Region-Aware Network: Model Human's Top-Down Visual Perception Mechanism for Crowd Counting[★]

Yuehai Chen[a,1], Jing Yang[a,b,1], Dong Zhang[a], Kun Zhang[a], Badong Chen[b] and Shaoyi Du[b,*]

[a]School of Automation Science and Engineering,Faculty of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an, 710049, Shanxi, China

[b]Institute of Artical Intelligence and Robotics, College of Artical Intelligence, Xi'an Jiaotong University, Xi'an, 710049, Shanxi, China

ARTICLE INFO

ABSTRACT

Background noise and scale variation are common problems that have been long recognized in crowd counting. Humans glance at a crowd image and instantly know the approximate number of human and where they are through attention the crowd regions and the congestion degree of crowd regions with a global receptive filed. Hence, in this paper, we propose a novel feedback network with Region-Aware block called RANet by modeling human's Top-Down visual perception mechanism. Firstly, we introduce a feedback architecture to generate priority maps that provide prior about candidate crowd regions in input images. The prior enables the RANet pay more attention to crowd regions. Then we design Region-Aware block that could adaptively encode the contextual information into input images through global receptive field. More specifically, we scan the whole input images and its priority maps in the form of column vector to obtain a relevance matrix estimating their similarity. The relevance matrix obtained would be utilized to build global relationships between pixels. Our method outperforms state-of-the-art crowd counting methods on several public datasets.

1. Introduction

Crowd counting is an essential work in the field of computer vision. It has wide range of applications such as video surveillance, urban planning, public safety et al. For example, with the rapid growth of population and urbanization, the situations of crowd gathering such as stadium, concerts and parades are more and more frequent. In these scenarios, crowd counting plays an indispensable role for public safety (Ali, Zhu & Zakarya, 2022).

Although crowd counting task is important and useful, the real usage remains limited since the dense crowd counting is challenging. One of the main challenges is background noise. The influence of complex and irrelevant background is not conducive to correctly identify the crowd. Scale variation may also hurt the counting performance. Since the scale of people varies dramatically in images and across different images, it is hard to extract effective features for density regression.

To alleviate noises caused by cluttered backgrounds, attention mechanism is usually introduced to focus on crowd regions (Liu, Long, Zou, Niu, Pan & Wu, 2019a; Rong & Li, 2021; Jiang, Zhang, Xu, Zhang, Lv, Zhou, Yang & Pang, 2020; Hossain, Hosseinzadeh, Chanda & Wang; Gao, Wang & Yuan, 2019b). For instance, (Liu et al., 2019a) proposed an attention-injective deformable convolutional network to generate crowd regions and congestion priors . With similar idea, (Rong & Li, 2021) presented a novel attention network by incorporating attention maps to better focus on the crowd area .

Although these attention-based methods have shown great success in dealing with issue about background noise, they ignore the context information accounting for scale variation in congested scenes by using pixel-wise product of image and corresponding attention map.

Meanwhile, as the heart of these CNNs-based approaches, standard convolutions always exploit the same filters and pooling operations over the whole image to obtain multiscale features. This operation leads to the situation that the actual size of receptive fields in the networks is much smaller than the theoretical size (Onoro-Rubio & López-Sastre, 2016a). It means that normal CNN networks may have limited effect for rapid scale changes in complex scenes as they may assign a fixed scale for large objects (Liu, Salzmann & Fua, 2019b).

People's Top-Down visual perception mechanism can solve these above problems well. As Fig. 1 shows, when human observer crowd scenes, they quickly scan the whole crowd scenes and focus on the crowd regions based on the prior knowledge. It is conducive to reduce the interference of background noise. Then human estimate the approximate number in crowd scenes through the congestion degree of crowd regions and global receptive fields accounting for scale variation.

Inspired by the human's Top-Down visual perception mechanism, we propose a novel method to address the issues above for precise crowd counting in this paper. First, we introduce a feedback architecture to generate priority map focusing on crowd regions for reducing interference caused by the background noise. Then, we design a Region-Aware block to adaptively encoder context information for understanding scale variation through global receptive field.

Fig. 1: Human's Top-Down visual system: According to the goal of the current task and previous prior knowledge, human scan a crowd image and instantly know the approximate number of human and where they are through attention the crowd regions and the congestion degree of crowd regions with a global receptive filed. The color is brighter, the crowd congestion degree is higher. 'GT' means Ground Truth.

In the Region-Aware block, we first do a similarity measurement between the input images and corresponding priority maps through scanning them in the form of column vector. Then, the obtained similarity measurement matrix will be embedded into input images to enhance the crowd regions and build global context relationship accounting for regional consistencies and scale variation.

The main contributions of this paper are as follows:

- We exploit the feature information from priority maps to focus on crowd regions for reducing interference caused by the background noise.

- Through the designed Region-Aware block, the network can encoder the context information and expand the size of receptive field for understanding scale variation.

- With the proposed framework, we achieve state-of-the-art performance on most crowd counting benchmarks.

The remainder of the paper is organized as follows. Section 2 outlines the related works including traditional counting methods, CNNs-based methods and column-based methods. Section 3 describes our proposed models. Section 4 shows the results of our experiments and section 5 concludes this paper.

2. Related work

2.1. Traditional counting methods

Traditional crowd counting algorithms are mainly divided into two categories: detection-based methods and regression-based methods.

Early researchers in crowd counting focus on detection-based methods. (Dollar, Wojek, Schiele & Perona, 2011) used sliding window based detection algorithms to estimate the number of people in images. Also some low-level features such as histogram oriented gradients (HOG), Haar

wavelets and edge were often extracted from human heads or human bodies for human detection (Dalal & Triggs; Viola & Jones, 2004; Wu & Nevatia, 2005). While for partially occluded pedestrians, detection is disappointing.

Hence, regression-based methods are gradually used to solve the problem of crowd counting. Regression-based methods aim to learn the mapping function from low-level features in images such as foreground and texture to the count or density (Ryan, Denman, Fookes & Sridharan, 2009; Chan & Vasconcelos, 2009; Fiaschi, Köthe, Nair & Hamprecht, 2012). These regression methods are more efficient than detection methods, however, they do not fully utilized information in images.

2.2. CNN-based counting methods

Recently, CNN-based methods have demonstrated significant improvements over the traditional methods. Different network architectures are designed to handle various challenges such as background noise and scale changes (Yang, Li, Wu, Su, Huang & Sebe, 2020; Wang, Lv, Zhao, Yang & Ruan, 2020c; Zhang, Li, Wang & Yang, 2015; Chen, Papandreou, Kokkinos, Murphy & Yuille, 2017; Sam, Peri, Sundararaman, Kamath & Radhakrishnan, 2020; Rodriguez-Vazquez, Alvarez-Fernandez, Molina & Campoy, 2022).

