Dense Scale Network for Crowd Counting

Feng Dai

fdai@ict.ac.cn

Key Laboratory of Trustworthy
Distributed Computing and Service,
Beijing University of Posts and
Telecommunications
& Key Laboratory of Intelligent
Information Processing, Institute of
Computing Technology, Chinese
Academy of Sciences
Beijing, China

Hao Liu

liuhao2018@ict.ac.cn

Key Laboratory of Intelligent
Information Processing, Institute of
Computing Technology, Chinese
Academy of Sciences
& University of Chinese Academy of
Sciences
Beijing, China

Yike Ma

ykma@ict.ac.cn

Key Laboratory of Intelligent
Information Processing, Institute of
Computing Technology, Chinese
Academy of Sciences
Beijing, China

Xi Zhang

zhangx@bupt.edu.cn

Key Laboratory of Trustworthy
Distributed Computing and Service,
Beijing University of Posts and
Telecommunications
Beijing, China

Qiang Zhao

zhaoqiang@ict.ac.cn

Key Laboratory of Intelligent
Information Processing, Institute of
Computing Technology, Chinese
Academy of Sciences
Beijing, China

ABSTRACT

Crowd counting has been widely studied by computer vision community in recent years. Due to the large scale variation, it remains to be a challenging task. Previous methods adopt either multi-column CNN or single-column CNN with multiple branches to deal with this problem. However, restricted by the number of columns or branches, these methods can only capture a few different scales and have limited capability. In this paper, we propose a simple but effective network called DSNet for crowd counting, which can be easily trained in an end-to-end fashion. The key component of our network is the dense dilated convolution block, in which each dilation layer is densely connected with the others to preserve information from continuously varied scales. The dilation rates in dilation layers are carefully selected to prevent the block from gridding artifacts. To further enlarge the range of scales covered by the network, we cascade three blocks and link them with dense residual connections. We also introduce a novel multi-scale density level consistency loss for performance improvement. To evaluate our method, we compare it with state-of-the-art algorithms on five crowd counting datasets (ShanghaiTech, UCF-QNRF, UCF_CC_50, UCSD and WorldExpo'10). Experimental results demonstrate that DSNet can achieve the best overall performance and make significant improvements.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

ICMR '21, August 21–24, 2021, Taipei, Taiwan

© 2021 Association for Computing Machinery.

ACM ISBN 978-1-4503-8463-6/21/08...\$15.00

<https://doi.org/10.1145/3460426.3463628>

CCS CONCEPTS

• **Computing methodologies** → **Computer vision**; • **Applied computing** → **Surveillance mechanisms**.

KEYWORDS

Crowd Counting, Dense Scale, Connections, Scale Consistency

ACM Reference Format:

Feng Dai, Hao Liu, Yike Ma, Xi Zhang, and Qiang Zhao. 2021. Dense Scale Network for Crowd Counting. In *Proceedings of the 2021 International Conference on Multimedia Retrieval (ICMR '21), August 21–24, 2021, Taipei, Taiwan*. ACM, New York, NY, USA, 9 pages. <https://doi.org/10.1145/3460426.3463628>

1 INTRODUCTION

With the rapid growth of population, crowd counting has gained considerable attention, because of its broad applications in video surveillance, traffic control and sport events. Earlier works estimate crowd counts via the detection of body or head [25, 37, 38], while some other methods learn a mapping from local or global hand-crafted feature to the actual count [4, 5, 26]. More recently, the problem of crowd counting is formalized into the regression of crowd density map, whose values are summed to give the count of crowd within that image. This approach can handle serious occlusions in dense crowd images. With the success of deep learning technology, convolutional neural networks (CNNs) are utilized to generate accurate crowd density map and can get better performance than traditional methods.

However, crowd counting still remains to be an extremely challenging task due to large scale variation, heavy occlusion, background noise and perspective distortion. Among them, scale variation is the main issue and has attracted the most attention of recent CNN-based methods [2, 3, 21, 27, 32, 42, 43]. A number of multi-column or multi-branch networks have been proposed to

Figure 1: Large scale variations exist in crowd counting datasets. Left: Input image and corresponding ground truth density map from ShanghaiTech dataset [43]. Right: Input image and corresponding ground truth density map from UCF-QNRF dataset [10].

handle scale variation for better accuracy. These architectures contain several columns of CNN or several branches from different stages of backbone network. The columns or branches have different receptive fields to capture variation in people sizes. Although these methods show good improvements, the scale diversity they captured is restricted by the number of columns or branches.

The key challenges of scale variation lie in two aspects. First, as shown in Figure 1 left, the people in crowd image often have very different sizes, ranging from several pixels to tens of pixels. This requires the network to be able to capture a *large range* of scales. Second, as shown in Figure 1 right, the scale usually varies continuously across the image, especially for high density images. This requires the network to be able to sample the captured scale range *densely*. However, *none* of existing methods can deal with these two challenges simultaneously.

