DADNet: Dilated-Attention-Deformable ConvNet for Crowd Counting

Dan Guo[1], Kun Li[1*], Zheng-Jun Zha[2], Meng Wang[1]
[1] School of Computer Science and Information Engineering, Hefei University of Technology, Hefei, China.
[2] University of Science and Technology of China, Hefei, China.
guodan@hfut.edu.cn,kunli.hfut@gmail.com,zhazj@ustc.edu.cn,eric.mengwang@gmail.com

ABSTRACT

Most existing CNN-based methods for crowd counting always suffer from large scale variation in objects of interest, leading to density maps of low quality. In this paper, we propose a novel deep model called Dilated-Attention-Deformable ConvNet (DADNet), which consists of two schemes: multi-scale dilated attention and deformable convolutional DME (Density Map Estimation). The proposed model explores a scale-aware attention fusion with various dilation rates to capture different visual granularities of crowd regions of interest, and utilizes deformable convolutions to generate a high quality density map. There are two merits as follows: (1) varying dilation rates can effectively identify discriminative regions by enlarging the receptive fields of convolutional kernels upon surrounding region cues, and (2) deformable CNN operations promote the accuracy of object localization in the density map by augmenting the spatial object location sampling with adaptive offsets and scalars. DADNet not only excels at capturing rich spatial context of salient and tiny regions of interest simultaneously, but also keeps a robustness to background noises, such as partially occluded objects. Extensive experiments on benchmark datasets verify that DADNet achieves the state-of-the-art performance. Visualization results of the multi-scale attention maps further validate the remarkable interpretability achieved by our solution.

CCS CONCEPTS

• **Computing methodologies** → **Computer vision**; **Machine learning**; *Interest point and salient region detections*; *Object detection*.

KEYWORDS

Crowd counting; density map estimation; scale-aware attention; dilated convolution; deformable convolution

*Corresponding author.

MM '19, October 21–25, 2019, Nice, France

ACM ISBN 978-1-4503-6889-6/19/10…$15.00
https://doi.org/10.1145/3343031.3350881

ACM Reference Format:
Dan Guo[1], Kun Li[1*], Zheng-Jun Zha[2], Meng Wang[1] . 2019. DADNet: Dilated-Attention-Deformable ConvNet for Crowd Counting. In *Proceedings of the 27th ACM International Conference on Multimedia (MM '19), October 21–25, 2019, Nice, France*. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3343031.3350881

1 INTRODUCTION

Crowd counting aiming at estimating the number of people in images, has attracted extensive attention due to the potential important applications in the real world, such as public security, traffic monitoring and video surveillance. This task still remains challenging due to the existence of scale variations, object occlusions, and background noises in the crowd scene.

Recently, convolutional neural network (CNN) based approaches have been widely explored for crowd counting. Some works [1, 26, 31, 40] have achieved significant successes by adopting the multi-column network architecture. For example, MCNN [40] is a typical multi-column architecture, which showed strong adaptability to discover variable-size objects (heads) using convolutions with different kernel sizes in each column. Based on the multi-column architecture, Sam *et al.* [1, 26] designed an auxiliary switch-CNN layer to vote the best appropriate CNN regressor column for each image patch for the density map generation. Moreover, Sindagi *et al.* [31] proposed the Global Context Estimator (GCE), Local Context Estimator (LCE), and Fusion-CNN modules to explicitly incorporate global and local contextual information of crowd images. Besides, a single-column structure network based on dilated convolutions is proposed [18], which deployed dilated convolutional layers for progressively extracting deeper contextual information of saliency. The CSRNet model [18] expanded the receptive field without losing resolution by the dilated convolutions.

In addition, a close task to crowd counting is the problem of vehicle counting [8, 22]. Vehicle counting aims at precisely estimate the number of vehicles in traffic congestion scenes [22, 39]. Onoro-Rubio *et. al.* [22] proposed an MCNN-like network called Hydra-3s to generate the density map. Zhang *et al.* [39] used a combination of Fully Convolutional Neural network (FCN) and Long Short-Term Memory network (LSTM) to estimate the vehicle density and count number jointly. Besides, the CSRNet model [18] for crowd counting has also verified its effectiveness on the vehicle counting dataset TRANCOS [8].

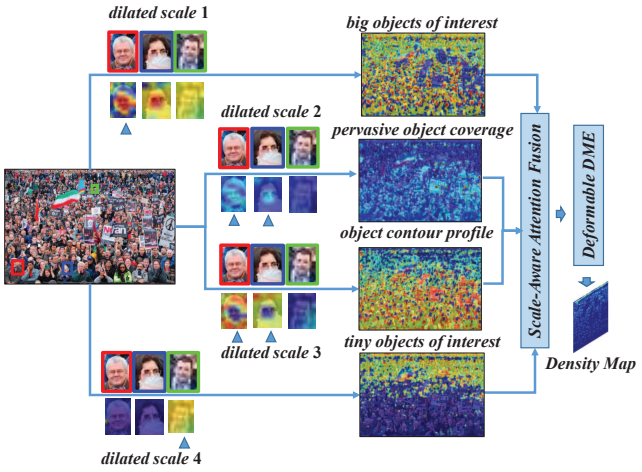


Figure 1: Illustration of scale-aware dilated attention fusion in the proposed DADNet. The screenshot of each cropped image subregion (*i.e.*, red, blue and green subregions) responses different attention maps under various dilation scales. Utilizing the complementary of all these attention maps is helpful to exactly locate objects in the crowd scenes.

Although these methods achieve remarkable progress, there are remains two defects: (1) they ignore the attention map differences among multi-scale feature maps; and (2) even they consider and fuse all these multi-scale feature maps, but always concatenate them directly [40]. This simple fusion strategy can not adapt to the complex issues of scale variations, object occlusions, and background noises in crowd scenes. In the paper, we propose a novel network framework called Dilated-Attention-Deformable ConvNet (DADNet), which designs adaptive dilated-CNN attention fusion and adaptive deformable-CNN DME (Density Map Estimation) modules to address the above issues.