The crowd density we want to estimate is the number of people per unit area. However, the density maps will be severely affected by background noise, that is, the background terms with similar texture features to congested crowd scenes will be mistaken as heads easily. With attention model succeeded in various computer vision tasks, many researchers attempted to use the attention method to deal with background noise in crowd counting (Rong & Li, 2021; Zhu, Zhao, Lu, Lin, Peng & Yao, 2019; Liu, Gao, Meng & Hauptmann, 2018; Hossain et al.; Gao et al., 2019b; Zhang, Yue, Shen, Zhu, Zhen, Cao & Shao, 2019b; Zhang, Shen, Xiao, Zhu, Zhen, Cao & Shao, 2019a; Jiang et al., 2020; Wang et al., 2020c). Some researchers use refinement-based algorithms to focus the crowd region and improve the quality of density maps (Rong & Li, 2021; Zhu et al., 2019; Gao et al., 2019b). (Zhu et al., 2019) proposed a dual path multiscale fusion network architecture to generate the final high-quality density maps by fusing attention maps and density maps. (Rong & Li, 2021) devised a from-coarse-to-fine progressive attention mechanism to better focus on the crowd area for people count estimation. (Gao et al., 2019b) introduced the Spatial-/Channel-wise Attention Models to alleviate the mistaken estimation for background regions. (Liu et al., 2019a) proposed an attention-injective deformable convolutional network for crowd understanding that can suppress background noise in highly congested noise scenes. Another way to improve the performance of crowd counting is to adopt the idea of classification (Liu et al., 2018; Zhang et al., 2019b,a; Jiang et al., 2020). (Liu et al., 2018) designed an attention module to adaptively estimate the crowd density based on its real density conditions with detection and regression. (Zhang et al., 2019b,a) incorporated non-local

Fig. 2: The architecture of the proposed region-aware network for crowd counting.

attention mechanism to conquer huge scale variations. (Jiang et al., 2020) provided different attention masks related to regions of different density levels aiming to attenuate the estimation errors in different regions. These attention methods could effectively reduce background interference.However, these attention methods use attention map as mask to do pixel-wise product, which may ignore the relationship between pixels.

Scale variation is also a problem that has been long recognized. Most works handled the large-scale variations issue using different architectures (Zhang, Zhou, Chen, Gao & Ma, 2016; Babu Sam, Surya & Venkatesh Babu, 2017; Onoro-Rubio & López-Sastre, 2016b). (Zhang et al., 2016) designed a Multi-Column Convolutional Neural Network (MCNN) architecture to estimate crowd number accurately in a single image from almost any perspective. Based on multi-scale CNN architecture, (Babu Sam et al., 2017) designed the independent CNN regressor with different receptive fields and trained a switch classifier relay the crowd scene patch to the best CNN regressor. In contrast to these methods that propose specific architectures directly addressing scale variations, the recent methods concentrate on incorporating related information like high-level semantic information (Boominathan, Kruthiventi & Babu, 2016; Sindagi & Patel, 2017) and contextual information (Shang, Ai & Bai, 2016; Liu et al., 2019b) respectively into the network. These related information is useful for network to understand the congested scenes. For example, (Boominathan et al., 2016) used a combination of high-level semantic information and the low-level features from deep learning framework to deal with large scale variations for estimating crowd density. (Sindagi & Patel, 2017) incorporated a high-level prior into the density estimation network enabling the network to learn globally relevant discriminative features for lower count error (Sindagi &

Patel, 2017). And (Shang et al., 2016) designed an end-to-end CNN architecture to predict both local and global count by making use of contextual information. (Liu et al., 2019b) proposed an end-to-end trainable deep architecture that can adaptively encodes the scale of the contextual information aiming to accurately predict crowd density. These methods could alleviate the problem of scale variation to a certain extent. However, these methods may not capture sufficient global contextual information as they focus on local regions.

2.3. Column-Based methods

The patch-based and super pixel-based operations are common and consistent with human visual system. This is consistent with common cognition. However, Deep Convolutional Neural Networks (DCNNs) likely achieve an object recognition competence through a set of mechanisms that are distinct from those in humans (Lonnqvist, Clarke & Chakravarthi, 2020). Thus, the patch-based and super pixel-based operations may not be optimal choices for some applications. In the task of place recognition within 3D point cloud, Scan Context (Kim & Kim, 2018), made a column-wise comparison to achieve effective localization for dynamic objects, by considering global information. In similar way, Kim proposed an analogous column scanning method named Scan Context Image, to improve the localization performance in SLAM task (Kim, Park & Kim, 2019). Image transformer flattens the input tensor in raster-scan order, and computes 1D local attention (similar to column-based operation) for generating natural-looking images (Parmar, Vaswani, Uszkoreit, Kaiser, Shazeer, Ku & Tran, 2018). Szeskin proposed a custom column based convolutional neural network, which is used in the classification of light scattering patterns in columns of vertical pixel-wide vectors in OCT slices (Szeskin, Yehuda, Shmueli, Levy & Joskowicz, 2021). These works use a similar idea of column-based operation

and have achieved good results, which shows that column-based operation is reasonable in some works. Thus, we consider designing column-based region-aware block, which is used to adaptively encoder the global context information into features.

3. Proposed method

As discussed above, we aim to deal with the issues of background noise and scale variation. Human's Top-Down visual perception mechanism can well deal with these issues. When human do crowd counting, according to the goal of the current task and prior knowledge, they firstly focus on crowd regions. Meanwhile, people would not trouble in scale variation as they will take the context information into consideration with a global receptive field. Inspired by this, we proposed a feedback structure with Region-Aware (RA) block modeling human's Top-Down visual perception mechanism for crowd region enhancement and context information capturing through global receptive field.

3.1. Model architecture

As shown in Fig. 2, the model architecture contains 4 components, VGG16 backbone, Dual path multi-scale fusion (Thanasutives, Fukui, Numao & Kijsirikul, 2021), Feedback to provide prior and Region-Aware block. Input images are first fed into VGG16 backbone feature map extractor to extract multi-scale features. Then, the features of high-level semantics information are passed through context-aware module (CAN) (Liu et al., 2019b) and atrous spatial pyramid pooling (ASPP) (Chen et al., 2017) to obtain the scale-aware contextual features. CAN module combines features obtained in Conv4_3 using multiple receptive field sizes of average pooling operation. The pooling output scales are 1,2,3,6. ASPP module applies dilated convolution with different rates (1,6,12,18) to features obtained in Conv5_3 for multi-scale information. There is a skip connection between conv5_3 and conv3_3 for embedding high-level semantical information (Thanasutives et al., 2021). Then, the Dual path multi-scale fusion uses concatenate and up-sample to fuse these multi-scale features to generate priority map. The priority map would provide input images prior information that where are the important regions through feedback. Then, input images and corresponding priority map are put into the Region-Aware block together to obtain new inputs which would enhance the crowd regions and contain global context information. The new inputs are passed through the encoder-decoder based deep convolutional neural networks again to generate feature map and corresponding attention map. At last, the feature map and corresponding attention map would be fused to generate the final high-resolution density map.

3.2. Feedback to provide prior

According to goal of current task and prior knowledge, human could focus on region of interest. With similar idea, we would like to boost the crowd regions in input images for reducing background interference. Given a set of N training images $\{Q_i\}_{1 \le i \le N}$, our goal is to train corresponding priority maps $\{A_i\}_{1 \le i \le N}$ to focus on crowd regions in input images. In order to combine more efficient information, we choose the output fusing the multi-scale features which generated from the VGG16 backbone as shown in Fig. 2:

$$f_i = F_{vgg}(Q), i = \{2, 3, 4, 5\} \tag{1}$$

$$A_i = F_{amg}(f_2, f_3, f_4, f_5) \tag{2}$$

where F_{vgg} is the VGG16 backbone that extracts multi-scale features f_i; F_{amg} is the decoder network that fuses multi-scale features by using bilinear interpolation and concatenation.