In this paper, we propose a novel dense scale single-column neural network called DSNet for crowd counting. DSNet is built upon the blocks consisting of densely connected dilated convolutional layers. Thus it can output features having different receptive fields and capture crowds at different scales. The block has similar structure with DenseASPP [40] used for semantic segmentation, but with different combination of dilation rates. We carefully select these rates for layers within the blocks, such that each block can *more densely* sample the continuously varied scale than DenseASPP. At the same time, the selected combination of dilation rates can use *all* pixels under the receptive field for feature computation, and let us step from gridding artifacts. To further increase the scale diversity captured by DSNet, we stack three dense dilated convolution blocks (DDCBs) and link them by utilizing the residual connection [8] densely. The final network can sample a very large scale range in a much denser manner, thus has the ability to deal with the large scale variation problem in crowd counting.

Another issue of most previous methods is that they use traditional Euclidean loss to train their networks, which is based on the pixel independence hypothesis. This loss ignores the global

and local coherence in density maps and would give poor performance in crowd counting. To overcome this issue, we incorporate a novel multi-scale density level consistency loss, which is used to ensure the global and local density level consistency between the estimated and ground truth crowd density map. The new proposed loss function can be quickly calculated and further improve the performance of DSNet.

In summary, the main contributions of our paper are:

- We propose the dense dilated convolution block (DDCB) with carefully selected dilation rates and residually connect them to tackle the scale variation problem in crowd counting.
- Besides of commonly used Euclidean loss, we incorporate a multi-scale density level consistency loss for network training, which can further improve the performance.
- We carry out extensive experiments on five public crowd counting datasets. Our approach can achieve the best performance compared with existing state-of-the-art methods.

2 RELATED WORK

Existing crowd counting methods can be classified into two categories: traditional counting methods and CNN-based methods. By combining deep learning, the CNN-based methods demonstrate powerful ability and outperform traditional methods.

2.1 Traditional methods

Most of the early traditional works focus on detection-based methods using body or part-based detector to locate people in the crowd image and count their number. However, severe occlusions of highly congested scenes limit the performance of these methods. To overcome the problem, regression-based methods are deployed to learn a mapping from the extracted feature to the number of objects directly. Following similar approaches, Idrees *et al.* [9] proposed a method that extracts features via Fourier analysis and SIFT [18] interest points based counting in local patches. Due to the overlooked saliency that causes inaccurate results in local regions, Lempitsky *et al.* [13] proposed a method that learns a linear mapping between features and its object density maps in the local region. Furthermore, considering the difficulty of learning an ideal linear mapping, Pham *et al.* [22] used random forest regression to learn a non-linear mapping instead of the linear one.

2.2 CNN-based methods

Due to the success of convolutional neural networks (CNNs), CNN-based methods are employed for the purpose of crowd counting and density estimation. Walach *et al.* [36] made use of layered boosting and selective sampling methods to reduce the estimation error. Instead of using patch-based training, Shang *et al.* [28] proposed an estimation method using CNNs which takes the whole image as input and directly outputs the final crowd count. Boominathan *et al.* [2] presented the first work purely using convolutional network and dual-column architecture to tackle the issue of scale variation for generating density map. Zhang *et al.* [43] introduced a multi-column architecture to extract features at different scales. Similarly, Onoro *et al.* [21] proposed a scale-aware, multi-column counting model called Hydra CNN for object density estimation. Recently,

Figure 2: The architecture of the proposed dense scale network (DSNet) for crowd counting. The DSNet consists of backbone network with the front ten layers of VGG-16, three dense dilated convolution blocks (DDCBs) with dense residual connections (DRCs), and three convolutional layers for crowd density map regression. The DDCBs with DRCs are used to enlarge scale diversity and receptive fields of features to handle large scale variations so that density maps can be estimated accurately. The symbol © means feature concatenation.

inspired by MCNN [43], Sam *et al.* [27] proposed a Switching-CNN that adaptively select the most optimal regressor among several independent regressors for a particular patch. Sindagi *et al.* [32] explored a new architecture where MCNN [43] is enriched with two additional columns capturing global and local context. Tian *et al.* [35] proposed PaDNet, which contains multiple subnetworks pretrained on scenarios with different densities. Jiang *et al.* [11] presented a multi-scale encoder and a multi-path decoder based network architecture for high-quality density map estimation.

Although these multi-column networks have the ability to estimate crowd count, several disadvantages also exist: they are hard to train and they have large amount of redundant parameters. Also the speed is slow as multiple CNNs need to be run. Taking all above drawbacks into consideration, recent works have focused on multi-scale, single column architectures. Zhang *et al.* [42] proposed a scale-adaptive CNN that combines feature maps extracted from multiple layers to produce the final density map. Li *et al.* [14] proposed a network for congested scene called CSRNet, which uses dilated kernels to deliver larger reception fields. Liu *et al.* [16] proposed ADCrowdNet, which adopts attention mechanism to address the accuracy degradation problem of highly congested noisy scenes in crowd counting. Cao *et al.* [3] presented scale aggregation network that improves the multi-scale representation and generates high-resolution density maps. Moreover, Idrees *et al.* [10] proposed a composition loss for counting and localization of people based on the architecture of multiple dense blocks after branching off the base network similarly. Yan *et al.* [39] proposed PGCNet, which uses perspective-guided convolution to overcome the dramatic intra-scene scale variations of people due to the perspective effect. Perspective information is integrated into density regression to provide additional knowledge of the person scale change in the crowd

image [29]. DADNet [7] explores scale-aware attention fusion with various dilation rates to capture different visual granularities of crowd regions. Tan *et al.* [34] applied multiple regressors on regions with different congestion degrees. All these single-column works can only capture several kinds of receptive fields, which limits the network to handle large variations in crowd images.

