Shallow Feature Based Dense Attention Network for Crowd Counting

Yunqi Miao,[1] **Zijia Lin,**[2] **Guiguang Ding,**[3] **Jungong Han**[1*]

[1]University of Warwick, Coventry, UK
[2]Microsoft Research, Beijing, China
[3]Tsinghua University, Beijing, China
[1]{Yunqi.Miao.1, Jungong.Han}@warwick.ac.uk
[2]zijlin@microsoft.com
[3]dinggg@tsinghua.edu.cn

Abstract

While the performance of crowd counting via deep learning has been improved dramatically in the recent years, it remains an ingrained problem due to cluttered backgrounds and varying scales of people within an image. In this paper, we propose a **S**hallow feature based **D**ense **A**ttention **Net**work (SDANet) for crowd counting from still images, which diminishes the impact of backgrounds via involving a shallow feature based attention model, and meanwhile, captures multi-scale information via densely connecting hierarchical image features. Specifically, inspired by the observation that backgrounds and human crowds generally have noticeably different responses in shallow features, we decide to build our attention model upon shallow-feature maps, which results in accurate background-pixel detection. Moreover, considering that the most representative features of people across different scales can appear in different layers of a feature extraction network, to better keep them all, we propose to densely connect hierarchical image features of different layers and subsequently encode them for estimating crowd density. Experimental results on three benchmark datasets clearly demonstrate the superiority of SDANet when dealing with different scenarios. Particularly, on the challenging UCF_CC_50 dataset, our method outperforms other existing methods by a large margin, as is evident from a remarkable 11.9% Mean Absolute Error (MAE) drop of our SDANet.

Introduction

Crowd counting aims to count the number of people by means of estimating the density distribution of the crowd in a single image. It is a very useful computer vision technique to facilitate a variety of applications, including crowd control, disaster management and public safety monitoring. However, it is not a trivial task due to great challenges in real-world situations caused by cluttered backgrounds and non-uniform people scale within an image.

Tremendous algorithms (Zhang et al. 2016; Li, Zhang, and Chen 2018; Jiang et al. 2019) have been proposed in the literature for estimating the crowd density distribution. The majority of them focused on addressing two problems when

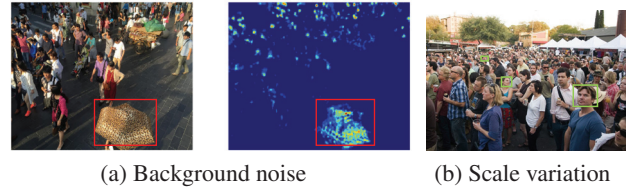(a) Background noise (b) Scale variation

Figure 1: Illustrations of the problems of cluttered backgrounds and varying scales of people. In (a), the right picture depicts the estimated density map of the left image, where backgrounds like the umbrella (in red box) could be mistakenly regarded as people in the density map and thus decrease the estimation accuracy. In (b), sizes of human heads (in green boxes) vary greatly within the image due to their different distances from the camera.

learning the mappings from image features to density distribution maps, *i.e.*, 1) how to eliminate the impacts of cluttered backgrounds, and 2) how to deal with varying scales of people within an image. Figure 1 illustrates both mentioned problems. Specifically, in Figure 1(a), the right picture depicts the estimated density map of the left image, derived by the MCNN model (Zhang et al. 2016). It can be noticed that backgrounds, *e.g.*, umbrellas, could be mistakenly regarded as people on the density map, thus decreasing the estimation accuracy. Meanwhile, as illustrated in Figure 1(b), sizes of human heads can vary greatly within an image, because of their different distances from the camera.

To eliminate noises caused by cluttered backgrounds, attention mechanism is usually introduced to re-weigh features or regions in terms of their probabilities of being the crowds. Generally, additional training samples and parameters are employed to train standalone classifiers indicating density levels (Sam, Surya, and Babu 2017) or head probability (Liu et al. 2019) as the metric to evaluate the importance of different features/regions within an image, on the basis of which, different weights are given to the features/regions. However, standalone networks with complex structures usually require millions of extra to-be-learned parameters, which can be a heavy burden for a real-life application.

By exploring the relationship between images and their corresponding normalized shallow feature maps generated by several baselines (Zhang et al. 2016; Boominathan, Kruthiventi, and Babu 2016; Li, Zhang, and Chen 2018) (Figure 2), we observe, for the first time, that backgrounds like stairs, trees and buildings, tend to have significantly different responses from those of the human crowds. For example, the backgrounds have stronger responses in Figure 2(a) but weaker ones in Figure 2(b), whereas human crowds' reactions are opposite (weaker responses in Figure 2(a) but stronger ones in Figure 2(b)). This tells us that backgrounds and human crowds are more separable on shallow-layer feature maps. An attention model based on shallow features has potential to generate more accurate attention maps. Therefore, instead of involving a sophisticated standalone attention model as previous works, we incorporate an attention module in our feature extraction networks, which effectively reuses the shallow features and enjoys less complex structures to diminish background noises.