As an illustration example shown in Figure 1, under the dilated scale 1 (dilation rate r=1), the 2D attention map highlights the big heads of interest, *e.g.*, the nearby red box subregion; while the dilated scale 4 (dilation rate r=9) highlights the small heads of interest, *e.g.*, the distant green box subregion. The attention maps at dilated scales 2 (dilation rate r=3) and 3 (dilation rate r=6) reflect the pervasive head coverages and the contour information of heads, respectively. The latter two dilated scales alleviate the noise issue in occlusion situations (*e.g.*, the blue box subregions). DADNet jointly learns the context variety under different dilated scale views and delivers a three-layer deformable DME to generate a high-quality density map.

In detail, as shown in Figure 2, the DADNet is realized by three steps. First, low-level feature maps of an image are extracted by a modified VGG-16 backbone network [29]. Secondly, a scale-aware attention fusion using dilated convolution is designed to discover the visual responses of big objects, tiny objects, and contour profile of objects in crowd or traffic scenes effectively. Finally, the accuracy of object localization is further promoted by utilizing a deformable convolution with adaptive receptive fields on object locations, which promises high-quality density maps. The main contributions of this paper are summarized as follows:

- The proposed DADNet generates high-quality density maps by effectively learning visual context cues of multi-scale features, which shows strong adaptability to resist scale variations, object occlusions, and background noises in crowd image.
- DADNet consists of a scale-aware attention fusion and a deformable DME. The former utilizes an adaptive 2D attention map mechanism on multi-scale features for exact visual representation, while the latter augments the flexibility of spatial sampling locations of objects with learnable offsets and scalars.
- Extensive experiments on three crowd counting benchmark datasets (*i.e.*, ShanghaiTech, UCF_CC_50, UCF-QNRF) and one vehicle counting dataset (TRANCOS) achieve the state-of-the-art performance. Ablation studies demonstrate the effectiveness of each module within the proposed model.

2 RELATED WORK

Traditional methods. Early approaches usually counted the number of people by detecting heads or bodies in the images. Most of them focused on extracting hand-crafted features (*e.g.*, Haar wavelets [34] and HOG [5]) from the human body or particular body parts [33]. Dollar *et al.* [6] used a slide window detector over the image to count the person. All these detection-based approaches are not applicable when crowds are extremely congested.

Meanwhile, some researchers explored regression-based approaches for crowd counting, which aimed at training regression models to directly map the visual features to the number of people. Idress *et al.* [12] proposed a regression model to learn multiply features (*i.e.*, head detection and SIFT [21] interest points) for extremely dense crowd images. They adopted the object counting value for optimization training, while neglecting the location information of objects. This training process needed massive image samples. Furthermore, to solve the sparse and imbalanced training data, Chen *et al.* [2] proposed a novel cumulative attribute-based regression model to effectively capture the scalar variation of objects, which showed a promising performance on sparse data and imbalanced training data yet suffering an expensive computational cost. By contrast, Lempitsky *et al.* [16] first introduced a density map to estimate the counting number accurately. They adopted a linear model to map the input image to the density map. Pham *et al.* [24] used a random forest regression to learn the non-linear mapping instead of the linear mapping.

CNN-based methods. Inspired by the success of deep CNN in the computer vision community, a variety of CNN-based approaches have been proposed for crowd counting. Wang *et al.* [35] designed a CNN-based regression network

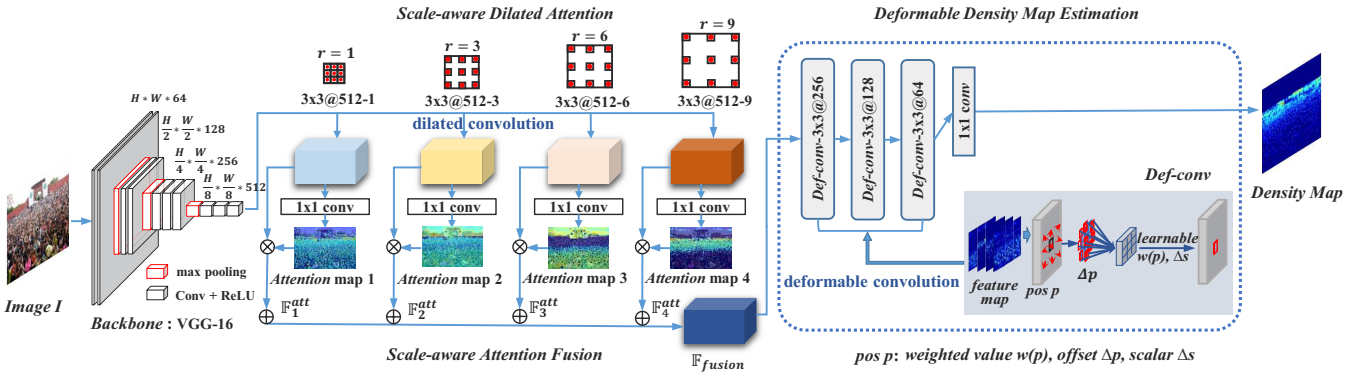


Figure 2: The architecture of the proposed DADNet. The convolutional parameter is denoted as "(kernel size)×(kernel size)@(filter size)-(dilation rate)" in the dilated convolutions, while "Def-conv-(kernel size)×(kernel size)@(filter size)" in the deformable convolutions.

to extract convolutional features and compared it with traditional hand-crafted features. Zhang *et al.* [38] designed a deep CNN network with two alternative loss objectives on both density map and counting number evaluations to address a novel cross-scene counting problem. With the alternative training, the proposed model was fine-tuned and adapted to new target scenarios. To handle the scale variance of crowds, researchers further employed a CNN-based multi-column architecture with different convolutional kernel sizes in multiple branches. Zhang *et al.* [40] proposed a Multi-column Convolutional Neural Network (MCNN) to tackle arbitrary image size inputs. Based on the MCNN framework, Sam *et al.* [1, 26] proposed a Switch-CNN layer to vote the best appropriate CNN regressor column for each image patch for the density map generation.