3 OUR APPROACH

The fundamental idea of our approach is to deploy an end-to-end single-column CNN with denser scale diversity to deal with the large scale variations and density level differences in both congested and sparse scenes. In this section, we first introduce the architecture of DSNet we proposed as shown in Fig. 2. And then we describe our novel multi-scale density level consistency loss, which enforces the estimated density maps to be consistent with the ground truth according to crowd density at multiple scale levels.

3.1 DSNet architecture

Our proposed DSNet contains backbone network as feature extractor, three dense dilated convolution blocks stacked by dense residual connections that enlarges denser scale diversity, and three convolutional layers for crowd density map regression.

Backbone network: We keep the first ten layers of VGG-16 [31] with only three pooling layers to be our backbone network. According to [31], using more convolutional layers with small kernels is more efficient than using fewer layers with larger kernels that have been employed in existing multi-column networks. Moreover, it achieves the best tradeoff between accuracy and the resource overhead, which is suitable for accurate and fast crowd counting.

Dense dilated convolution block(DDCB): As discussed in Section 1, the scale often varies continuously across the image and has

large range. Considering these challenges, multi-columns CNNs or single-column CNNs with multiple branches have restricted capabilities. This is because they all have limited number of columns or branches, and can only process crowd images with a few different scales. Thus to tackle the challenges of scale variation, we need a network architecture that can capture a large scale range in the manner as dense as possible. DenseASPP [40] stacks dilated layers by dense connections to enlarge scale diversity and receptive field for semantic segmentation of high-resolution street images. However, the large dilation rates in their network has led to large gap between the sizes of different receptive field, i.e. 6 pixels. This is not desired for crowd counting. Because the scale variation of crowd scenes caused by camera perspective is nearly continuous, we need more densely sampled scale range. Here, we propose a new dense dilated convolution block that contains three dilated convolutional layers with increasing dilation rate of 1, 2, 3. This setting preserves information from denser scales with small gap of receptive field size, i.e. 2 pixels. We add a 1×1 convolutional layer before every dilated convolutional layer to prevent the dense block from growing too wide. Each dilated layer within the block is densely connected with others, so that each layer can access to all the subsequent layers and pass on information that needs to be preserved. After dense connection, the acquired scale diversity is increased. A standard convolution layer with 3×3 filter size is adopted to fuse concatenated features from previous layers and reduce channel number in the end. We further enlarge the scale range by linking multiple blocks with dense residual connections, which will be discussed later.

Another advantage of our selected dilation rates is that it can overcome the *gridding artifacts*. In DenseASPP [40], a dilated convolution layer with dilation rate of 6 is put below the layer with dilation rate of 3. The final result of one pixel after these two layers only obtains information from 7 pixels in one-dimensional case. This phenomenon gets worse when the input data is two-dimensional. As a result, the final pixel can only view original information in a gridding fashion and lose a large portion (86.4%) of information. This is harmful for crowd counting to capture detailed features, as local information of original feature maps is completely missing and the information can be irrelevant across large distances due to large dilation rate. By adopting the new combination of dilation rates, the top layer can cover all pixel information of original feature map and avoid irrelevant information across large distances caused by large dilation rates of intermediate layers. This is critical for crowd counting to estimate density maps with high quality and accurate counting.

Dense residual connection(DRC): Although the proposed DDCB provides dense scale diversity, the hierarchical features between different blocks are not fully utilized. Hence, we improve the architecture by adding dense residual connections to further improve the information flow. Also, they can prevent the network from being wider compared with the conventional dense connections. By doing this, the output of one DDCB has direct access to each layer of the subsequent DDCBs, resulting in a contiguous information pass. Compared with the residual connection without dense style, its scale diversity is further enlarged and the suitable features for specific scenes are preserved adaptively during the flow process of information. The ablation experiments in Section 5.2 demonstrate

that the dense residual connections (DRCs) can further improve the performance.

Density map estimator(DME): After the DDCBs extract detailed features with large scale diversity and receptive field, these features are passed to an estimator consisting of two 3×3 convolution layers and one 1×1 convolution layer to estimate the density map. ReLU [6] is applied after every convolution layer except the last one.

3.2 Loss function

Most previous works use Euclidean loss as the loss function for crowd counting, which only takes care of pixel error but ignores the global and local correlations between estimated and ground truth density maps. In this paper, we incorporate multi-scale density level consistency loss.

Euclidean loss: Similar to other works [2, 14, 27, 43], we choose the Euclidean distance to measure the dissimilarity between the estimated density map and the ground truth one, i.e.

$$L_e = \frac{1}{N} \sum_{i=1}^N \|G(X_i; \theta) - D_i^{GT}\|_2^2, \quad (1)$$

where N is the number of images in the batch, $G(X_i; \theta)$ is the estimated density map for training image X_i with parameters θ , D_i^{GT} is the ground truth density map for X_i .