Regarding the problem of varying scales of people within an image, some works (Zhang et al. 2016; Deb and Ventura 2018) adopted "multi-column" frameworks to extract multi-scale information from images, where each branch extracts features of a specific scale by adopting filters with a certain size. Others exploit some convolutional operations, like dilated (Li, Zhang, and Chen 2018; Deb and Ventura 2018) and deformable convolution kernels (Liu et al. 2019; Zou et al. 2018), to capture multi-scale information by expanding the receptive field of filters. Yet most of them extracted features layer by layer, and thus the features of the current layer may lose information of features in some preceding layers.

Actually, the most representative features of people across different scales can appear in different layers of the feature extraction networks. For example, the most representative features of people in a smaller scale can probably be extracted in an earlier layer, while those of people in a larger scale can be extracted in a later layer. Thus, it is vital to keep information of features in all different layers. Therefore, densely-connected structure that enables each layer to process features from all preceding layers seems like an appropriate structure, on which features corresponding to all scales can be well preserved and better encoded to facilitate the estimation of the crowd density.

Based on the observations above, we propose a new method for crowd counting, termed **S**hallow feature based **D**ense **A**ttention **Net**work (SDANet). SDANet consists of three components, *i.e.*, low-level feature extractor, high-level feature encoder, and attention map generator. As mentioned above, the attention map generator reduces the noises caused by backgrounds via re-weighing specific regions with attention maps generated with shallow features. Moreover, multi-scale information is well preserved via densely connecting the features of different layers in the high-level feature encoder. Extensive experiments on benchmark datasets also clearly demonstrate the superiority of SDANet.

Contributions of our work are summarized as follows:

- We observe, for the first time, that shallow features contain distinguishable information between backgrounds
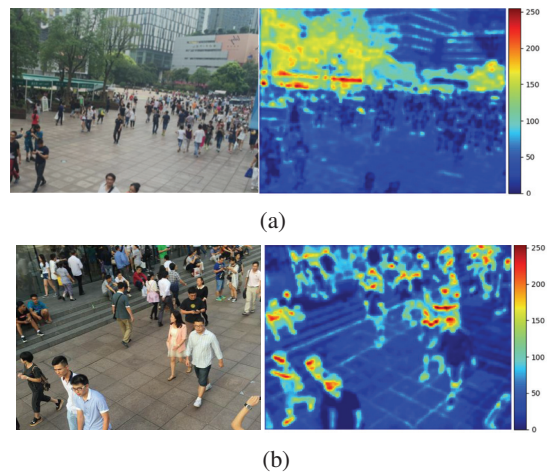

(a)

(b)

Figure 2: Images and their corresponding shallow feature maps from several baselines. The shallow feature maps are linearly normalized to [0,255] by their maximums, which are shown as heat maps. It can be seen that the backgrounds and the human crowds have significantly different responses in (a) and (b).

and human crowds, which allows us to utilize a lightweight network to generate even more accurate attention maps.

- We propose to employ densely connected structures in feature extraction/encoding networks, such that multi-scale information in different layers can be well kept to facilitate the estimation of the crowd density.

- We propose a novel crowd counting method termed SDANet. And experiments conducted on three benchmark datasets show that SDANet achieves the state-of-the-art performance for crowd counting.

Related Works

Over the last few years, researchers have attempted to address the issue of crowd counting by density estimation with a variety of approaches (Sindagi and Patel 2018), where a mapping from image features to crowd density is learned and then the counted number is the summation over an estimated density map. Existing density estimation methods can be generally categorized as hand-crafted feature based ones and deep feature based ones, where latter ones tend to incorporate attention mechanism recently.