More recently, many advanced methods have been used to improve density map quality. Shen *et al.* [27] explored a Generative Adversarial Network (GAN) model to constrain the cross-scale density consistency in the crowd estimation. Shi *et al.* [28] designed a Negative Correlation Learning (NCL) based ensemble network for the density map generation. Liu *et al.* [20] proposed a self-supervised approach by incorporating unlabelled crowd images during training. As the attention mechanism has been widely used in various computer vision tasks, Liu *et al.* [19] designed a framework named DecideNet to introduce visual attention for crowd counting. MA Hossain *et al.* [10] proposed a joint attention network to model both global and local visual contextual information.

Dilated convolutions. The dilated convolution operation is widely used in a variety of vision tasks, such as semantic segmentation [3, 37], image de-raining [17], and object detection [32]. Both Yu *et al.* [37] and Chen *et al.* [3] designed dilated-CNN based convolution networks to capture contextual features to segment objects robustly. Li *et al.* [17] proposed a recurrent feature aggregation module involving dilation convolution operations for image de-raining. Recently, Song *et al.* [32] utilized a pyramid dilated convolution module to extract multi-scale spatial features for the salient object detection simultaneously. The most related to ours is [18] for crowd counting. The CSRNet model [18] based on progressively dilated convolutions expanded contextual information in the back-end DME module. Specifically, it [18] adopted a fixed dilation rate.

Contrary to previous approaches, the paper proposes a dilated-CNN based multi-column (scale) architecture to generate a high-quality density map. The dilated-CNN with various dilation rates contributes to learning the rich contextual information for the front-end feature representation. Moreover, an adaptive deformable convolution is introduced in the back-end DME module to precisely locate objects' positions.

3 PROPOSED METHOD

In this paper, we propose a Dilated-Attention-Deformable ConvNet (DADNet) for crowd counting. As illustrated in Figure 2, DADNet mainly consists of two modules: a scale-aware attention module and a deformable density map estimation module. We first employ the first 10 layers of VGG-16 as the backbone which generates the VGG feature maps of each image. Based on the VGG feature maps, DADNet captures various visual attention granularities of crowd regions of interest with different dilation scale convolutions, and implements a scale-aware attention fusion. Finally, it feeds the fused feature maps into the deformable Density Map Estimation (DME) module, which utilizes deformable convolutions by augmenting the spatial sampling locations with learnable offsets and scalars to generate high-quality density maps. Specifically, DADNet is an attention-injective deformable ConvNet, which addresses the large variations in object sizes with various dilation scales.

3.1 Scale-aware Attention Fusion

Current top-down CNN-based methods [18, 40] usually focus on discriminative object regions and ignore non-discriminative areas. In this paper, we devote to realizing the dense object location correlation by transferring spatial information from discriminative regions to adjacent non-discriminative regions.

Figure 3: Visualization of multi-scale dilated attention maps of three images in the ShanghaiTech Part A dataset. As for crowd counting, Figures (a), (b), (c), and (d) reflect different attention appearances on the big objects of interest (nearby people's heads, r=1), pervasive areas of objects (r=3), contour profiles of objects (r=6), and tiny objects of interest (distant people's heads, r=9), respectively.

To this end, the dilated convolution, also called atrous convolution, is adopted, which provides a larger receptive field than normal convolution operations and remains the unchanged number of model parameters. Given a kernel size $k \times k$ and a dilation rate r, the receptive field of a dilated convolution operation is enlarged to $k + (k-1)(r-1)$. In other words, it can capture border surrounding areas and richer context information than normal convolution units. The head areas of people in an image always vary drastically in the crowd scenes. A unitary receptive field can not adapt to handle the crowd density variation.

To address the problem of crowd density variation, we design a multi-scale dilated convolution attention module as shown in Figure 2. We use different dilation scales to discover visual context cues. The core idea is to enable different visual context cues to perform the spatial referring on non-discriminative areas. In this paper, we set the number of varying scales $S = 4$ with corresponding dilation rate $r \in \{1, 3, 6, 9\}$. At each dilated rate (r_i, $i \in [1, S]$), we calculate dilated feature maps by $\mathbb{F}_{r_i} = \mathcal{F}_r(\mathbb{F}_{vgg})$, where $\mathbb{F}_{vgg}$ denotes the VGG feature maps and $\mathcal{F}_r$ denotes the dilated convolutional operation on $\mathbb{F}_{vgg}$.

Based on feature maps $\{\mathbb{F}_{r_i}\}(i \in [1, S])$, we conduct a scale-aware fusion based on an attention weighting mechanism to fuse different feature branches. The fusion measures the contribution of the spatial context under different dilation scales. The corresponding 2D attention map $\mathbb{I}_{r_i}$ of feature map $\mathbb{F}_{r_i} \in \mathbb{R}^{H \times W \times \#ch}$ is formulated as:

$$\mathbb{I}_{r_i} = Sigmoid(\mathcal{F}_{\{1\times1\}}(\mathbb{F}_{r_i}, \Theta_{\mathcal{F}})) \in \mathbb{R}^{H \times W}, \quad (1)$$

where $H \times W$ is the dimension of feature maps, $\#ch$ is the channel number, $Sigmoid$ denotes the sigmoid activation function, $\mathcal{F}_{\{1\times1\}}$ denotes the 1×1 convolution operation, and $\Theta_{\mathcal{F}}$ is the model parameters of $\mathcal{F}_{\{1\times1\}}$.

To make a scale-aware attention fusion, we normalize the attention maps $[\mathbb{I}_1, \ldots, \mathbb{I}_S]$ at each scale. At dilation rate r_i, the normalized 2D attention map $\mathbb{W}_{r_i}$ is defined as follow:

$$\mathbb{W}_{r_i} = \mathbb{I}_{r_i} ./ \sum\nolimits_{r_i=1}^{S} \mathbb{I}_{r_i} \in \mathbb{R}^{H \times W}, \quad (2)$$

where "./" is the element-wise division operation.