Multi-scale density level consistency loss: We also take the consistency of global and local density level between estimated density maps and ground truth into consideration. Here, the multi-scale density level consistency loss is defined as

$$L_c = \frac{1}{N} \sum_{i=1}^N \sum_{j=1}^S \frac{1}{k_j^2} \|P_{ave}(G(X_i; \theta), k_j) - P_{ave}(D_i^{GT}, k_j)\|_1, \quad (2)$$

where S is the number of scale levels for consistency checking, P_{ave} is average pooling operation, k_j is the output size of average pooling. In our experiments, the loss L_c has $S = 3$ scale levels with output size of 1×1 , 2×2 and 4×4 respectively for best trade off between training speed and estimation accuracy. The grid average mean absolute error (GAME) proposed in [24] is similar to our multi-scale density level consistency loss. However, GAME is used as an evaluation metric, while in our work L_c is used for performance improvement.

Final objective: By weighting the above two loss functions, the entire network is trained using the following objective function:

$$L = L_e + \lambda L_c, \quad (3)$$

where λ is the weight parameter.

4 IMPLEMENTATION DETAILS

In this section, we provide specific details of generating ground truth, training process and evaluation process. Due to the good design of network architecture and joint loss function, our network can be trained end-to-end easily.

4.1 Ground truth generation

Following the method of generating density maps in [43], the geometry-adaptive kernels are adopted to tackle the datasets with congested scenes including ShanghaiTech Part_A [43], UCF-QNRF [10] and UCF_CC_50 [9], while the fixed Gaussian kernels are used

Table 1: The comparison of different methods on ShanghaiTech, UCF-QNRF, UCF_CC_50 and UCSD datasets: the bold font indicates the best performance and the underlined one indicates the second best performance.

Method	Year	Venue	ShanghaiTech A		ShanghaiTech B		UCF-QNRF		UCF_CC_50		UCSD	
			MAE ↓	RMSE ↓	MAE ↓	RMSE ↓	MAE ↓	RMSE ↓	MAE ↓	RMSE ↓	MAE ↓	RMSE ↓
D-ConvNet [30]	2018	CVPR	73.5	112.3	18.7	26.0	-	-	288.4	404.7	-	-
IG-CNN [1]	2018	CVPR	72.5	118.2	13.6	21.1	-	-	291.4	349.4	-	-
ic-CNN [23]	2018	ECCV	68.5	116.2	10.7	16.0	-	-	260.9	365.5	-	-
CSRNet [14]	2018	CVPR	68.2	115.0	10.6	16.0	-	-	266.1	397.5	1.16	1.47
SANet [3]	2018	ECCV	67.0	104.5	8.4	13.6	-	-	258.4	334.9	1.02	1.29
PaDNet [35]	2019	TIP	59.2	98.1	8.1	12.2	96.5	170.2	185.8	278.3	<u>0.85</u>	1.06
ADCrowdNet [16]	2019	CVPR	63.2	98.9	8.2	15.7	-	-	266.4	358.0	1.39	1.68
TEDnet [11]	2019	CVPR	64.2	109.1	8.2	12.8	113	188	249.4	354.5	-	-
PACNN [29]	2019	CVPR	66.3	106.4	8.9	13.5	-	-	267.9	357.8	0.89	<u>1.18</u>
DADNet [7]	2019	ACM MM	64.2	99.9	8.8	13.5	113.2	189.4	285.5	389.7	-	-
MRNet [34]	2019	ACM MM	63.3	97.8	7.5	11.5	111.1	182.8	232.3	314.8	-	-
BL [20]	2019	ICCV	62.8	101.8	7.7	12.7	<u>88.7</u>	<u>154.8</u>	229.3	308.2	-	-
DSSINet [15]	2019	ICCV	60.6	96.0	<u>6.9</u>	<u>10.7</u>	99.1	159.2	216.9	302.4	-	-
PGCNet [39]	2019	ICCV	57.0	86.0	8.8	13.7	-	-	244.6	361.2	-	-
HyGNN [19]	2020	AAAI	60.2	<u>94.5</u>	7.5	12.7	100.8	185.3	184.4	270.1	-	-
AMRNet [17]	2020	ECCV	61.6	98.4	7.0	11.0	86.6	152.2	<u>184.0</u>	<u>265.8</u>	-	-
DSNet (ours)	-	-	61.3	97.3	6.7	10.5	91.4	160.4	183.3	240.6	0.82	1.06

to generate density maps for the datasets with relatively sparse crowd including ShanghaiTech Part_B [43] and UCSD [4]. And the ground-truth density maps of WorldExpo'10 [41] are generated following [41, 43]. By blurring each head annotation using a Gaussian kernel (which is normalized to 1), the ground truth density maps can be generated. The setups of fixed standard deviation for the datasets with sparse crowd follow the configuration in [14]. And the geometry-adaptive kernel used for datasets with congested crowd is defined as follows:

$$D^{GT} = \sum_{i=1}^N \delta(x - x_i) \times G_{\sigma_i}(x), \text{ with } \sigma_i = \beta \bar{d}_i, \quad (4)$$

where x is the position of pixel and N is the number of head annotations in the image. For each targeted object x_i in the ground truth δ , we use $\bar{d}_i$ to indicate the average distance of k nearest neighbors. Then we convolve $\delta(x - x_i)$ with a Gaussian kernel with parameter σ_i . Following [43], we set $\beta = 0.3$ and $k = 3$.