Hand-Crafted Feature based Methods

Early works usually extract hand-crafted features implying global image characteristics, such as local binary pattern (LBP) and gray level co-occurrence matrices (GLCM), and learn its mapping to the density by regression models, ranging from linear ones to non-linear ones. Lempitsky *et al.*(Lempitsky and Zisserman 2010) utilized linear models to describe the mapping from image features to the density in a local region, which is applied in bacteria counting
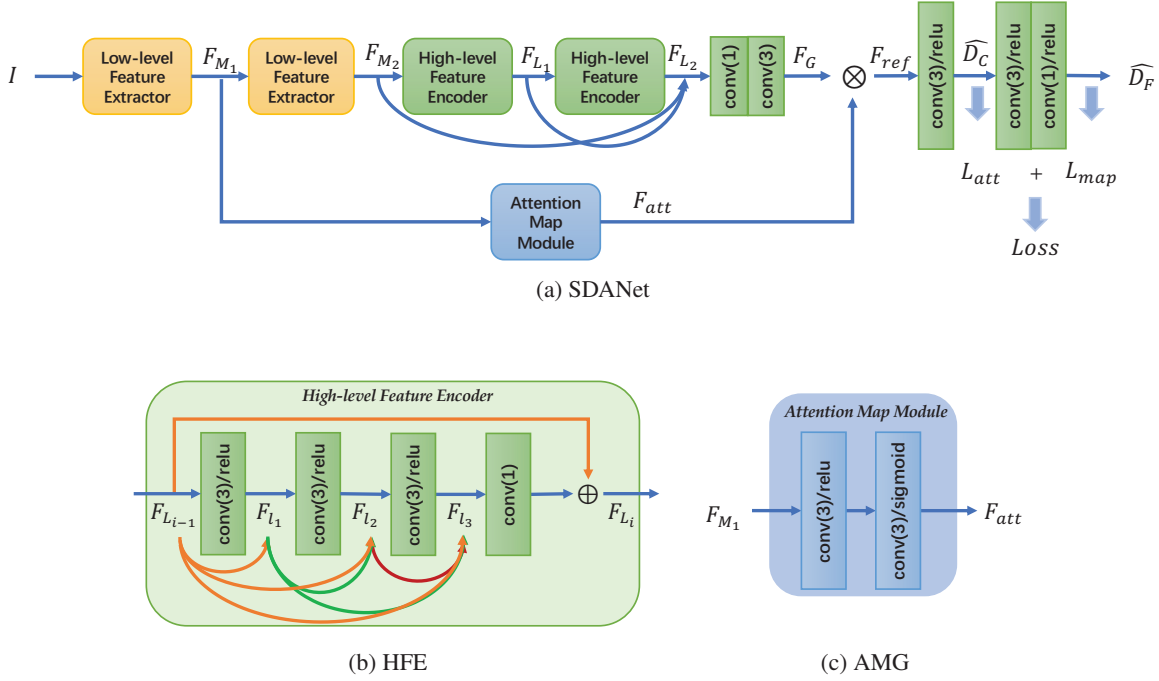
Figure 3: (a) The architecture of SDANet. (b) The architecture of HFE. (c) The architecture of AMG.

and crowd counting with a relatively sparse density. Idress *et al.*(Idrees et al. 2013) explored features from three sources, *i.e.*, Fourier, interest points and head detection combined with their respective confidences to get counts at localized patches and adopted a Markov Random Field (MRF) framework to obtain an estimated count for the entire image.

Deep Feature based Methods

Inspired by the huge success of convolutional neural networks (CNN) in image classification (Krizhevsky, Sutskever, and Hinton 2012), recently deep features have been leveraged for density estimation. Owing to their superior performance, deep learning based methods (Wang et al. 2015; Zhang et al. 2016; Deb and Ventura 2018; Li, Zhang, and Chen 2018; Hossain et al. 2019) quickly dominate the research in crowd counting.

Zhang *et al.*(Zhang et al. 2016) proposed a multi-column based architecture (MCNN), where each column adopts a filter with a certain size to extract features of the corresponding scale. Instead of training all patches with the same paralleled network, Sam *et al.*(Sam, Surya, and Babu 2017) proposed a switching CNN that adaptively selects the optimal branch for an image patch according to its density. A classifier indicating patch density is trained beforehand and empowers density estimation networks by providing prior knowledge. Recently, dilated kernels have also been involved in multi-column frameworks to further deliver larger reception fields (Li, Zhang, and Chen 2018).

Attention mechanism in crowd counting Recently, attention mechanism is widely incorporated to enhance the crowd counting performance. The idea is to roughly approximate the regions in the image where people are likely appeared. To do so, an attention model is learned to assign larger weights to pixels/regions of being human crowds. (Liu et al. 2018; Kang and Chan 2018; Hossain et al. 2019; Liu et al. 2019; Zhu et al. 2019).