Finally, we employ the scale-aware attention maps $[\mathbb{W}_{r_1}, \ldots, \mathbb{W}_{r_S}]$ to generate the fused feature maps $\mathbb{F}_{fusion}$:

$$\mathbb{F}_{fusion} = \sum_{i=1}^{S} \mathbb{F}_i^{att} = \sum_{i=1}^{S} \mathbb{F}_{r_i} \odot \mathbb{W}_{r_i} \in \mathbb{R}^{H \times W \times \#ch}, \quad (3)$$

where $\odot$ means element-wise product operation, and $\mathbb{F}_i^{att}$ denotes the scale-aware feature map at dilation scale i. Note that the feature dimensions of $\mathbb{F}_i^{att}$ and $\mathbb{F}_{fusion}$ are the same as $\mathbb{F}_{vgg}$.

Figure 3 depicts the normalized multi-scale dilated attention maps $\{\mathbb{W}_{r_i}\}|r_i \in \{1, 3, 6, 9\}$ of three image samples for crowd counting. We can see that DADNet captures different attention responses effectively, which is adaptive to tackle diversified crowd distributions, complex backgrounds, and various occlusions in crowd scenes. From a theoretical view, the solution of scale-aware dilated attention is flexible, which can be extended to arbitrary branches.

3.2 Deformable Density Map Estimation

After adaptively fusing feature maps in Section 3.1, in this subsection, a deformable DME is proposed to modulate the accurate spatial transformation of the object locations in the to-be-generated density map. As illustrated in Figure 2, the

DME module consists of a three-layer deformable convolution (*def-conv*) network, in which the sampling location weight $\mathbf{w}(\mathbf{p})$, scalar $\Delta\mathbf{s}$ and offset $\Delta\mathbf{p}$ on each *def-conv* layer are all the learning parameters. With the help of these parameters, the convolutional grid can self-adapt to obtain useful location cues on the fused feature map to generate the high-quality density map.

As a basic convolution, the sampling location $\mathbf{p}_k$ with a convolution kernel of 3×3 can be expressed as $\mathbf{p}_k \in \mathcal{K} = \{(-1,-1),(0,-1),\ldots,(1,0),(1,1)\}$. For a location $\mathbf{p}$ in the input feature map $\mathbf{x}$, the output feature map $\mathbf{y}(\mathbf{p})$ is formulated as:

$$\mathbf{y}(\mathbf{p}) = \sum_{k=1}^{\mathcal{K}} \mathbf{w}(\mathbf{p}_k) \cdot \mathbf{x}(\mathbf{p} + \mathbf{p}_k) \tag{4}$$

where $\mathbf{w}(\mathbf{p})$ weighted the summation of sampled values [4].

Different from the uniform sampling with the fixed $\mathbf{p}_k$ in normal convolutions, the adaptive learnable offset $\Delta\mathbf{p}$ and modulation scalar $\Delta\mathbf{s}$ are added. Both $\Delta\mathbf{s}$ and $\Delta\mathbf{p}$ in the deformable convolution can be optimized via training [4, 41]. In the paper, we adopt the deformable convolution in [41]. With the same sampling location set $\mathcal{K}$, the output feature map $\mathbf{y}(\mathbf{p})$ in the deformable convolution is expressed as follow:

$$\mathbf{y}(\mathbf{p}) = \sum_{k=1}^{\mathcal{K}} \mathbf{w}(\mathbf{p}_k) \cdot \mathbf{x}(\mathbf{p} + \mathbf{p}_k + \Delta\mathbf{p}_k) \cdot \Delta\mathbf{s}_k\,, \tag{5}$$

where $\Delta\mathbf{p}_k$ and $\Delta\mathbf{s}_k$ denote the offset and the modulation scalar at the k-th location in $\mathcal{K}$, respectively.

Technically speaking, the deformable DME is realized as follows. We use three consecutive *def-covn* layers with the same kernel size of 3×3 and add the ReLU activation [15] function after each *def-covn* layer to continuously refine the to-be-generated density map. Finally, we use 1×1 convolution on the *def-covn* layer to generate the density maps. This kind of dynamic sampling scheme is appropriate for crowd counting, especially for crowd scenes with some congested noises. The proposed DME adjusts the sampling locations in the image automatically, which promotes the accuracy of object location in the to=be-generated density map.

3.3 Loss Optimization

We adopt the pixel-wise mean square error (MSE) loss as the optimization objective, which measures the difference between a predicted density map and the groundtruth. Given an image I_i, the model parameter Θ of DADNet is optimized as follow:

$$Loss = \frac{1}{2B} \sum_{i=1}^{B} \|F(I_i, \Theta) - Y_i\|_2^2\,, \tag{6}$$

where B is the batch size, $F(I_i, \Theta)$ denotes the predicted density map, and Y_i is the groundtruth.

4 EXPERIMENTS

4.1 Experiment Setup

4.1.1 Datasets. In this paper, we test and verify the proposed model on both crowd counting and vehicle counting datasets.

ShanghaiTech [40] is divided into Part A and Part B. Part A includes 482 images downloaded from the Internet, while Part B contains 716 images which are shot from streets in Shanghai. Images in Part A are more crowded than in Part B. Parts A and B contain 300/182 and 400/316 images for training/testing sets, respectively.

UCF_CC_50 [12] contains only 50 images collected from the Internet. The number of people ranges from 94 to 4,543 in each image, and the average count is 1,280. The limited training data and a wide margin of the crowd density variation make the dataset extremely challenging. For a fair comparison, experiments on this dataset are conducted with the standard settings in [12], *i.e.*, a 5-fold cross-validation.

UCF-QRNF [13] is a new and the largest dataset for crowd counting, which contains 1,535 images. The minimum and maximum numbers of people are 49 and 12,865, respectively. And the average count is 815. The UCF-QRNF dataset has the largest crowd density variation.

TRANCOS [8] is a vehicle counting dataset collected from different traffic congested scenes with various viewpoints. It consists of 1,244 images captured by surveillance cameras. The regions of interest (ROI) in each image are annotated.