The priority map could find the boundary between persons and background as Fig. 3(b) shows. Taking the priority map to boost input images will provide prior about candidate crowd region. The prior enables the inputs pay more attention to those crowd regions. Traditional attention map is an image-sized weight map where crowd regions have higher values (Liu et al., 2019a). They often take an element-wise multiplication while ignore the context information in the crowd region. Different from the general attention mechanism, we enhance the input images by learning the similarity between input images and priority maps and embedding the obtained relevance into inputs as Fig. 4 shows. The details will be illustrated in the following Region-Aware Block.

(a) Input image

(b) Priority map

(c) New input

(d) Difference between (a) Input images and (c) New inputs

Fig. 3: Visualization of the outputs of RANet. (a) is input image; (b) is the corresponding priority map; (c) is the new input obtained from the Region-Aware Block; (d) is obtained by subtracting (a) input images from (c) new inputs.

3.3. Region-Aware Block

As the heart of convolution neural networks, standard convolutions always exploit the same filters and pooling

Fig. 4: The proposed region-aware block.

operations over the whole image. This means that standard convolutions can only capture the local spatial correlation (except the ones at top-most layers). Lack of global receptive field is hard to account for large scale variation and capture context information fully.

In order to deal with scale variation, we introduce global context information into the input images by designing the Region-Aware Block. Specifically, as Fig. 4 shows, we scan the whole input images and its priority maps in the form of column vector to estimate their similarity for ontaining relevance matrix. Then we embed the relevance matrix W into the input images to build global relationships between pixels. By doing this, the network can encoder context information for understanding scale variation and expand the size of receptive field.

3.3.1. Similarity Measurement: Global Information Capturing

Similarity Measurement aims to obtain global relationships between pixels in input images by estimating the similarity between each column in the input images and that in its priority maps.

Traditional attention mechanism takes pixel-wise product between the input image and its attention map. However, such a pixel-wise product may cause lack of context information and local receptive field as they focus on a single pixel and ignore the relationship between pixels. This would not account for scale variation well in images.

In order to capture global relationships between pixels, we scan the whole input images and its priority maps in the form of column vector to estimate their similarity. That is, we estimate the similarity between each column vector in input images and that in its priority maps to obtain weight matrix representing relevance between pixels.

More specifically, the sizes of input image Q and its priority map A are both $N \times M$, denoted as:

$$Q = \begin{bmatrix} q_{11} & \cdots & q_{1m} \\ \vdots & \ddots & \vdots \\ q_{n1} & \cdots & q_{nm} \end{bmatrix} \tag{3}$$

$$A = \begin{bmatrix} a_{11} & \cdots & a_{1m} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nm} \end{bmatrix} \tag{4}$$

Then for each column $Q(:,i)$ in Q and $A(:,j)$ in A, we calculate the similarity $s_{i,j}$ between these two columns by inner product:

$$s_{i,j} = \langle Q(:,i), A(:,j) \rangle = \sum_{r=1}^{n} q_{ri} a_{rj} \tag{5}$$

$$S = Q^T \times A = \begin{bmatrix} s_{11} & \cdots & s_{1m} \\ \vdots & \ddots & \vdots \\ s_{m1} & \cdots & s_{mm} \end{bmatrix} \tag{6}$$

Then, we use softmax function to get the relevance matrix W from similarity matrix S.

$$W = soft\max(S) = soft\max(Q^T \times A)$$
$$= \begin{bmatrix} w_{11} & \cdots & w_{1m} \\ \vdots & \ddots & \vdots \\ w_{m1} & \cdots & w_{mm} \end{bmatrix} \tag{7}$$

The value of w_{ij} can be regarded as an index, which represents the relevance between $i-th$ column in input image and $j-th$ column in corresponding priority map. Consequently, the relevance matrix W can provide input image with access to global information. In other word, the relevance matrix W is further embedded into input images to adaptively build relationship between pixels..

3.3.2. Relevance Embedding: Adaptive Recalibration

We proposed a relevance embedding module to exploit relevance matrix obtained from similarity measurement to build global relationships between pixels. That is, we use the similarity between column vectors in input image and priorty map to adaptively recalibrate relationship between pixels.

More specifically, we calculate the output O by inner product between the input image Q and relevance matrix W as follows:

$$O = Q \times W^T = \begin{bmatrix} o_{11} & \cdots & o_{1m} \\ \vdots & \ddots & \vdots \\ o_{n1} & \cdots & o_{nm} \end{bmatrix} \tag{8}$$

$$o_{ij} = \sum_{r=1}^{m} q_{ir} w_{jr} = \sum_{r=1}^{m} q_{ir} \left[soft\max \left(\sum_{l=1}^{n} q_{lj} \cdot k_{lr} \right) \right] \quad (9)$$

Discussion: we can build global relationship between pixels through similarity measurement and relevance embedding. The new input is displayed in Fig. 3(c). Compare Fig. 3(c) and Fig. 3(d), we could find that the new input could enhance the crowd regions.

Afterward, we train the encoder-decoder networks based deep convolutional neural networks again using the output as new input.

3.4. Loss Function

As we introduce the priors and context information into input images, we use the Bayesian loss as the loss function (Ma, Wei, Hong & Gong, 2019). Suppose that $x\left(x_m = m : m = 1, 2, ..., M\right)$ is a random variable that represents the spatial location and $y\left(y_n = n : n = 1, 2, ..., N\right)$ is a random variable that denotes the annotated head points, y_0 is the background pixels. According to bayes' theorem, the posterior probability of x_m obtaining the annotation y_n and background label y_0 can be calculated as:

$$p\left(x_m | y_n\right) = \frac{1}{\sqrt{2\pi}\delta} \exp\left(-\frac{\|x_m - y_n\|_2^2}{2\delta^2} \right) \quad (10)$$

$$p\left(x_m | y_0\right) = \frac{1}{\sqrt{2\pi}\delta} \exp\left(-\frac{\left(d - \|x_m - y_n^m\|_2\right)^2}{2\delta^2} \right) \quad (11)$$

$$p\left(y_n \mid \mathbf{x}_m\right) = \frac{p\left(\mathbf{x}_m \mid y_n\right) p\left(y_n\right)}{\sum_{n=1}^{N} p\left(\mathbf{x}_m \mid y_n\right) p\left(y_n\right) + p\left(\mathbf{x}_m \mid y_0\right) p\left(y_0\right)}$$
$$= \frac{p\left(\mathbf{x}_m \mid y_n\right)}{\sum_{n=1}^{N} p\left(\mathbf{x}_m \mid y_n\right) + p\left(\mathbf{x}_m \mid y_0\right)} \quad (12)$$

where δ is the variance of a 2D Guassian distribution, y_n^m denotes the nearest head point of x_m and d is a parameter that defines the background points by controlling the distance between the head and background points. Due to the different density of scenes in different datasets, the choice of parameter d is also different which will be shown in Table 2.