ADCrowdNet (Liu et al. 2019) employs an attention map generator trained on additional negative samples and then applies it to detect crowd regions in the images. Hossain *et al.*(Hossain et al. 2019) proposed a Scale-Aware Attention Networks (SAAN), which utilizes attention mechanism to re-weigh multi-scale features learned by multi-columns. SFANet (Zhu et al. 2019) generates an attention map with the same size of the image by an additional CNN branch, where each pixel indicates its probability of being the head. Alternatively, DecideNet (Liu et al. 2018) uses a learned attention map to combine the two maps generated by the regression branch and the detection branch.

The proposed SDANet in this paper is also a deep feature based method with attention mechanism incorporated. However, different from previous works that learned a standalone attention model with sophisticated structures, by observing that shallow features can have strong signals to distinguish backgrounds and human crowds, we propose to use shallow features to build an attention module in SDANet with simpler network structures. Moreover, instead of encoding multi-scale features layer by layer that has the risk of losing feature information of some preceding layers, we propose to densely connect outputs of each layer in SDANet, so that multi-scale features of different layers can be better kept and encoded to facilitate the estimation of crowd density.
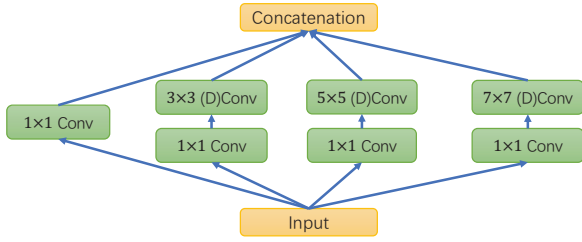
Figure 4: The architecture of the second module of LFE. **(D)Conv** represents the convolution layer with dilated kernels.

Our Approach

The framework of SDANet is illustrated schematically in Figure 3(a), which mainly consists of three components: Low-level Feature Extractor (LFE), High-level Feature Encoder (HFE), and Attention Map Generator (AMG).

Low-level Feature Extractor (LFE)

Most existing methods use separate branches with different size filters to extract multi-scale information from images, which may introduce redundant structures into the pipeline (Li, Zhang, and Chen 2018). Inspired by the success of SANet (Cao et al. 2018) in feature extraction, the Inception module (Szegedy et al. 2015), a tool to process visual information of various scales, is used as the shallow feature extractor of SDANet.

Specifically, LFE consist of two feature extractor blocks and each of them contains four branches with filter sizes of 1×1, 3×3, 5×5, and 7×7 respectively, as shown in Figure 4. Each branch focuses on a certain scale and generates the same number of feature maps. To further enhance model's capability to capture multiple scales information, dilated convolution, which can enlarge the receptive field without involving extra computations, is employed in the second block. Additionally, expect for the 1×1 branch, there is an extra 1×1 filter added before the other three branches to reduce the feature channels by half. Moreover, ReLU activate function is applied after each convolution layer in LFE to avoid negative values.

As a departure from most of works, we remove the pooling layers between the inception modules to avoid the reduction in spatial resolution caused by the pooling operation and the additional complexity brought by subsequent deconvolutional layers. Considering the trade-off between resource consumption and model accuracy, we instead adopt dilated filters with the dilated rate of 2 to replace the pooling layer (Chen et al. 2018). Features from different branches, covering multi-scale appearance of people in images, are subsequently concatenated together for the feature encoding.

High-level Feature Encoder (HFE)

The structure of HFE is shown in Figure 3(b), which takes shallow features extracted from the second block of LFE as input. While encoding features, such a structure can well preserve multi-scale information.

HFE is compose of two blocks, where each block consists of three convolution layers with the filter size of 3×3 followed by a ReLU activate function. Particularly, the input of a specific convolution layer $F_{l_j}(j = 1, 2, 3)$ is the concatenation of all outputs from preceding layers, i.e., $F_{l_j} = concat(F_1, ..., F_{l_{j-1}})$, which are indicated by different colors in the figure. The dense connection between layers ensures that multi-scale information in the shallow features can be preserved. At the bottom of each block, a 1×1 convolution layer is applied to integrate the concatenated hierarchical features and reduce feature channels to the same dimension as the input, which is indicated by $Conv_{1 \times 1}$. Therefore, the output of the i-th block in HFE F_{L_i} can be obtained by,

$$F_{L_i} = F_{L_{i-1}} + Conv_{1 \times 1}[concat(F_{l_1}, F_{l_2}, F_{l_3})]. \quad (1)$$

Finally, the input of each block is added onto the output, which will in turn become the input of the next block.