4.1.2 Evaluation metrics. We adopt five metrics as follows: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM) [36], and Grid Average Mean Absolute Error (GAME) [8].

Given an image I, the groundtruth density map Y is obtained by the method in [31]. Y is calculated by convolving each pixel with a Gaussian kernel, which is formalized as follows:

$$Y = \sum_{x_i \in Q} \delta(x - x_i) \times G_{\mu,\sigma^2}(x)\,, \tag{7}$$

where x is a pixel position in image, x_i is the i-th head position in the annotation set Q. μ and σ represent the kernel size and standard deviation parameters of Gaussian kernel (G_{μ,σ^2}), respectively. We set $\mu = 15$ and $\sigma = 4$ for all datasets. Besides, as the size of feature maps is cut down to 1/8 of the input image, the same is the density map. We utilize bilinear interpolation [18] to resize it to the size of the groundtruth.

The most commonly used metrics MAE and RMSE are calculated as follows:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|\tilde{Y}_i - \hat{Y}_i|\,,\ RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}|\tilde{Y}_i - \hat{Y}_i|^2}\,, \tag{8}$$

where N is the number of test images, $\tilde{Y}_i$ and $\hat{Y}_i$ denote the predicted and the groundtruth counts of the i-th test image, respectively. $\tilde{Y}_i$ is calculated by summing over the predicted density map.

Besides, the GAME metric is used to evaluate the TRANCOS dataset. GAME(L) splits the density map into 4^L non-overlapping subregions and calculates the MAE value in each subregion, where L is the split level. The sum of the MAE values in all these subregions is the GAME value. GAME is equal to the MAE metric when $L = 0$.

$$GAME(L) = \frac{1}{N}\sum_{i=1}^{N}(\sum_{l=1}^{4^L}|\tilde{Y}_i^l - \hat{Y}_i^l|)\,, \tag{9}$$

where $\tilde{Y}_i^l$ represents the predicted count of the i-th test image within region l, while $\hat{Y}_i^l$ is the groundtruth count. The higher L, the more restrictive the GAME metric will be [8]. In this paper, we experiment with $L = 0, 1, 2$, and 3.

4.1.3 Implementation details. The proposed model is implemented on the platform of PyTorch [23]. To promote the diversity and variety of training data, we crop four patches with a fixed size (256×256) at some random positions in each training image. The training patches are horizontally flipped with a probability of 0.5 for data augmentation. In the proposed model, we adopt the Adam optimizer [14] and set the batch size to 50, the initialized learning rate to $1e^{-5}$, and the dropout rate to 0.5 over every 50 epochs.

4.2 Ablation Study

Here we experiment and discuss the ablation study to investigate the effectiveness of each component within the proposed model.

- **DADNet w/o dil-cov** replaces the dilated convolutional operation by the normal convolutional operation in the scale-aware dilated attention module. To be fair, the kernel sizes of these convolutional layers are equivalently transformed to $\{3, 7, 13, 19\}$, respectively.
- **DADNet w/o def-cov** replaces the deformable convolution operation by the normal convolution operation in the DME module.
- **DADNet w/o scale-fusion** removes the scale-aware attention fusion in DADNet. Under this condition, we directly sum the feature maps of each dilated convolution branch.

As the experimental results are shown in Table 1 on the ShanghaiTech Part A dataset, the overall DADNet has obvious performance improvement compared with **DADNet w/o dil-cov** and **DADNet w/o def-cov**. We observe obvious increases at the MAE and RMSE values of **DADNet w/o dil-cov**. It verifies that the dialed convolution can increase the receptive field and context information, which improves the accuracy of the density map estimation. Compared with DADNet, the MAE value of **DADNet w/o def-cov** rises from 64.2 to a much bigger value 67.5. The reason is that the learnable parameters (offset $\Delta\mathbf{p}$ and scalar $\Delta\mathbf{s}$) in the deformable convolution are helpful to generate density maps. The parameters ensure the adaptive object location sampling in the surrounding spatial context. Besides, even with both the dilated and the deformable convolutions, **DADNet w/o scale-fusion** still has a performance drop compared to DADNet. It verifies that fusion effectively integrates different visual cues in the crowd scenes.

Table 1: Ablation study on the ShanghaiTech Part A dataset.

Model	MAE↓	RMSE↓
DADNet w/o dil-cov	65.5	105.5
DADNet w/o def-cov	67.5	106.5
DADNet w/o scale-fusion	66.7	104.2
DADNet	**64.2**	**99.9**

Table 2: Performance evaluation on the ShanghaiTech and UCF_CC_50 datasets.

Method	Part A		Part B		UCF_CC_50	
	MAE↓	RMSE↓	MAE↓	RMSE↓	MAE↓	RMSE↓
Zhang *et al.* [38]	181.8	277.7	32.0	49.8	467.0	498.5
MCNN [40]	110.2	173.2	26.4	41.3	377.6	509.1
CMTL [30]	101.3	152.4	20.0	31.1	322.8	341.1
Switching-CNN [26]	90.4	135.0	21.6	33.4	318.1	439.2
CP-CNN [31]	73.6	106.4	20.1	30.1	295.8	**320.9**
ACSCP [27]	75.7	102.7	17.2	27.4	291.0	404.6
Liu *et al.* [20]	73.6	112.0	13.7	21.4	337.6	434.3
D-ConvNet [28]	73.5	112.3	18.7	26.0	288.4	404.7
IG-CNN [1]	72.5	118.2	13.6	21.1	291.4	349.4
ic-CNN [25]	68.5	116.2	10.7	16.0	**260.9**	365.5
CSRNet [18]	68.2	115.0	10.6	16.0	266.1	397.5
DADNet	**64.2**	**99.9**	**8.8**	**13.5**	285.5	389.7

4.3 Main Comparison

4.3.1 Comparison results on crowd counting datasets. As shown in Table 2, our approach achieves the best MAE and RMSE values on the ShanghaiTech dataset, especially for Part A which have a great scale view variance in the crowd scenes. As for Part B, our approach achieves the best MAE of 8.8 and RMSE of 13.5 too. Experimental results on the UCF-QNRF dataset are summarized in Table 3. DADNet achieves the lowest MAE and RMSE, as well as a 14.2% improvement of MAE. These results indicate that DADNet adapts to the scale variability and has good robustness.