4.2 Training details

Similar to other recent crowd counting works [2, 14, 27, 33], the backbone network with ten layers is fine-tuned from a well-trained VGG-16 [31]. All new layers are initialized from a Gaussian distribution with zero mean and 0.01 standard deviation. Adam [12] optimizer is applied with fixed learning rate at $5e-6$ and weight decay of $5e-4$ because it shows faster convergence than stochastic gradient descent with momentum. And the network is trained with batch size of 1.

During training, we crop image patch of 1/4 size of the original image at four quarters of the image without overlapping and other patches are randomly cropped from the image. After that, it is horizontal flipped randomly with probability 0.5. The image resolution of UCF-QNRF [10] dataset is larger than all other datasets, causing GPU out of memory in the training process. We resize large images to a maximum size of 720p before data augmentation. On

the contrary, the image resolution of UCSD [4] dataset is 238×158 , which is too small to generate high-quality density maps. Hence, we enlarge each image to 952×632 resolution.

4.3 Evaluation details

At test time, we do not crop image patches from the images and instead we feed the whole image into the network to generate the estimated density map. The mean absolute error (MAE) and root mean square error (RMSE) are adopted to evaluate the performance of our network. The MAE reflects the model's accuracy while the RMSE reflects the model's robustness. Lower value means better performance. These two metrics are defined as

$$MAE = \frac{1}{N} \sum_{i=1}^N |C_i - C_i^{GT}|, \quad (5)$$

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N |C_i - C_i^{GT}|^2}, \quad (6)$$

where N is the number of images in test set, C_i represents the estimated count calculated by summation of the whole output density map while C_i^{GT} is the ground truth count.

5 EXPERIMENTS

In this section, we compare our proposed approach with state-of-the-art methods on the five publicly available datasets. Then, we conduct ablation experiments on ShanghaiTech dataset to demonstrate the effectiveness of each component of our network architecture and the proposed density level consistency loss. The networks with different combinations of dilation rates in DDCBs are also trained to prove the validity of our carefully selected ones.

5.1 Comparisons with State-of-the-Art

To demonstrate the effectiveness of our proposed approach, we conduct comparative experiments on five public challenging crowd

Figure 3: The sample images, density maps estimated by DSNet and ground truth density maps for (a) ShanghaiTech Part_A, (b) ShanghaiTech Part_B, (c) UCF-QNRF, (d) UCF_CC_50 and (e) World-Expo'10 datasets.

Table 2: The comparison of different methods on World-Expo'10 dataset. Only the MAE evaluation metric are reported.

Method	S1	S2	S3	S4	S5	Avg.
D-ConvNet [30]	<u>1.9</u>	12.1	20.7	8.3	2.6	9.1
IG-CNN [1]	2.6	16.1	10.2	20.2	7.6	11.3
ic-CNN [23]	17.0	12.3	9.2	<u>8.1</u>	4.7	10.3
CSRNet [14]	2.9	11.5	<u>8.6</u>	16.6	3.4	8.6
SANet [3]	2.6	13.2	9.0	13.3	3.0	8.2
PGCNet [39]	2.5	12.7	8.4	13.7	3.2	8.1
ADCrowdNet [16]	1.7	14.4	11.5	7.9	3.0	7.7
TEDnet [11]	2.3	10.1	11.3	13.8	2.6	8.0
PACNN [29]	2.3	12.5	9.1	11.2	3.8	<u>7.8</u>
DSNet (ours)	2.2	<u>10.6</u>	9.9	12.9	<u>2.9</u>	7.7

counting datasets. The experimental results on ShanghaiTech [43], UCF-QNRF [10], UCF_CC_50 [9] and UCSD [4] are shown in Table 1. The results on WorldExpo'10 [41] are shown in Table 2. It is obvious that our proposed method achieves the best overall performance compared with existing state-of-the-art crowd counting methods. Our method can get the best performance on three datasets and achieve comparable performance on the left dataset.

For ShanghaiTech dataset, our model achieves the lowest MAE and RMSE in Part_B compared to other state-of-the-art methods. For UCF-QNRF dataset, which is the largest dataset and the one with the largest variation of head size, our model attains the third best performance, which is much similar to that of BL [20]. For UCF_CC_50 dataset, DSNet achieves 9.5% RMSE improvement compared with AMRNet [17]. DSNet also achieves the best performance on UCSD dataset, which demonstrates that our approach not only achieves superior performance on congested crowd scenes but also on sparse crowd dataset. On the WorldExpo'10 dataset, although

Table 3: The performance of different variations of our proposed network on ShanghaiTech Dataset. The number in bracket indicates how many dense dilated convolution blocks are used.

	ShanghaiTech A		ShanghaiTech B	
Network	MAE↓	RMSE↓	MAE↓	RMSE↓
VGG+DME	70.9	112.2	15.21	22.96
VGG+DDCB(1)+DME	66.2	105.4	10.71	15.89
VGG+DDCB(2)+DME	64.8	104.4	8.67	13.76
VGG+DDCB(3)+DME	63.2	103.2	7.33	12.07
VGG+DDCB(3)+DRC+DME	62.6	102.8	7.06	12.01
VGG+DDCB(3)+DRC+ L_c +DME	61.3	97.3	6.74	10.48

Table 4: The performance of our network on ShanghaiTech Dataset when different connections are used to link the three DDCBs.