On top of that, to further preserve multi-scale information, shallow features obtained by low-level feature extractor (F_{M_2}) and the output of each block in HFE (F_{L_i}) are concatenated together, which is F_g in Eq. (2), as the input for the feature integration in global level. In the integration, a 1×1 and a 3×3 convolution layer are employed to integrate high-level features in a global level, which is indicated by G in Eq. (2). Henceforth, the output of HFE can be calculated by,

$$F_G = G(F_g). \quad (2)$$

Rather than widening the network, the proposed densely connected structure takes full advantage of features from all layers and well preserves the scale information in shallow features, which efficiently eliminates the problem of scale variation. In the paper, the dimension of F_{M_2} and F_G are both set to 64 according to the extensive experiments, which is less than most of the state-of-the-art methods.

Attention Map Generator (AMG)

In light of the observation that backgrounds on shallow feature maps tend to have significantly different responses, compared to the crowds, we generate attention maps based on low-level features only. Specifically, AMG takes shallow features from the first block of LFE (F_{M_1}) as input and generates pixel-wise attention maps (F_{att}) on which crowd regions are always "brighter" than the backgrounds, i.e.,

$$F_{att} = AMG(F_{M_1}). \quad (3)$$

Here, two convolution layers followed by a sigmoid function, as shown in Figure 3(c), are used to ensure that all the computed weights are within the range of 0 to 1. L_{att}, the summation of pixel-wise Euclidean distance between refined feature maps F_{ref} and ground-truth density map D, conveys the supervision information to the learning process of the attention module. Subsequently, the attention map F_{att} is employed to refine the encoded feature F_G by element-wise multiply ($\otimes$) as follows,

$$F_{ref} = F_G \otimes F_{att}, \quad (4)$$

where F_{ref} is taken as the input of the last two convolution layers whose filter sizes are 1×1 and 3×3 respectively to generate the high-quality density map $\hat{D}$ under the supervision of a combination of several losses.

Figure 5: Qualitative results on ShanghaiTech Dataset. For each group of images, pictures in the middle and on the right are corresponding ground truth and estimated density map of the image on the left, where the number on the top right corner indicates the ground truth (GT) and the estimated number of people (PRE) respectively. It can be seen that SDANet has a strong adaptability to different density levels with a error less than 4%.

Table 1: Comparison results of different methods on the UCF_CC_50 dataset.

Method	MAE	MSE
FHSc+MRF	468.0	590.3
MCNN	377.6	509.1
Switching-CNN	318.1	439.2
SANet (Cao et al. 2018)	258.4	334.9
CSRNet (Li, Zhang, and Chen 2018)	266.1	397.5
SAAN (Hossain et al. 2019)	271.6	391.0
SDANet (ours)	**227.6**	**316.4**

Loss Function

The density maps generator in the SDANet adopts a coarse-to-fine strategy. Concretely, the loss is composed of two terms: L_{att} and L_{map} in the Figure 3(a) respectively.

Firstly, a convolution layer with the filter size of 3×3 is employed to learn a coarse mapping between combined feature maps (F_{ref}) from the HFE and AMG to the density maps, and meanwhile, prepare coarse density maps for further process. In order to supervise the learning process of attention maps and the generation of coarse density maps, L_{att}, measuring the Euclidean distance between coarse density maps ($\hat{D}_C$) and the ground-truth density map D, is adopted. Explicitly, L_{att} is defined as,

$$L_{att} = \frac{1}{M} \sum_M ||\hat{D}_C - D||_2^2, \tag{5}$$

where M is the dimension of $\hat{D}_C$, and is set to 32 throughout all experiments.

Subsequently, two convolution layers with filter sizes of 3×3 and 1×1 are involved to further refine the quality of coarse density map, thus enhancing the accuracy of crowd counting. Noticeably, the ReLU activation function is employed after convolution layers to avoid appearance of negative values. Last, L_{map} is introduced to supervise the refinement process and generate the fine-grained density map ($\hat{D}_F$). Concretely, L_{map} is composed by an Euclidean loss

(L_E) and a Counting loss (L_C), which are somewhat complementary to each other. Initially, L_E is adopted to improve the quality of density map by minimizing the Euclidean distance between the fine-grained density map and the ground-truth, which can be described by,

$$L_E = \frac{1}{N} \sum_{i=1}^{N} ||\hat{D}_{F_i} - D_i||_2^2, \tag{6}$$

where $\hat{D}_{F_i}$ and D_i are estimated density map and ground truth of the i-th image I_i, respectively, and N refers to the number of training samples. However, sharp edges and outliers in coarse density maps might be blurry in fine-grained maps. To remedy this situation, L_C is added as a compensation, which is defined by,

$$L_C = \frac{1}{N} \sum_{i=1}^{N} (\frac{C_i - \hat{C}_i}{C_i + \epsilon})^2, \tag{7}$$

where $\hat{C}_i$ and C_i represent, respectively, the estimated number of people and the ground truth of the i-th training sample, which are the integral over all pixels p on the corresponding density map, i.e., $C_i = \sum_p D_i$. Additionally, $\epsilon = 0.0001$ is set to avoid the denominator being zero. L_C not only accelerates the convergence process but improves the counting accuracy. In summary, L_{map} is expressed as,

$$L_{map} = L_E + \alpha L_C, \tag{8}$$

where $\alpha = 0.01$ is the empirical weight for L_C.