The last equation is simplified with the assumption $p\left(y_n\right) = p\left(y_0\right) = 1/\left(N + 1\right)$. Then the estimated counts for each person and background are defined as:

$$E\left[c_n\right] = \sum_{m=1}^{M} p\left(y_n \mid \mathbf{x}_m\right) \mathbf{D}^{est}\left(\mathbf{x}_m\right)$$
$$E\left[c_0\right] = \sum_{m=1}^{M} p\left(y_0 \mid \mathbf{x}_m\right) \mathbf{D}^{est}\left(\mathbf{x}_m\right) \quad (13)$$

where $\mathbf{D}^{est}\left(\mathbf{x}_m\right)$ is the estimated density map and $\mathbf{x}_m$ denotes a 2D pixel location.

In this case, we would like the foreground count at each annotation point equals to one and the background count to be zero. Thus, the final loss function is as follows:

$$\mathcal{L}^{\text{Bayes}+} = \sum_{n=1}^{N} \mathcal{F}\left(1 - E\left[c_n\right]\right) + \mathcal{F}\left(0 - E\left[c_0\right]\right) \quad (14)$$

4. Experiment

In this section, we present the experimental details and evaluation results on 4 public challenging datasets: ShanghaiTech (Zhang et al., 2016), UCF_CC_50 (Bansal & Venkatesh, 2015), UCF-QRNF (Idrees, Tayyab, Athrey, Zhang, Al-Maadeed, Rajpoot & Shah, 2018), JHU-CROWD++ (Sindagi, Yasarla & Patel, 2020) and NWPU (Wang, Gao, Lin & Li, 2020b).

4.1. Evaluation Metrics and Implementation Details

We use the following metrics mean absolute error (MAE), mean square error (MSE) and mean normalized absolute error (NAE) to evaluate the performance of our method.

$$MAE = \frac{1}{N} \sum_{i=1}^{N} \left| C_i - C_i^{GT} \right| \quad (15)$$

$$MSE = \sqrt{\frac{1}{N} \sum_{i=1}^{N} \left| C_i - C_i^{GT} \right|^2} \quad (16)$$

$$NAE = \frac{1}{N} \sum_{i=1}^{N} \frac{\left| C_i - C_i^{GT} \right|}{C_i^{GT}} \quad (17)$$

where N is the number of test images, C_i and C_i^{GT} are the estimated count and the ground truth respectively.

We do the image augmentation using random crop. As shown in Table 2, the sizes of the cropped images differ across datasets. Therefore, when training different datasets, we use different learning rates and batch sizes.

4.2. ShanghaiTech dataset

ShanghaiTech (Zhang et al., 2016) crowd counting dataset contains 1198 labeled images with 330165 people annotated totally. The dataset is divided into two parts named Part A and Part B. Part A consists of 482 (300 for train, 182 for test) images with highly congested scenes collected from the internet. The images in Part A are highly dense with crowd counts between 33 to 3139. While Part B contains 716 (400 for train, 316 for test) images taken from busy streets in Shanghai. The images in Part B are less dense with the number of people varying from 9 to 578. Because of limited numbers of training samples, we pre-train our models on UCF-QNRF.

Table 1
Estimation errors on ShanghaiTech dataset

Method	Part A		Part B	
	MAE	MSE	MAE	MSE
MCNN (Zhang et al., 2016)	110.2	173.2	26.4	41.3
Switch-CNN (Babu Sam et al., 2017)	90.4	135.0	21.6	33.4
CSR-Net (Li, Zhang & Chen, 2018)	68.2	115.0	10.6	16.0
SA-Net (Cao, Wang, Zhao & Su, 2018)	67.0	104.2	8.4	13.6
CAN (Liu et al., 2019b)	62.3	100	7.8	12.2
MBTTBF (Sindagi & Patel, 2019)	60.2	94.1	8.0	15.5
ADCcrowdNet (Liu et al., 2019a)	70.9	115.2	7.7	12.9
LSC-CNN (Sam et al., 2020)	66.4	117.0	8.1	12.7
BL (Ma et al., 2019)	62.8	101.8	7.7	12.7
SFCN (Wang, Gao, Lin & Yuan, 2019)	67.0	104.5	8.4	13.6
CG-DRCN-CC-Res101 (Sindagi et al., 2020)	60.2	**94.0**	7.5	12.1
M-SFANet (Thanasutives et al., 2021)	62.49	106.11	**6.38**	**10.22**
OURS	**57.92**	99.23	7.15	11.86

Table 2
training settings for each datasets

Dataset	Learning rate	Batch size	Crop size	d
ShanghaiTech PartA	$1e-6$	8	256×256	0.1
ShanghaiTech PartB	$1e-6$	8	400×400	0.1
UCF_CC_50	$5e-4$	5	512×512	0.15
UCF-QNRF	$5e-4$	5	512×512	0.15
JHU-CROWD	$5e-4$	5	512×512	0.15
NWPU	$5e-4$	5	512×512	0.15

Note: "d" is the parameter that described in the loss function.

We evaluate our model and compare it to other 12 recent methods in Table 1. The results in Table 1 show that our model can achieve better performance with 7.31% MAE and 6.48% MSE improvement compared with base model(M-SFANet) on Part A. It can also be observed that the proposed method is able to achieve comparable performance with the base model (M-SFANet) on Part B. The effect of our model is similar to that of other models in ShanghaiTech dataset. The reason may be that ShanghaiTech dataset contains less people than other popular datasets and scale variation are not so dramatic compared to other challenging datasets.

To further explore the reasons for the different performance of our method on the two datasets, we select representative images from the two datasets and compare the estimated density maps in Fig. 5. Compared to Part A, the scenes in Part B are more simple and monotonous. More specifically, the scenes of Part B are mainly composed of busy streets and consist of less people. Therefore, the issues of background noise and scale variation are not prominent in Part B dataset. This makes our method still performs honorably but looses its edge compared to the others in Part B dataset. On the contrary, in Part A dataset which is collected from the internet including various diverse scenarios, our model performs better. This further illustrates the superiority

and robustness of our approach under a variety of diverse scenarios.

Fig. 5: Visualization of input images from ShanghaiTech and corresponding estimated density maps. The first and third rows are samples images from ShanghaiTech Part B and ShanghaiTech Part A, respectively. The second and fourth rows are the corresponding estimated density maps from our proposed method.

4.3. UCF_CC_50 dataset

UCF_CC_50 (Bansal & Venkatesh, 2015) is an extremely dense crowd dataset including 50 images of different resolutions. The numbers of head annotations range from 94 to 4543 with an average number of 1280. To better evaluate model performance, 5-fold cross-validation is performed following the standard setting in (Bansal & Venkatesh, 2015). Because of limited numbers of training samples, we pre-train our models on UCF-QNRF to speed up convergence.