On the UCF_CC_50 dataset, our approach achieves comparable performances with the MAE of 285.5 and the RMSE of 389.7, which is better than most of the comparative methods except [25]and [31]. UCF_CC_50 is challenging due to its sparse training data, extremely dense crowds, and noise backgrounds. To solve the issues, some models have some additional data-processing. [25] used the original image and a low-resolution density map to optimize a high-resolution density map. [31] adopted both local and global context estimators to optimize the feature extraction. Our model is just a simple end-to-end network based on VGG features without additional data-processing.

By observing Figure 4, there are two interesting conclusions. (1) **Figure 4 (a)**. The persons in boxed regions are surrounded with many background noises. For examples, the color of clothes is similar to head, and the color of hand is the same to the face. It is difficult to locate the real head

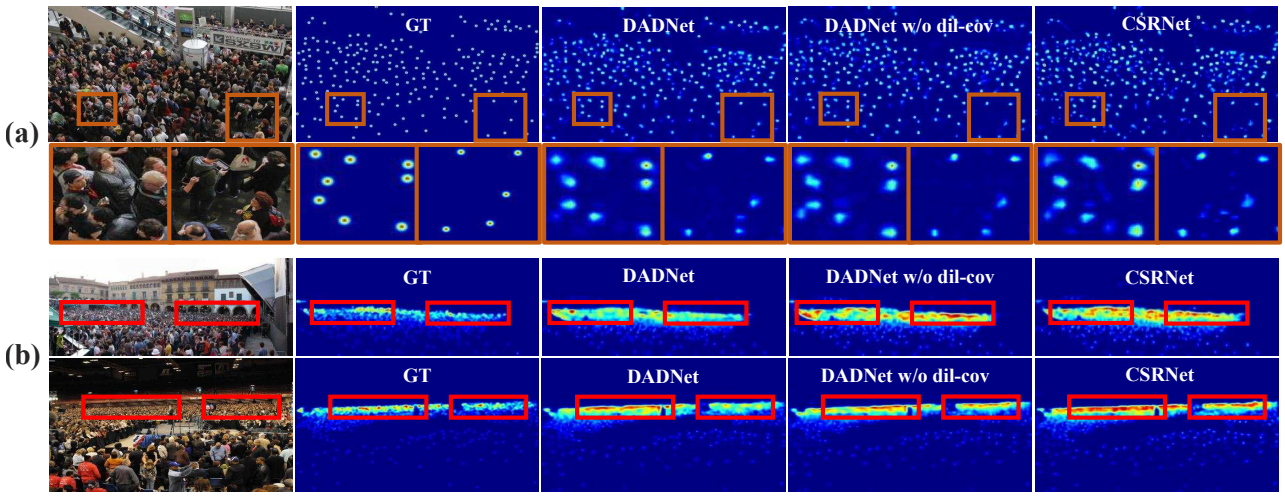


Figure 4: Comparison of the proposed DADNet model to CSRNet [18] and the groundtruth on the generated density maps. GT denotes the goundtruth. (a) Local subregion comparison. (b) Global dentistry map comparison.

Table 3: Performance evaluation on the UCF-QNRF dataset.

Model	MAE↓	RMSE↓
Idrees *et al.* [12]	315.0	508.0
MCNN [40]	277.0	426.0
ResNet101 [9]	190.0	277.0
CMTL [30]	252.0	514.0
Switching-CNN [26]	228.0	445.0
DenseNet201 [11]	163.0	226.0
Idrees *et al.* [13]	132.0	191.0
DADNet	**113.2**	**189.4**

location. Our density maps have high-quality, which are approximate to the groundtruth. Experimental results indicate that DADNet has good robustness to generate clear density maps. (2) **Figure 4 (b)**. Detecting tiny regions is tough. Some algorithms pay much more attention to context and have high response regions overly, resulting in the overestimation of human count. For example, CSRNet overestimates the real count, which adopted a single scale dilated convolution. It verifies that DADNet adapts to the scale variability. Our multi-scale dilated attention can extract richer visual context cues. In addition, Figure 5 demonstrates that DADNet has close count number to the groundtruth too. The predicted density map of the proposed model exhibits good clarity and accuracy in both crowded and sparse areas.

4.3.2 Comparison results on vehicle counting dataset. We make comparison with the state-of-the-art methods [7, 16, 18, 22, 39]. As shown in Table 4, our approach achieves the best performance, especially with a great improvement on GAME3. The higher $L = 3$, the more restrictive the GAME metric will be. Compared with the results of CSRNet [18], our approach is 21.6% lower on GAME0 (L =0), 19.7% lower