Therefore, the overall loss of SDANet is,

$$Loss = L_{att} + L_{map}. \tag{9}$$

Adam (Kingma and Ba 2014) algorithm with the initial learning rate of 1e-4 is adopted to optimize the SDANet.

Experiments

Evaluation Metrics

Similar to the previous work, the mean absolute error (MAE) and mean squared error (MSE) metrics are used for algo-

Table 2: Comparison results of different methods on 5 scenes (S1∼S5) in the WorldExpo′10 dataset in terms of MAE.

Method	S1	S2	S3	S4	S5	Average
Cross-scene (Zhang et al. 2015)	9.8	14.1	14.3	22.2	3.7	12.9
MCNN (Zhang et al. 2016)	3.4	20.6	12.9	13.0	8.1	11.6
Switching-CNN (Sam, Surya, and Babu 2017)	4.4	15.7	10.0	11.0	5.9	9.4
SANet (Cao et al. 2018)	2.6	13.2	9.0	13.3	3.0	8.2
CSRNet (Li, Zhang, and Chen 2018)	2.9	**11.5**	**8.6**	16.6	3.4	8.6
SaCNN (Zhang, Shi, and Chen 2018)	2.6	13.5	10.6	12.5	3.3	8.5
SDANet (ours)	**2.0**	14.3	12.5	**9.5**	**2.5**	**8.1**

rithm evaluation, which are defined as:

$$MAE = \frac{1}{N} \sum_{i=1}^{N} \left| C_i - \hat{C}_i \right|, \qquad (10)$$

$$MSE = \sqrt{\frac{1}{N} \sum_{i=1}^{N} \left(C_i - \hat{C}_i \right)^2}, \qquad (11)$$

where N represents the total number of images involved in testing, C_i and $\hat{C}_i$ are the ground truth and estimated number of people for the i-th image respectively.

Datasets

In the experiment, three crowd counting benchmark datasets, the UCF_CC_50 dataset, the WorldExpo′10 dataset, and the ShanghaiTech dataset, are used to evaluate the performance of SDANet, each being elaborated below.

UCF_CC_50 dataset   (Idrees et al. 2013) contains 50 images with various perspectives and resolutions. The number of annotated people per image ranges from 94 to 4543 with an average number of 1280, which is a challenging dataset in the field of crowd counting.

WorldExpo′10 dataset   (Zhang et al. 2015) consists of 3980 annotated frames from 1132 video sequences captured by 108 different surveillance cameras, which is divided into a training set (3380 frames) and a test set (600 frames). The region of interest (ROI) is also provided for the whole dataset.

ShanghaiTech dataset   (Zhang et al. 2016) consists of 1198 annotated images with a total amount of 330,165 annotated people. The dataset contains two parts: Part_A and Part_B. Part_A includes 482 internet images with highly congested scenes while Part_B includes 716 images with relatively sparse crowd scenes taken from streets in Shanghai.

Experiment Settings

Taking the computation cost and data variety into account, we adopted the patch-wise training strategy. Following the previous work (Zhang et al. 2016), 9 patches, where each patch is $1/4$ of the image size, are cropped from each image to generate the training set. The first four patches contain four quarters of the image without overlapping while the other five patches are randomly cropped from the image. During the test, non-overlapping patches are cropped from

each image in the test set and compute individually. The final density map of the image is the concatenation of its patches' predictions. Additionally, images are further augmented by randomly horizontal flipping.

Besides, we generated the ground-truth from head annotations given by datasets (Zhang et al. 2016). Each head annotation is blurred with a Gaussian kernel, whose summation is normalized to one and the number of people is the integral over the density map.

The implementation of SDANet is based on the PyTorch framework. As we train the whole network from scratch, all parameters are randomly initialized by Gaussian distribution with mean of zero and standard deviation of 0.01.

Results and Analysis

On each dataset, we follow the standard protocol to generate ground truth and compare our method with the state-of-the-art algorithms. Furthermore, we conduct extensive ablation experiments on the WorldExpo′10 dataset to analyze the effects of different components in SDANet. We explain experimental settings and show results as follows.