Table 3
Estimation errors on UCF_CC_50 dataset

Method	MAE	MSE
MCNN (Zhang et al., 2016)	377.6	509.1
Switch-CNN (Babu Sam et al., 2017)	318.1	439.2
CSR-Net (Li et al., 2018)	266.1	397.5
SA-Net (Cao et al., 2018)	258.4	334.9
CAN (Liu et al., 2019b)	107	183
MBTTBF (Sindagi & Patel, 2019)	233.1	300.9
ADCcrowdNet (Liu et al., 2019a)	273.6	362.0
LSC-CNN (Sam et al., 2020)	225.6	302.7
BL (Ma et al., 2019)	229.3	308.2
SFCN (Wang et al., 2019)	258.4	334.9
M-SFANet (Thanasutives et al., 2021)	162.33	276.76
OURS	**155.01**	**219.45**

Table 4
Estimation errors on UCF-QNRF dataset

Method	MAE	MSE
MCNN (Zhang et al., 2016)	277.0	126.0
Switch-CNN (Babu Sam et al., 2017)	228	445
CSR-Net(Li et al., 2018)	120.3	208.5
CAN (Liu et al., 2019b)	212.2	243.7
MBTTBF (Sindagi & Patel, 2019)	97.5	165.2
LSC-CNN (Sam et al., 2020)	120.5	218.2
BL (Ma et al., 2019)	88.7	154.8
SFCN (Wang et al., 2019)	102.0	171.4
CG-DRCN-CC-Res101 (Sindagi et al., 2020)	95.5	164.3
M-SFANet (Thanasutives et al., 2021)	85.6	151.23
OURS	**83.38**	**141.79**

Table 5
Estimation errors on JHU-CROWD++ dataset

Method	MAE	MSE
MCNN (Zhang et al., 2016)	188.9	483.4
CSR-Net(Li et al., 2018)	85.9	309.2
SA-Net (Cao et al., 2018)	91.1	320.4
MBTTBF (Sindagi & Patel, 2019)	81.8	299.1
LSC-CNN (Sam et al., 2020)	112.7	454.4
SFCN (Wang et al., 2019)	82.3	328.0
CG-DRCN-CC-Res101 (Sindagi et al., 2020)	71.0	278.6
OURS	**59.36**	**257.56**

Table 3 shows the results on UCF_CC_50 dataset. The proposed method is compared with other recent works. It can be observed that our model obtains the best performance with 4.51% MAE and 20.7% MSE improvement compared with the second best approach M-SFANet.

On the UCF_CC_50 dataset, we consistently and clearly outperform all other methods. As shown in Fig. 6, images in UCF_CC_50 dataset are mostly extremely dense crowd images. This makes context more informative and our approach state-of-the-art. What's more, compared to MAE, the improvement of MSE is more remarkable. Compared with MAE, MSE assesses the estimated deviation of the overall data. The smaller the MSE, the more accurate our estimation of the number of people in each image. This could further prove the robustness of our method. In other word, our approach can evaluate approximate number close to ground truth in mostly extremely dense crowd scenes. This is consistent with the behavior of humans observing dense scenes, aiming to get the approximate number of people in dense scenes.

Fig. 6: Some representative sample images from UCF_CC_50 datasets. These sample images are all extremely dense crowd images. More precisely, images in this dataset are almost extremely dense scenes.

4.4. UCF-QNRF dataset

UCF-QNRF (Idrees et al., 2018) is a large and challenging dataset due to the extremely congested scenes. The dataset contains 1535 (1201 for train, 334 for test) jpeg images with 1251642 people in them. What's more, it has a wide range of counts, complex environment and image resolutions. We train our model on UCF-QNRF with VGG-16bn pre-trained weights.

We evaluate our model and compare it to other recent works and results in Table 4. The results in Table 4 indicate that our model can achieve better performance with 2.59% MAE and 6.24% MSE improvement compared with the second best approach M-SFANet.

In Fig. 7, we show input images form UCF-QNRF, along with the density maps generated by the M-SFANet and our proposed method. Compared to accurately localizing each person in M-SFANet model, our method pays more attention to the crowd regions. This means that our method could take context information into consideration for more precise crowd counting. This is exactly in line with human's Top-Down visual system: human scan a crowd image and instantly know the approximate number of human through overall perception of congestion degree of crowd regions. What's more, as Fig. 7 shows, our method can obtain a larger received field which is more closely match the distribution of the crowd regions.

4.5. JHU-CROWD++ dataset

JHU-CROWD++ (Sindagi et al., 2020) contains 4372 images containing a total of 1.51 million dot annotations with an average of 346 dots per image and a maximum of 25000 dots. In comparison to most datasets, the JHU-CROWD++ dataset is a large dataset collecting under a variety of diverse scenarios and environment conditions.

We evaluate our model and compare it to other recent works and results in Table 5. The results in Table 5 show

GT Count: 975 — Estimate: 1248.4 — Estimate: 1114.4

GT Count: 923 — Estimate: 641.0 — Estimate: 797.9

GT Count: 2338 — Estimate: 2320.5 — Estimate: 2668.0

(a) Input Image   (b) M-SFANet   (c) Our method

Fig. 7: Input images form UCF-QNRF, along with the density maps generated by the M-SFANet and our proposed method. From left to right: The left column are input images; the middle column are density maps generated by M-SFANet and the right column are density maps generated by our proposed method. Compared to accurately localizing each person in M-SFANet, our method pays more attention to the crowd regions. What's more, our method can obtain a larger received field for crowd regions.

that our model can achieve better performance with 16.39% MAE and 7.55% MSE improvement compared with the second best approach CG-DRCN-CC-Res101.

In Fig. 8, we show input images form JHU-CROWD++, along with the density maps generated by our proposed method. We can find that our proposed method has a good ability to estimate number in different dense scenarios. In other word, through introducing global context information into the input images by designing the Region-Aware Block, our approach can deal with scale variation well.

4.6. NWPU dataset

The NWPU dataset is the largest-scale and most challenging crowd counting dataset publicly available (Wang et al., 2020b). It is a large-scale congested crowd counting dataset that consists of 5,109 images crawled from the Internet, elaborately annotating 2,133,375 instances. The ground truth for test images set are not released and researchers could submit their results online for evaluation.

We evaluate our model and compare it to other recent works and results in Table 6. The results in Table 6 show that our model significantly outperforms the state-of-the-art methods. Notably, on the NWPU test (obtained by submitting to the evaluation server), our model reduces the MAE

Table 6
Comparison with state-of-the-art methods on NWPU validation and test sets.

Method	Validation set		Test set		
	MAE	MSE	MAE	MSE	NAE
MCNN (Zhang et al., 2016)	218.5	700.6	232.5	714.6	1.063
CSRNet(Li et al., 2018)	104.8	433.4	121.3	387.8	0.604
CAN(Liu et al., 2019b)	93.5	489.9	106.3	386.5	0.295
BL(Ma et al., 2019)	93.6	470.3	105.4	454.2	0.203
SFCN(Wang et al., 2019)	95.4	608.3	105.7	424.1	0.254
DM-Count(Wang, Liu, Samaras & Hoai, 2020a)	70.5	**357.6**	88.4	388.6	**0.169**
OURS	**65.3**	432.9	**77.5**	**365.8**	0.228

Fig. 8: Visualization of estimated density maps from JHU-CROWD++. From left to right: the number of human gradually increases.

and MSE by a large margin, from 88.4 to 77.5 in MAE and from 388.6 to 356.8 in MSE.

4.7. Ablation Study

4.7.1. complexity of the model

To evaluate the complexity of our method, we have conducted ablation study on NWPU dataset in Table 7. To exclude interference from other factors, we conducted the experiment on the same experimental environment, and reported the results in the largest online benchmark NWPU (Wang et al., 2020b). As shown in Table 7, our model does not have an advantage in model parameters and inference speed. However, our model has achieved better performance in the crowd counting. Moreover, our model could also achieve real-time crowd counting at a speed of 0.095 seconds per picture. It does not affect the application of our method in reality.