	ShanghaiTech A		ShanghaiTech B	
Connections	MAE↓	RMSE↓	MAE↓	RMSE↓
Without Connections	62.5	101.2	6.86	10.85
Residual Connection	61.8	99.6	6.81	10.54
Dense Residual Connection	61.3	97.3	6.74	10.48

our method does not deliver the best accuracy for each type of scene, it gives the best average performance.

The crowd counting results of our approach on the datasets are presented in Fig.3. It is obvious that our method performs well on counting the number of people in images. It also validates that the large varied heads with different sizes can be captured so that DSNet becomes more robust and accurate for not only congested crowd scenes but also sparse scenes.

5.2 Ablation experiments

In this section we conduct ablation experiments of our network components as well as the configuration of loss function and compare the performance of three different combined styles of dilation rates in DDCB to choose the best configuration. All these experiments are conducted on the ShanghaiTech dataset because Part_A set is comprised of congested crowd scenes while Part_B contains relatively sparse scenes captured by surveillance cameras of realistic streets, which can demonstrate the effectiveness of our network components for both kinds of cases.

Network architecture: Our proposed approach contains backbone network, dense dilated convolution block, dense residual connection and multi-scale density level consistency loss. To demonstrate their effectiveness, we conduct experiments by adding these components incrementally. The results are shown in Table 3.

We use the backbone network and density map estimator to be the baseline model. It achieves 70.9 MAE on ShanghaiTech Part_A and 15.21 MAE on Part_B, which are the worst across all entries in the table. By only adding the proposed DDCB incrementally to enrich the baseline model, the MAE decreases to 63.2 on ShanghaiTech Part_A and to 7.33 on ShanghaiTech Part_B, which improves by a big margin. This illustrates that the features with dense scales and

large receptive fields caused by incremental dense dilated convolution block are essential and beneficial to count crowd accurately and robustly. When our proposed network contains only two DDCBs, it has similar number of parameters as CSRNet [14], which also uses dilated convolution for crowd counting. However our network can achieve MAE of 64.8 and RMSE of 104.4 without utilizing dense residual connections and density level consistency loss, which achieves much better performance than CSRNet.

We then add dense residual connections between three dense dilated convolution blocks. These connections also bring an improvement and the MAE further decreases to 62.6 on ShanghaiTech Part_A and 7.06 on ShanghaiTech Part_B as shown in Table 3, which indicates that the dense residual connections can enlarge the scale diversity by reusing features from different DDCB.

Finally, we incorporate the multi-scale density level consistency loss to train the whole network. It further decreases the MAE to 61.3 on ShanghaiTech Part_A and 6.74 on ShanghaiTech Part_B, which is the best performance of all the variations. It demonstrates that the proposed loss can enforce the density level of estimated density map to be consistent with ground truth at global and local level.

Type of Connections between DDCBs: We also invest the influence of dense residual connections by additionally testing the performance of networks without connections and with ordinary residual connections. The experimental results are reported in Table 4. By utilizing residual connections, the estimated error decreases to 61.8 on ShanghaiTech Part_A and 6.81 on ShanghaiTech Part_B due to the reused features from the previous one block. However it ignores the features from other blocks with different scales. To address the issue, dense residual connections are adopted to further decrease the MAE to 61.3 on ShanghaiTech Part_A and 6.74 on ShanghaiTech Part_B, which indicates that scale diversity is further enlarged and features are more effective.

Combination of Dilation Rates: To choose the best combination of dilation rates in DDCB, we also perform an ablation study by comparing the performance of selected one with that of the other two. As shown in Table 5, the selected one, i.e. the combination 1-2-3, achieves the best performance among the three combinations. The other two combinations are adopted by following CSRNet [14] and DenseASPP [40] respectively. According to the reported results, the carefully selected dilation rates that provide more densely sampled scale range and overcome the gridding artifacts can extract more abundant feature information for crowd counting, while the other two combinations that lose some information because of gridding artifacts perform worse effect on ShanghaiTech. Interestingly, the combination of dilation rates 2-2-2 loses less information of original features than the combination 3-6-12 and acquires better results, which validates our analysis that gridding artifact is harmful for crowd counting task.

Loss function: Our proposed novel loss adopts a configuration with three scale levels (i.e. 1×1 , 2×2 , 4×4 output size of average pooling operation). We conduct experiments with these three levels to present that every scale level can regularize the consistence between the estimated density map and ground truth. Experimental results are presented in Table 6. Before we add the multi-scale density level consistency loss, the proposed network achieves 62.6 MAE for ShanghaiTech Part_A and 7.06 MAE for ShanghaiTech

Table 5: The performance of our network on ShanghaiTech Dataset when different dilation rates are selected. The numbers in the first column are the dilation rates in each DDCB.

Dilation Rates	ShanghaiTech A		ShanghaiTech B	
	MAE↓	RMSE↓	MAE↓	RMSE↓
2-2-2	62.1	100.1	6.85	11.45
3-6-12	62.7	103.0	6.96	11.93
1-2-3	61.3	97.3	6.74	10.48

Table 6: The performance of our network on ShanghaiTech Dataset when different scale levels are set in consistency loss L_c . The number in the first column is the output size of average pooling operation.