Experimental Evaluations

Quantitative results   On the **UCF_CC_50** dataset, we performed a 5-fold cross-validation to evaluate the proposed method as suggested by (Idrees et al. 2013). Table. 1 shows the comparison of the results of our method with contemporary state-of-the-art works on UCF_CC_50 dataset, which illustrates the proposed SDANet is able to deal with crowd scenes with varying densities and achieves a superior performance over other approaches. Specifically, our method achieves 11.91% MAE reduction and 5.52% MSE reduction. This clearly demonstrates that SDANet is super robust against the scale and density changes.

The comparison results of SDANet with contemporary state-of-the-art work on the 5 scenes (S1∼S5) in the test set of **WorldExpo′10** dataset are shown in Table. 2. The challenging test set is a combination of different densities, ranging from sparse to dense, and various backgrounds including squares, stations, *etc*. From the result, it can be seen that the proposed SDANet scores the best in Scene1, Scene4 and Scene5 as well as the best accuracy on average, which again proves the strong adaptability of SDANet against different scenarios with varying density levels.

On the **ShanghaiTech** dataset, SDANet is evaluated and compared with other recent works and results are shown in Table. 3. Again, the proposed method attains the lowest

Table 3: Comparison results of different methods on the ShanghaiTech dataset.

Method	Part_A		Part_B	
	MAE	MSE	MAE	MSE
Cross-scene	181.8	277.7	32.0	49.8
MCNN	110.2	173.2	26.4	41.3
Switching-CNN	90.4	135.0	21.6	33.4
CP-CNN (Sindagi and Patel 2017)	73.6	106.4	20.1	30.1
DecideNet (Liu et al. 2018)	-	-	21.5	32.0
ACSCP (Shen et al. 2018)	75.7	102.7	17.2	27.4
CSRNet	68.2	115.0	10.6	16.0
SANet	67.0	104.5	8.4	13.6
TEDnet (Jiang et al. 2019)	64.2	109.1	8.2	12.8
SDANet (ours)	**63.6**	**101.8**	**7.8**	**10.2**

Table 4: Ablation study results on the WorldExpo′10 dataset.

Models	MAE	MSE
SDANet without AMG	12.89	15.28
SDANet without Dense Structure	10.14	13.25
SDANet without Refinement	9.64	13.19
SDANet	8.10	12.90

MAE and MSE as well. Specifically, our approach outperforms the latest work TEDnet by 4.87% and 20.31% over the MAE and MSE metric respectively on the ShanghaiTech Part_B dataset.

Visualization results   We firstly analyzed the attention maps generated by AMG and obtained some statistical results. Taking the attention map of Figure 2(a) as an example, the average attention value of crowd region (center-right) is 0.874 (GT=1) while that for background region (left corner) is 0.253 (GT=0), which proves that the attention maps reduce the background noise by arranging background regions with relatively low weights.

To demonstrate the performance of SDANet on scenes with cluttered backgrounds and varying head sizes, we choose, in particular, the ShanghaiTech dataset for estimated density maps visualization, which are shown in Figure 5. For each group of images, pictures in the middle and on the right are corresponding ground truth and estimated density map of the image on the left, where the number on the top right corner indicates the ground truth (GT) and the estimated number of people (PRE) respectively. Here, we display the estimated density maps of various scenarios, ranging from 103 persons to 1067 persons, to demonstrate that the proposed SDANet performs decently in both dense and sparse scenes. It can be seen that SDANet has a strong adaptability to different density levels with a error less than 4%.

Ablation Study

To validate the effectiveness of key components in the SDANet, we also conducted ablation studies on the WorldExpo′10 dataset which is more realistic and challenging due to the fact that all images are acquired from real surveillance scenes.

Effectiveness of AMG   We explore the performance improvement offered by AMG by removing the attention module from the SDANet and compare it with the network with AMG. The result is indicated by *SDANet without AMG* in Table. 4. There are 37% increase in MAE and 15% increase in MSE if AMG is dropped out, clearly demonstrating that AMG has made a significant contribution in diminishing background noise.

Effectiveness of densely-connected structure   In order to shed light on how the densely connecting structure preserves multi-scale features, we conduct an experiment on the same dataset without the dense connection between layers and the result is indicated by *SDANet without Dense Structure* in Table. 4. It can be seen that the removal of the dense connection between layers leads to an over 20.1% drop in the counting accuracy, which means that densely-connected structure reinforces the diversity of features and improve the performance of SDANet.