To better evaluate the complexity of our method, we simply employ VGG16 with Feature Pyramid fusion like U-Net as backbone. We add our proposed feedback architecture and Region-Aware block on the VGG16 backbone. We named these two models as VGG16+PFN and VGG16+PFN+RA.

The quantitative results of VGG16+FPN and VGG16+FPN+RA on ShanghaiTech have been reported in Table 8. As shown in Table 8, our VGG16+PFN+RA model is superior to the base model VGG16+PFN. More specifically, on the ShanghaiTech A, our model reduces the MAE and MSE, from 68.01 to 64.84 in MAE and from 109.94 to 103.58 in MSE. On the ShanghaiTech B, our model reduces the MAE and MSE, from 7.48 to 6.30 in MAE and from 12.08 to 10.14 in MSE. Compared to the improvement of the result, it is acceptable that our method has an increase the inference time.

4.7.2. normal CNN networks for scale variation

To verity the statement that normal CNN networks may have limited effects for rapid scale changes in complex scenes as they may assign a fixed scale for large objects, we simply employ two models named VGG16 and VGG16+SFN respectively to conduct experiments on ShanghaiTech dataset and UCF-QNRF dataset. For VGG16, we simply employ the first pretrained 13 layers of VGG16 with batch normalization as encoder header. Then, we put the output of the backbone to a decoder header which consists of three 3×3 convolutional layers with 256, 64 and 32 channels respectively, and 1×1 convolutional layers to get final density map. While for VGG16+FPN, considering that different layers may focus on different abstract level features, we up-sample and cascade these multiple features as Feature Pyramid Networks.

As can be seen from Table 9, the normal network VGG16 is able to achieve a similar performance with VGG16+FPN which fuses multiple features of different layers in ShanghaiTech A and UCF-QNRF datasets. The reason may be that features of low layers contribute less to crowd counting in complex dense scenes. While in ShanghaiTech B which contains less people and scale variation are not prominent, VGG16+FPN with fusing multiple features of different layers performs better than normal VGG16. These results further indicate that normal networks may have limited effects for rapid scale variation in complex scenes.

4.7.3. Comparison of different boost strategies

The work (Lonnqvist et al., 2020) shows that, Deep Convolutional Neural Networks (DCNNs) and human visual system are not necessarily equivalent models in object

Table 7

The parameters, FLOPs, inference speed and results in NWPU (Wang et al., 2020b) of various models.

Method	Backbone	Parameters(M)	FLOPs(G)	Inference speed(s)	MAE In NWPU
MCNN (Zhang et al., 2016)	FS	**0.13**	**7.05**	**0.008**	232.5
PCC-Net(Gao, Wang & Li, 2019a)	FS	0.51	43.87	0.013	167.4
CSR-Net(Li et al., 2018)	VGG16	16.26	108.34	0.038	121.3
CAN(Liu et al., 2019b)	VGG16	18.10	114.83	0.047	106.3
SCAR(Cao et al., 2018)	VGG16	16.29	108.44	0.047	110.0
SFANet(Zhu et al., 2019)	VGG16	15.92	93.27	0.043	–
M-SFANet(Thanasutives et al., 2021)	VGG16	22.88	115.14	0.058	–
SFCN(Wang et al., 2019)	ResNet101	38.60	162.03	0.096	105.7
RANet(ours)	VGG16	22.88	205.97	0.095	**77.5**

Note: The parameters and FLOPs are computes with the input size of 512×512 on a single NVIDIA 3090 GPU. The inference time is the average time of 100 runs on testing 1024×768 sample. "FS" represents that the model is trained From Scratch.

Table 8

The parameters and inference speeds of two models. The parameters are computes with the input size of 512×512 on a single NVIDIA 3090 GPU. The inference time is the average time of 100 runs on testing 1024×768 sample.

Method	Parameters(M)	Inference speed(s)	ShanghaiTechA (MAE/MSE)	ShanghaiTechB (MAE/MSE)
VGG16+PFN	15.86	0.041	68.01/109.94	7.48/12.08
VGG16+PFN+RA (OURS)	15.87	0.076	64.84/103.58	6.30/10.14

Table 9

Quantitative results of VGG16 and VGG16+FPN in ShanghaiTech and UCF-QNRF.

Method	ShanghaiTech A (MAE/MSE)	ShanghaiTech B (MAE/MSE)	UCF-QNRF (MAE/MSE)
VGG16	67.64/109.46	7.70/12.55	89.52/154.60
VGG16+PFN	68.01/109.94	7.48/12.08	88.48/155.84

recognition. DCNNs likely achieve an object recognition competence through a set of mechanisms that are distinct from those in humans (Lonnqvist et al., 2020). As a result, the patch-based and superpixel-based operations which are common and consistent with human visual system, may not be optimal choices for some applications. There are some column-based methods, which are applied in 3D point cloud (Kim & Kim, 2018), localization task (Kim et al., 2019) and classification task (Szeskin et al., 2021) and have achieved good performances. In order to explore which method is suitable for our dense crowd counting task, we compare the column-based operation with patch-based and superpixel-based operations from theoretical analysis and experimental verification.

Firstly, column-based operation is able to boost the important information of the image. If two vectors are similar, then they will get a high score and will be boosted in our column-based method. In dense crowd counting task, the crowd areas occupy most of the content so that the crowd regions would be highlighted after column-based operation. As a result, column-based operation would boost the crowd

regions and would not destroy the semantic information inside the image. Secondly, dividing the dense image into patches for operation may divide the crowd into different areas which may cause the separation of semantic information. Moreover, patch-based operation is somewhat similar to convolution, and the extracted information may be similar to the features obtained by convolution. Finally, superpixel-based operation focus on individual pixels which may be a lack of consideration of context in dense scenes.

We perform ablation studies on ShanghaiTech A dataset in Table 10 to evaluate the effectiveness of each proposed component. In superpixel-based method, we do superpixel-based operation between input images and obtained maps. For patch-based method, we divide the input images and maps into 16 × 16 patches, then we do similarity measurement and relevance embedding for these patches as we introduced in Region-Aware Block. Comparing base model with superpixel-based method, patch-based method and our column-based method, we find that using feedback to enhance inputs could improve the performance of base model. As can be seen from Table 10, our column-based method performs better than common superpixel-based method and patch-based method in dense scenes. This further evaluates the effectiveness of our proposed Region-Aware block.

To further explore the difference of different boost strategies, we visualize the outputs of these three models in Fig. 9. We respectively visualize the superpixel-based method, patch-based method and our column-based method in the first row, second row and third row. Comparing the differences between input image and new input, we find that superpixel-based method and our column-based method

Fig. 9: Visualization of outputs of superpixel-based method, patch-based method and our column-based method. (a) is the corresponding priority maps; (b) is the new inputs obtained from the networks; (c) is obtained by subtracting input image from (b) new inputs; (d) is the predicted density maps.

Table 10
Comparison of different boost strategies. Superpixel-based method mean that we do superpixel-based operation between input images and obtained maps. For patch-based method, we divide the input images and maps into patches, then we do similarity measurement and relevance embedding for these patches as we introduced in Region-Aware Block.