Loss	ShanghaiTech A		ShanghaiTech B	
	MAE↓	RMSE↓	MAE↓	RMSE↓
w/o L_c	62.6	102.8	7.06	12.01
Level 1	62.3	101.4	6.95	11.38
Level 1,2	61.9	100.8	6.88	11.35
Level 1,2,4	61.3	97.3	6.74	10.48

Part_B. By adopting single level with the output size of 1×1 , which is the global context that represents the density level of the whole input image, the MAE decreases to 62.3 for ShanghaiTech Part_A and 6.95 for ShanghaiTech Part_B. Furthermore, the performance is continued to be improved due to the influence of the local levels with the output size of 2×2 and 4×4 that decreases the MAE to 61.9 and 61.3 individually for ShanghaiTech Part_A as well as 6.88 and 6.74 for ShanghaiTech Part_B. All these incremental experiments indicates that both global and local regularization of density levels can help the estimated density maps to be consisted with the ground truth maps at different scale levels so that high-quality density map can be generated.

6 CONCLUSION

In this paper, we propose a novel end-to-end single-column model called DSNet to estimate crowd count accurately, based on the dense dilated convolution blocks with dense residual connections. These two components enlarge scale diversity and receptive field of features that can handle the issue of large scale variations to perform well on counting the number of people in images. We further introduce a novel loss to enforce the density level of estimated density maps to be consisted with corresponding ground truth maps at different scale levels. Our proposed approach achieves state-of-the-art results on five public challenging crowd counting datasets, on all evaluation metrics.

ACKNOWLEDGMENTS

This work was supported by National Natural Science Foundation of China (61771458).

REFERENCES

- [1] Deepak Babu Sam, Neeraj N Sajjan, R Venkatesh Babu, and Mukundhan Srinivasan. 2018. Divide and grow: capturing huge diversity in crowd images with incrementally growing CNN. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 3618–3626.
- [2] Lokesh Boominathan, Srinivas SS Kruthiventi, and R Venkatesh Babu. 2016. Crowdnet: A deep convolutional network for dense crowd counting. In *Proceedings of the ACM International Conference on Multimedia*. ACM, 640–644.
- [3] Xinkun Cao, Zhipeng Wang, Yanyun Zhao, and Fei Su. 2018. Scale aggregation network for accurate and efficient crowd counting. In *Proceedings of the European Conference on Computer Vision*. 734–750.
- [4] Antoni B Chan, Zhang-Sheng John Liang, and Nuno Vasconcelos. 2008. Privacy preserving crowd monitoring: Counting people without people models or tracking. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. IEEE, 1–7.
- [5] Antoni B Chan and Nuno Vasconcelos. 2009. Bayesian poisson regression for crowd counting. In *Proceedings of the IEEE International Conference on Computer Vision*. IEEE, 545–551.
- [6] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. 2011. Deep sparse rectifier neural networks. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*. 315–323.
- [7] Dan Guo, Kun Li, Zheng-Jun Zha, and Meng Wang. 2019. DADNet: Dilated-Attention-Deformable ConvNet for Crowd Counting. In *Proceedings of the ACM International Conference on Multimedia*. 1823–1832.
- [8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 770–778.
- [9] Haroon Idrees, Imran Saleemi, Cody Seibert, and Mubarak Shah. 2013. Multi-scale multi-scale counting in extremely dense crowd images. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 2547–2554.
- [10] Haroon Idrees, Muhammad Tayyab, Kishan Athrey, Dong Zhang, Somaya Al-Maadeed, Nasir Rajpoot, and Mubarak Shah. 2018. Composition loss for counting, density map estimation and localization in dense crowds. In *Proceedings of the European Conference on Computer Vision*. 532–546.
- [11] X. Jiang, Z. Xiao, B. Zhang, X. Zhen, X. Cao, D. Doermann, and L. Shao. 2019. Crowd Counting and Density Estimation by Trellis Encoder-Decoder Networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 6126–6135.
- [12] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint:1412.6980* (2014).
- [13] Victor Lempitsky and Andrew Zisserman. 2010. Learning to count objects in images. In *Proceedings of Conference on neural information processing systems*. 1324–1332.
- [14] Yuhong Li, Xiaofan Zhang, and Deming Chen. 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.
- [15] Lingbo Liu, Zhilin Qiu, Guanbin Li, Shufan Liu, Wanli Ouyang, and Liang Lin. 2019. Crowd Counting with Deep Structured Scale Integration Network. In *Proceedings of the IEEE Conference on Computer Vision (ICCV)*.
- [16] N. Liu, Y. Long, C. Zou, Q. Niu, L. Pan, and H. Wu. 2019. ADCrowdNet: An Attention-Injective Deformable Convolutional Network for Crowd Understanding. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 3220–3229.
- [17] Xiyang Liu, Jie Yang, and Wenrui Ding. 2020. Adaptive Mixture Regression Network with Local Counting Map for Crowd Counting. In *Proceedings of the European Conference on Computer Vision*.
- [18] David G Lowe. 1999. Object recognition from local scale-invariant features. In *Proceedings of the IEEE International Conference on Computer Vision*. IEEE, 1150.
- [19] Ao Luo, Fan Yang, Xin Li, Dong Nie, Zhicheng Jiao, Shangchen Zhou, and Hong Cheng. 2020. Hybrid Graph Neural Networks for Crowd Counting. In *Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI)*.
- [20] Z. Ma, X. Wei, X. Hong, and Y. Gong. 2019. Bayesian Loss for Crowd Count Estimation With Point Supervision. In *Proceedings of the IEEE International Conference on Computer Vision*. 6141–6150.
- [21] Daniel Onoro-Rubio and Roberto J López-Sastre. 2016. Towards perspective-free object counting with deep learning. In *Proceedings of the European Conference on Computer Vision*. Springer, 615–629.
- [22] Viet-Quoc Pham, Tatsuo Kozakaya, Osamu Yamaguchi, and Ryuzo Okada. 2015. Count forest: Co-voting uncertain number of targets using random forest for crowd density estimation. In *Proceedings of the IEEE International Conference on Computer Vision*. 3253–3261.
- [23] Viresh Ranjan, Hieu Le, and Minh Hoai. 2018. Iterative crowd counting. In *Proceedings of the European Conference on Computer Vision*. 270–285.
- [24] Roberto Lopez-Sastre Ricardo Guerrero-Gomez-Olmedo, Beatriz Torre-Jimenez and Daniel Onoro-Rubio. 2015. Extremely overlapping vehicle counting. In *In Iberian Conference on Pattern Recognition and Image Analysis (IbPRIA)*. 423–431.
- [25] Mikel Rodriguez, Ivan Laptev, Josef Sivic, and Jean-Yves Audibert. 2011. Density-aware person detection and tracking in crowds. In *Proceedings of the IEEE International Conference on Computer Vision*. IEEE, 2423–2430.
- [26] David Ryan, Simon Denman, Clinton Fookes, and Sridha Sridharan. 2009. Crowd counting using multiple local features. In *Digital Image Computing: Techniques and Applications*. IEEE, 81–88.