Effectiveness of estimation refined layers   Furthermore, we study the refinement ability of the last two layers and the loss term L_{map}. We screen out the last two convolution layers in SDANet and train the network with solely L_{att}, whose result is indicated by *SDANet without Refinement* in Table. 4. Without the refinement layers, there is a nearly 16% decline in the MAE. Therefore, the coarse-to-fine strategy involved in the loss function can further enhance the performance of the network.

Conclusion

In this paper, we have presented a brand-new Shallow feature based Dense Attention Network (SDANet) aiming to automatically count the number of people in an image. Our SDANet is characterized by: 1) diminishing the impact of backgrounds via involving a lightweight attention model, and 2) capturing multi-scale information via densely connecting hierarchical image features. Extensive experiments have been carried out and the results on three benchmark datasets validate the adaptability and robustness of the SDANet when varying crowd scenes from sparse to dense.

References

Boominathan, L.; Kruthiventi, S. S.; and Babu, R. V. 2016. Crowdnet: A deep convolutional network for dense crowd counting. In *Proceedings of the 2016 ACM on Multimedia Conference*, 640–644. ACM.

Cao, X.; Wang, Z.; Zhao, Y.; and Su, F. 2018. Scale aggregation network for accurate and efficient crowd counting. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 734–750.

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

Deb, D., and Ventura, J. 2018. An aggregated multicolumn dilated convolution network for perspective-free counting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 195–204.

Hossain, M.; Hosseinzadeh, M.; Chanda, O.; and Wang, Y. 2019. Crowd counting using scale-aware attention networks. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 1280–1288. IEEE.

Idrees, H.; Saleemi, I.; Seibert, C.; and Shah, M. 2013. Multi-source multi-scale counting in extremely dense crowd images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2547–2554.

Jiang, X.; Xiao, Z.; Zhang, B.; Zhen, X.; Cao, X.; Doermann, D.; and Shao, L. 2019. Crowd counting and density estimation by trellis encoder-decoder network. *arXiv preprint arXiv:1903.00853*.

Kang, D., and Chan, A. 2018. Crowd counting by adaptively fusing predictions from an image pyramid. *arXiv preprint arXiv:1805.06115*.

Kingma, D. P., and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, 1097–1105.

Lempitsky, V., and Zisserman, A. 2010. Learning to count objects in images. In *Advances in neural information processing systems*, 1324–1332.

Li, Y.; Zhang, X.; and Chen, D. 2018. Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1091–1100.

Liu, J.; Gao, C.; Meng, D.; and Hauptmann, A. G. 2018. Decidenet: Counting varying density crowds through attention guided detection and density estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5197–5206.

Liu, N.; Long, Y.; Zou, C.; Niu, Q.; Pan, L.; and Wu, H. 2019. Adcrowdnet: An attention-injective deformable convolutional network for crowd understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3225–3234.

Sam, D. B.; Surya, S.; and Babu, R. V. 2017. Switching convolutional neural network for crowd counting. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 4031–4039. IEEE.

Shen, Z.; Xu, Y.; Ni, B.; Wang, M.; Hu, J.; and Yang, X. 2018. Crowd counting via adversarial cross-scale consistency pursuit. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5245–5254.

Sindagi, V. A., and Patel, V. M. 2017. Generating high-quality crowd density maps using contextual pyramid cnns. In *Proceedings of the IEEE International Conference on Computer Vision*, 1861–1870.

Sindagi, V. A., and Patel, V. M. 2018. A survey of recent advances in cnn-based single image crowd counting and density estimation. *Pattern Recognition Letters* 107:3–16.

Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; and Rabinovich, A. 2015. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1–9.

Wang, C.; Zhang, H.; Yang, L.; Liu, S.; and Cao, X. 2015. Deep people counting in extremely dense crowds. In *Proceedings of the 23rd ACM international conference on Multimedia*, 1299–1302. ACM.

Zhang, C.; Li, H.; Wang, X.; and Yang, X. 2015. Cross-scene crowd counting via deep convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 833–841.

Zhang, Y.; Zhou, D.; Chen, S.; Gao, S.; and Ma, Y. 2016. Single-image crowd counting via multi-column convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 589–597.

Zhang, L.; Shi, M.; and Chen, Q. 2018. Crowd counting via scale-adaptive convolutional neural network. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 1113–1121. IEEE.

Zhu, L.; Zhao, Z.; Lu, C.; Lin, Y.; Peng, Y.; and Yao, T. 2019. Dual path multi-scale fusion networks with attention for crowd counting. *arXiv preprint arXiv:1902.01115*.

Zou, Z.; Su, X.; Qu, X.; and Zhou, P. 2018. Da-net: Learning the fine-grained density distribution with deformation aggregation network. *IEEE Access* 6:60745–60756.