Method	ShanghaiTech A	
	MAE	MSE
Base Model	65.84	110.22
superpixel-based	60.13	100.34
patch-based (16×16)	64.65	106.79
column-based (OURS)	57.92	99.23

could effectively boost the dense crowd regions. This verifies that our column-based method could boost the important information in input images in dense scenes. While for patch-based method, it has limited effect on boosting the crowd regions by observing the difference between input image and new input. The reason may be that the extracted feature of patch-based method may be similar to the features extracted by convolution. Moreover, the obtained new input of column-based method is similar to input image. This further proves that our column-based method would not destroy the semantic information inside the image. As shown in the third column and fourth column in Fig. 9, we observe that our column-based method could extract more reliable detailed edge texture feature from global receptive field and performs better than superpixel-based method in dense scenes which indicates our column-based method could

consider more contextual information than superpixel-based method in dense scenes.

4.7.4. Visualization of the density maps of base model and RANet

We first introduce a novel feedback architecture to generate priority map focusing on crowd regions. Then we scan the images in the form of column vector to obtain global contextual information and boost the crowd regions in dense scenes for counting according to priority maps. The priority map makes our model pay more attention to crowd regions. And the column vector scanning could extract reliable edge and texture information of targets from global receptive field. The combination of priority map and column vector scanning could synergistically boost the crowd regions effectively. The boosted regions are shown in Fig. 10 (d). Therefore, our model could pay more attention to crowd regions so that it is able to relieve the problem of background noise. As shown in the red circles in the first and the second rows of Fig. 10, we could observe that base model is easier to mistake background for people. While our RANet could distinguish people and background. Compared with base model, the results of our RANet have more consistent responds for crowd regions. This shows that our RANet could relieve the problem of background noise.

Moreover, our model could extract reliable edge and texture information of targets from global receptive field with scanning the images in the form of column vector. As shown in Fig. 10 (d), our model could extract the size of each target. As a result, our model has the ability to implicitly encode the size of each target into the feature. Thus, our region-aware block could relieve the problem of

Fig. 10: Visualization of the density maps of base model and RANet.

scale variation. We could also observe that the respond ranges of base model are larger than scales of targets in green circle 2, green circle 4 and green circle 6. Our RANet has different responds ranges for different targets, which reflects the scale variation. Moreover, as shown in green circle 3 and green circle 5, the density map of base model has too large responds ranges to distinguish each people. While our RANet has suitable respond ranges and is able to distinguish each target. These results further indicate that our RANet has suitable responds ranges for different targets, which reflects our method could relieve the problem of scale variation.

5. Conclusion

In this paper, we proposed a novel feedback architecture model with Region-Aware block modeling human's Top-Down visual perception mechanism, name RANet, aiming to deal with background noise and scale variation. Firstly, we introduce a feedback architecture to train priority map that provide prior about candidate crowd region in input images. This prior would be fully utilized in Region-Aware block to reduce background noise and capture global context information. More specifically, we scan the whole input images and its priority maps in the form of column vector to obtain a relevance matrix for measuring their similarity. The relevance matrix obtained would be utilized to build global relationships between pixels. In other word, the Region-Aware block could adaptively encode the contextual information into input images through global receptive field. So that the RANet shows powerful ability to estimate different dense scenes through attention the crowd regions and the congestion degree of crowd regions with a global receptive filed.

References

Ali, A., Zhu, Y., & Zakarya, M. (2022). Exploiting dynamic spatio-temporal graph convolutional neural networks for citywide traffic flows prediction. *Neural networks*, *145*, 233–247.

Babu Sam, D., Surya, S., & Venkatesh Babu, R. (2017). Switching convolutional neural network for crowd counting. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 5744–5752).

Bansal, A., & Venkatesh, K. S. (2015). People counting in high density crowds from still images. arXiv:1507.08445.

Boominathan, L., Kruthiventi, S. S., & Babu, R. V. (2016). Crowdnet: A deep convolutional network for dense crowd counting. In *Proceedings of the 24th ACM international conference on Multimedia* (pp. 640–644).

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

Chan, A. B., & Vasconcelos, N. (2009). Bayesian poisson regression for crowd counting. In *IEEE 12th international conference on computer vision* (pp. 545–551). IEEE.

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., & Yuille, A. L. (2017). Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, *40*, 834–848.

Dalal, N., & Triggs, B. (). Histograms of oriented gradients for human detection. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)* (pp. 886–893). Ieee volume 1.

Dollar, P., Wojek, C., Schiele, B., & Perona, P. (2011). Pedestrian detection: An evaluation of the state of the art. *IEEE transactions on pattern analysis and machine intelligence*, *34*, 743–761.

Fiaschi, L., Köthe, U., Nair, R., & Hamprecht, F. A. (2012). Learning to count with regression forest and structured labels. In *Proceedings of the 21st International Conference on Pattern Recognition* (pp. 2685–2688). IEEE.

Gao, J., Wang, Q., & Li, X. (2019a). Pcc net: Perspective crowd counting via spatial convolutional network. *IEEE Transactions on Circuits and Systems for Video Technology*, *30*, 3486–3498.

Gao, J., Wang, Q., & Yuan, Y. (2019b). Scar: Spatial-/channel-wise attention regression networks for crowd counting. *Neurocomputing*, *363*, 1–8.

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

Idrees, H., Tayyab, M., Athrey, K., Zhang, D., Al-Maadeed, S., Rajpoot, N., & Shah, M. (2018). Composition loss for counting, density map estimation and localization in dense crowds. In *Proceedings of the European Conference on Computer Vision (ECCV)* (pp. 532–546).

Jiang, X., Zhang, L., Xu, M., Zhang, T., Lv, P., Zhou, B., Yang, X., & Pang, Y. (2020). Attention scaling for crowd counting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 4706–4715).

Kim, G., & Kim, A. (2018). Scan context: Egocentric spatial descriptor for place recognition within 3d point cloud map. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (pp. 4802–4809). IEEE.

Kim, G., Park, B., & Kim, A. (2019). 1-day learning, 1-year localization: Long-term lidar localization using scan context image. *IEEE Robotics and Automation Letters*, *4*, 1948–1955.

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

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

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

Liu, W., Salzmann, M., & Fua, P. (2019b). Context-aware crowd counting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 5099–5108).

Lonnqvist, B., Clarke, A. D., & Chakravarthi, R. (2020). Crowding in humans is unlike that in convolutional neural networks. *Neural Networks*, *126*, 262–274.

Ma, Z., Wei, X., Hong, X., & Gong, Y. (2019). Bayesian loss for crowd count estimation with point supervision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 6142–6151).

Onoro-Rubio, D., & López-Sastre, R. J. (2016a). Towards perspective-free object counting with deep learning. In *European conference on computer vision* (pp. 615–629). Springer.

Onoro-Rubio, D., & López-Sastre, R. J. (2016b). Towards perspective-free object counting with deep learning. In *European conference on computer vision* (pp. 615–629). Springer.

Parmar, N., Vaswani, A., Uszkoreit, J., Kaiser, L., Shazeer, N., Ku, A., & Tran, D. (2018). Image transformer. In *International Conference on Machine Learning* (pp. 4055–4064). PMLR.

Rodriguez-Vazquez, J., Alvarez-Fernandez, A., Molina, M., & Campoy, P. (2022). Zenithal isotropic object counting by localization using adversarial training. *Neural Networks*, *145*, 155–163.

