, pages 1215–1219, 2016.

[22] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *Proc. ICLR*, 2015.

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

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

[25] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *Proc. IEEE CVPR*, pages 1–9, 2015.

[26] P. Viola and M. J. Jones. Robust real-time face detection. *International journal of computer vision*, 57(2):137–154, 2004.

[27] E. Walach and L. Wolf. Learning to count with cnn boosting. In *Proc. Springer ECCV*, pages 660–676, 2016.

[28] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

[29] Y. Wei, H. Xiao, H. Shi, Z. Jie, J. Feng, and T. S. Huang. Revisiting dilated convolution: A simple approach for weakly- and semi-supervised semantic segmentation. In *Proc. IEEE CVPR*, pages 7268–7277, 2018.

[30] C. Zhang, H. Li, X. Wang, and X. Yang. Cross-scene crowd counting via deep convolutional neural networks. In *Proc. IEEE CVPR*, pages 833–841, 2015.

[31] S. Zhang, G. Wu, J. P. Costeira, and J. M. Moura. Fcn-rlstm: Deep spatio-temporal neural networks for vehicle counting in city cameras. In *Proc. IEEE ICCV*, pages 3687–3696, 2017.

[32] Y. Zhang, D. Zhou, S. Chen, S. Gao, and Y. Ma. Single-image crowd counting via multi-column convolutional neural network. In *Proc. IEEE CVPR*, pages 589–597, 2016.

[33] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. Learning deep features for discriminative localization. In *Proc. IEEE CVPR*, pages 2921–2929, 2016.