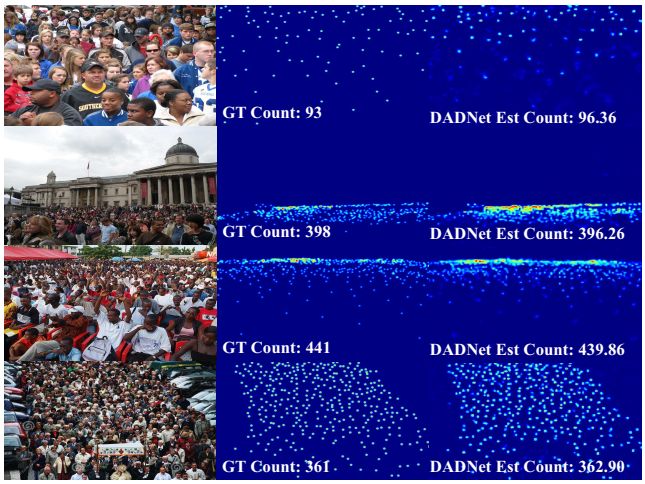
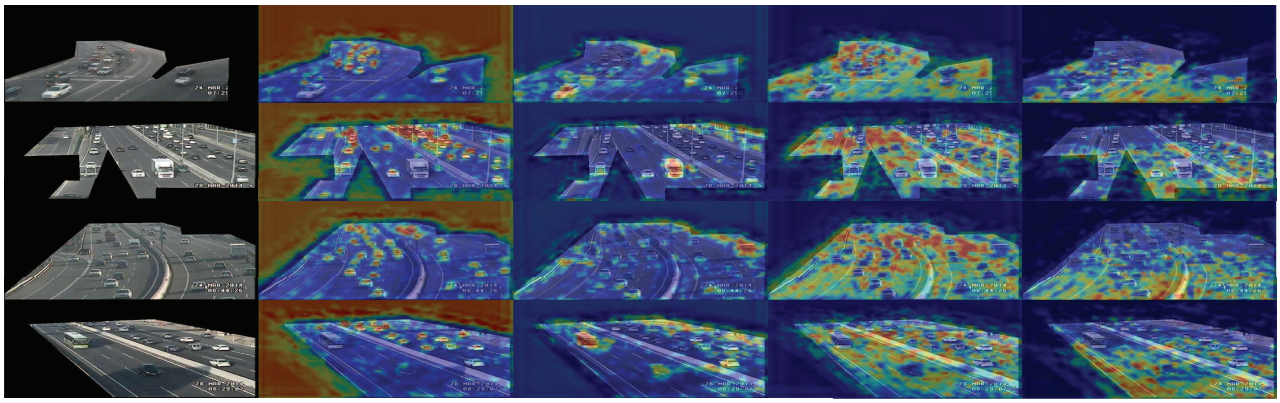


Figure 5: Visualization of density maps of four examples in the ShanghaiTech Part A dataset for crowd counting.

on GAME1 (L =1), 24.8% lower on GAME2 (L =2), and 38.4% lower on GAME3 (L =3).

As illustrated in Figures 6 and 7, DADNet generates high-quality density map for vehicle counting. Although the attention response regions in these attention maps are different from crowd counting, it still has a strong ability to identify different visual context cues. Vehicles are different from people heads. The vehicle dataset does not contain the noises of partial body occlusion and similar background as in the crowd datasets, *e.g.*, the similar skin color of other body parts (arms) to the face. The attention maps of vehicles mainly focus on pervasive coverage and contour profile of the objects, *i.e.*, the main body and surrounding areas of the cars, including nearby big cars and distant small cars. The



Input image (a) dilation rate r =1 (b) dilation rate r =3 (c) dilation rate r =6 (d) dilation rate r =9

Figure 6: Visualization of multi-scale dilated attention maps of four images in the TRANCOS dataset. As for vehicle counting, Figures (a), (b), (c), and (d) reflect different attention appearances on the tiny objects of interest (distant cars, r=1), big objects of interest (nearby cars, r=3), contour profiles of tiny objects (r=6) and big objects (r=9), respectively.

Table 4: Performance evaluation on the TRANCOS dataset.

Model	GAME0↓	GAME1↓	GAME2↓	GAME3↓
Fiaschi *et al.* [7]	17.77	20.14	23.65	25.99
Lempitsky *et al.* [16]	13.76	16.72	20.72	24.36
Hydra-3s [22]	10.99	13.75	16.69	19.32
FCN-HA [39]	4.21	-	-	-
CSRNet [18]	3.56	5.49	8.57	15.04
DADNet	**2.79**	**4.41**	**6.43**	**9.27**

proposed DADNet has great robustness and generalization, which applies to both crowd and vehicle counting.

4.4 Density Map Quality

To evaluate the image quality of the predicted density map intuitively, we adopt the PSNR and SSIM metrics to assess our model. The higher the PSNR score is, the higher-quality the predict density map is. The SSIM metric is used to measure the similarity between the predict density map and the groundtruth. Table 5 shows that DADNet achieves the highest PSNR and SSIM values (24.16 and 0.81) for crowd counting, respectively. As shown in Figure 4, when we zoom two red boxes, DADNet generates much more accurate head position and clearer density map than CSRNet.

5 CONCLUSIONS

The paper proposes a Dilated-Attention-Deformable ConvNet (DADNet) framework for both crowd and vehicle counting, which generates high-quality multi-scale attention maps. To capture multi-scale contextual cues in the image, the proposed DADNet designs a scale-aware dilated attention mechanism to effectively discover different visual appearances of nearby objects, distant objects and contour profile of the objects in

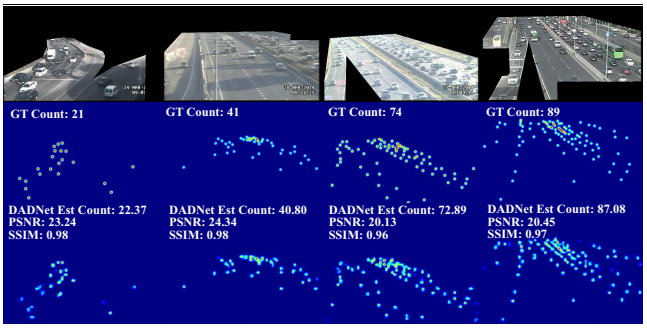


Figure 7: Visualization of density maps of four examples in the TRANCOS dataset for vehicle counting.

Table 5: Qualitative evaluation on the ShanghaiTech Part A dataset.

Model	PSNR↑	SSIM↑
MCNN [40]	21.40	0.52
CP-CNN [31]	21.72	0.72
CSRNet [18]	23.79	0.76
DADNet	**24.16**	**0.81**

crowd or traffic scenes. In addition, a deformable convolution module with adaptive receptive fields is used to further promote the accuracy of object localization, which promises high-quality density maps. Extensive experiments are conducted on the crowd datasets ShanghaiTech, UCF_CC_50, and UCF-QNRF, and the vehicle dataset TRANCOS. Experimental results show that the proposed DADNet achieves superior performance than the state-of-the-art approaches.