- [27] Deepak Babu Sam, Shiv Surya, and R Venkatesh Babu. 2017. Switching convolutional neural network for crowd counting. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. IEEE, 4031–4039.
- [28] Chong Shang, Haizhou Ai, and Bo Bai. 2016. End-to-end crowd counting via joint learning local and global count. In *Proceedings of the IEEE International Conference on Image Processing*. IEEE, 1215–1219.
- [29] M. Shi, Z. Yang, C. Xu, and Q. Chen. 2019. Revisiting Perspective Information for Efficient Crowd Counting. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 7271–7280.
- [30] Zenglin Shi, Le Zhang, Yun Liu, Xiaofeng Cao, Yangdong Ye, Ming-Ming Cheng, and Guoyan Zheng. 2018. Crowd counting with deep negative correlation learning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 5382–5390.
- [31] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint:1409.1556* (2014).
- [32] Vishwanath A Sindagi and Vishal M Patel. 2017. Cnn-based cascaded multi-task learning of high-level prior and density estimation for crowd counting. In *IEEE International Conference on Advanced Video and Signal Based Surveillance*. IEEE, 1–6.
- [33] Vishwanath A Sindagi and Vishal M Patel. 2017. Generating high-quality crowd density maps using contextual pyramid cnns. In *Proceedings of the IEEE International Conference on Computer Vision*. 1861–1870.
- [34] Xin Tan, Chun Tao, Tongwei Ren, Jinhui Tang, and Gangshan Wu. 2019. Crowd Counting via Multi-Layer Regression. In *Proceedings of the ACM International Conference on Multimedia*. 1907–1915.
- [35] Y. Tian, Y. Lei, J. Zhang, and J. Z. Wang. 2019. PaDNet: Pan-Density Crowd Counting. *IEEE Transactions on Image Processing* (2019).
- [36] Elad Walach and Lior Wolf. 2016. Learning to count with CNN boosting. In *Proceedings of the European Conference on Computer Vision*. Springer, 660–676.
- [37] Meng Wang and Xiaogang Wang. 2011. Automatic adaptation of a generic pedestrian detector to a specific traffic scene. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. IEEE, 3401–3408.
- [38] Bo Wu and Ram Nevatia. 2005. Detection of multiple, partially occluded humans in a single image by bayesian combination of edgelet part detectors. In *Proceedings of the IEEE International Conference on Computer Vision*. IEEE, 90–97.
- [39] Z. Yan, Y. Yuan, W. Zuo, X. Tan, Y. Wang, S. Wen, and E. Ding. 2019. Perspective-Guided Convolution Networks for Crowd Counting. In *Proceedings of the IEEE International Conference on Computer Vision*. 952–961.
- [40] Maoke Yang, Kun Yu, Chi Zhang, Zhiwei Li, and Kuiyuan Yang. 2018. DenseASPP for semantic segmentation in street scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3684–3692.
- [41] Cong Zhang, Hongsheng Li, Xiaogang Wang, and Xiaokang Yang. 2015. Cross-scene crowd counting via deep convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 833–841.
- [42] Lu Zhang, Miaoqing Shi, and Qiaobo Chen. 2018. Crowd counting via scale-adaptive convolutional neural network. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*. IEEE, 1113–1121.
- [43] Yingying Zhang, Desen Zhou, Siqin Chen, Shenghua Gao, and Yi Ma. 2016. Single-image crowd counting via multi-column convolutional neural network. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*.