ACKNOWLEDGMENTS

This work is supported by the National Natural Science Foundation of China (NSFC) under grants 61725203, 61732008, 61876058, and 61632007.

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 (CVPR)*. 3618–3626.

[2] Ke Chen, Shaogang Gong, Tao Xiang, and Chen Change Loy. 2013. Cumulative attribute space for age and crowd density estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2467–2474.

[3] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. 2018. Deeplab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)* 40, 4 (2018), 834–848.

[4] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. 2017. Deformable convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 764–773.

[5] Navneet Dalal and Bill Triggs. 2005. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition (CVPR)*, Vol. 1. IEEE Computer Society, 886–893.

[6] Piotr Dollar, Christian Wojek, Bernt Schiele, and Pietro Perona. 2012. Pedestrian detection: An evaluation of the state of the art. *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)* 34, 4 (2012), 743–761.

[7] Luca Fiaschi, Ullrich Köthe, Rahul Nair, and Fred A Hamprecht. 2012. Learning to count with regression forest and structured labels. In *Proceedings of the 21st International Conference on Pattern Recognition (ICPR)*. IEEE, 2685–2688.

[8] Ricardo Guerrero-Gómez-Olmedo, Beatriz Torre-Jiménez, Roberto López-Sastre, Saturnino Maldonado-Bascón, and Daniel Onoro-Rubio. 2015. Extremely overlapping vehicle counting. In *Iberian Conference on Pattern Recognition and Image Analysis*. Springer, 423–431.

[9] 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 (CVPR)*. 770–778.

[10] Mohammad Hossain, Mehrdad Hosseinzadeh, Omit Chanda, and Yang Wang. 2019. Crowd Counting Using Scale-Aware Attention Networks. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 1280–1288.

[11] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. 2017. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 4700–4708.

[12] Haroon Idrees, Imran Saleemi, Cody Seibert, and Mubarak Shah. 2013. Multi-source multi-scale counting in extremely dense crowd images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2547–2554.

[13] Haroon Idrees, Muhmmad 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 (ECCV)*. 532–546.

[14] Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. *International Conference on Learning Representations (ICLR)* (2015).

[15] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems (NIPS)*. 1097–1105.

[16] Victor Lempitsky and Andrew Zisserman. 2010. Learning to count objects in images. In *Advances in Neural Information Processing Systems (NIPS)*. 1324–1332.

[17] Xia Li, Jianlong Wu, Zhouchen Lin, Hong Liu, and Hongbin Zha. 2018. Recurrent squeeze-and-excitation context aggregation net for single image deraining. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 254–269.

[18] 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 (CVPR)*. 1091–1100.

[19] Jiang Liu, Chenqiang Gao, Deyu Meng, and Alexander G Hauptmann. 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 (CVPR)*. 5197–5206.

[20] Xialei Liu, Joost van de Weijer, and Andrew D Bagdanov. 2018. Leveraging unlabeled data for crowd counting by learning to rank. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 7661–7669.

[21] David G Lowe. 1999. Object recognition from local scale-invariant features. In *Proceedings of the Seventh IEEE International Conference on Computer Vision (ICCV)*, Vol. 99. 1150–1157.

[22] 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 (ECCV)*. Springer, 615–629.

[23] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in PyTorch. In *NIPS Autodiff Workshop*.

[24] 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 (ICCV)*. 3253–3261.

[25] Viresh Ranjan, Hieu Le, and Minh Hoai. 2018. Iterative crowd counting. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 270–285.

[26] 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 (CVPR)*. IEEE, 4031–4039.

[27] Zan Shen, Yi Xu, Bingbing Ni, Minsi Wang, Jianguo Hu, and Xiaokang Yang. 2018. Crowd Counting via Adversarial Cross-Scale Consistency Pursuit. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 5245–5254.

[28] 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 (CVPR)*. 5382–5390.

[29] K. Simonyan and A. Zisserman. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *International Conference on Learning Representations (ICLR)*.

[30] 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 *2017 14th IEEE International Conference on Advanced Video and Signal-Based Surveillance (AVSS)*. IEEE, 1–6.

[31] 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 (ICCV)*. 1861–1870.

[32] Hongmei Song, Wenguan Wang, Sanyuan Zhao, Jianbing Shen, and Kin-Man Lam. 2018. Pyramid dilated deeper convlstm for video salient object detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 715–731.

[33] Venkatesh Bala Subburaman, Adrien Descamps, and Cyril Carincotte. 2012. Counting people in the crowd using a generic head detector. In *2012 IEEE Ninth International Conference on Advanced Video and Signal-Based Surveillance (AVSS)*. IEEE, 470–475.

[34] Paul Viola and Michael J Jones. 2004. Robust real-time face detection. *International Journal of Computer Vision (IJCV)* 57, 2 (2004), 137–154.

[35] Chuan Wang, Hua Zhang, Liang Yang, Si Liu, and Xiaochun Cao. 2015. Deep people counting in extremely dense crowds. In *Proceedings of the 23rd ACM International Conference on Multimedia*. ACM, 1299–1302.

[36] Zhou Wang, Alan C Bovik, Hamid R Sheikh, Eero P Simoncelli, et al. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing (TIP)* 13, 4 (2004), 600–612.

[37] Fisher Yu and Vladlen Koltun. 2016. Multi-Scale Context Aggregation by Dilated Convolutions. In *International Conference on Learning Representations (ICLR)*.

[38] 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 (CVPR)*. 833–841.

[39] Shanghang Zhang, Guanhang Wu, Joao P Costeira, and José MF Moura. 2017. Fcn-rlstm: Deep spatio-temporal neural networks for vehicle counting in city cameras. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 3667–3676.

[40] 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 (CVPR)*. 589–597.

[41] Xizhou Zhu, Han Hu, Stephen Lin, and Jifeng Dai. 2019. Deformable convnets v2: More deformable, better results. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 9308–9316.