Rong, L., & Li, C. (2021). Coarse-and fine-grained attention network with background-aware loss for crowd density map estimation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* (pp. 3675–3684).

Ryan, D., Denman, S., Fookes, C., & Sridharan, S. (2009). Crowd counting using multiple local features. In *Digital Image Computing: Techniques and Applications* (pp. 81–88). IEEE.

Sam, D. B., Peri, S. V., Sundararaman, M. N., Kamath, A., & Radhakrishnan, V. B. (2020). Locate, size and count: Accurately resolving people in dense crowds via detection. *IEEE transactions on pattern analysis and machine intelligence*, *PP*, 1–1.

Shang, C., Ai, H., & Bai, B. (2016). End-to-end crowd counting via joint learning local and global count. In *IEEE International Conference on Image Processing (ICIP)* (pp. 1215–1219). IEEE.

Sindagi, V., Yasarla, R., & Patel, V. M. (2020). Jhu-crowd++: Large-scale crowd counting dataset and a benchmark method. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (pp. 1–1).

Sindagi, V. A., & Patel, V. M. (2017). Cnn-based cascaded multi-task learning of high-level prior and density estimation for crowd counting. In *14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)* (pp. 1–6). IEEE.

Sindagi, V. A., & Patel, V. M. (2019). Multi-level bottom-top and top-bottom feature fusion for crowd counting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 1002–1012).

Szeskin, A., Yehuda, R., Shmueli, O., Levy, J., & Joskowicz, L. (2021). A column-based deep learning method for the detection and quantification of atrophy associated with amd in oct scans. *Medical Image Analysis*, (p. 102130).

Thanasutives, P., Fukui, K.-i., Numao, M., & Kijsirikul, B. (2021). Encoder-decoder based convolutional neural networks with multi-scale-aware modules for crowd counting. In *25th International Conference on Pattern Recognition (ICPR)* (pp. 2382–2389). IEEE.

Viola, P., & Jones, M. J. (2004). Robust real-time face detection. *International journal of computer vision*, *57*, 137–154.

Wang, B., Liu, H., Samaras, D., & Hoai, M. (2020a). Distribution matching for crowd counting. In *Advances in Neural Information Processing Systems*.

Wang, Q., Gao, J., Lin, W., & Li, X. (2020b). Nwpu-crowd: A large-scale benchmark for crowd counting and localization. *IEEE transactions on pattern analysis and machine intelligence*, *43*, 2141–2149.

Wang, Q., Gao, J., Lin, W., & Yuan, Y. (2019). Learning from synthetic data for crowd counting in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 8198–8207).

Wang, X., Lv, R., Zhao, Y., Yang, T., & Ruan, Q. (2020c). Multi-scale context aggregation network with attention-guided for crowd counting. In *15th IEEE International Conference on Signal Processing (ICSP)* (pp. 240–245). IEEE volume 1.

Wu, B., & Nevatia, R. (2005). Detection of multiple, partially occluded humans in a single image by bayesian combination of edgelet part detectors. In *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1* (pp. 90–97). IEEE volume 1.

Yang, Y., Li, G., Wu, Z., Su, L., Huang, Q., & Sebe, N. (2020). Reverse perspective network for perspective-aware object counting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 4374–4383).

Zhang, A., Shen, J., Xiao, Z., Zhu, F., Zhen, X., Cao, X., & Shao, L. (2019a). Relational attention network for crowd counting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 6788–6797).

Zhang, A., Yue, L., Shen, J., Zhu, F., Zhen, X., Cao, X., & Shao, L. (2019b). Attentional neural fields for crowd counting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 5714–5723).

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

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

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

References

Ali, A., Zhu, Y., & Zakarya, M. (2022). Exploiting dynamic spatio-temporal graph convolutional neural networks for citywide traffic flows prediction. *Neural networks*, *145*, 233–247.

Babu Sam, D., Surya, S., & Venkatesh Babu, R. (2017). Switching convolutional neural network for crowd counting. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 5744–5752).

Bansal, A., & Venkatesh, K. S. (2015). People counting in high density crowds from still images. arXiv:1507.08445.

Boominathan, L., Kruthiventi, S. S., & Babu, R. V. (2016). Crowdnet: A deep convolutional network for dense crowd counting. In *Proceedings of the 24th ACM international conference on Multimedia* (pp. 640–644).

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

Chan, A. B., & Vasconcelos, N. (2009). Bayesian poisson regression for crowd counting. In *IEEE 12th international conference on computer vision* (pp. 545–551). IEEE.

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., & Yuille, A. L. (2017). Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, *40*, 834–848.

Dalal, N., & Triggs, B. (). Histograms of oriented gradients for human detection. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)* (pp. 886–893). Ieee volume 1.

Dollar, P., Wojek, C., Schiele, B., & Perona, P. (2011). Pedestrian detection: An evaluation of the state of the art. *IEEE transactions on pattern analysis and machine intelligence*, *34*, 743–761.

Fiaschi, L., Köthe, U., Nair, R., & Hamprecht, F. A. (2012). Learning to count with regression forest and structured labels. In *Proceedings of the 21st International Conference on Pattern Recognition* (pp. 2685–2688). IEEE.

Gao, J., Wang, Q., & Li, X. (2019a). Pcc net: Perspective crowd counting via spatial convolutional network. *IEEE Transactions on Circuits and Systems for Video Technology*, *30*, 3486–3498.

Gao, J., Wang, Q., & Yuan, Y. (2019b). Scar: Spatial-/channel-wise attention regression networks for crowd counting. *Neurocomputing*, *363*, 1–8.

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

Idrees, H., Tayyab, M., Athrey, K., Zhang, D., Al-Maadeed, S., Rajpoot, N., & Shah, M. (2018). Composition loss for counting, density map estimation and localization in dense crowds. In *Proceedings of the European Conference on Computer Vision (ECCV)* (pp. 532–546).

Jiang, X., Zhang, L., Xu, M., Zhang, T., Lv, P., Zhou, B., Yang, X., & Pang, Y. (2020). Attention scaling for crowd counting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 4706–4715).

Kim, G., & Kim, A. (2018). Scan context: Egocentric spatial descriptor for place recognition within 3d point cloud map. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (pp. 4802–4809). IEEE.

Kim, G., Park, B., & Kim, A. (2019). 1-day learning, 1-year localization: Long-term lidar localization using scan context image. *IEEE Robotics and Automation Letters*, *4*, 1948–1955.

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

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

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

Liu, W., Salzmann, M., & Fua, P. (2019b). Context-aware crowd counting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 5099–5108).

Lonnqvist, B., Clarke, A. D., & Chakravarthi, R. (2020). Crowding in humans is unlike that in convolutional neural networks. *Neural Networks*, *126*, 262–274.

Ma, Z., Wei, X., Hong, X., & Gong, Y. (2019). Bayesian loss for crowd count estimation with point supervision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 6142–6151).

Onoro-Rubio, D., & López-Sastre, R. J. (2016a). Towards perspective-free object counting with deep learning. In *European conference on computer vision* (pp. 615–629). Springer.

Onoro-Rubio, D., & López-Sastre, R. J. (2016b). Towards perspective-free object counting with deep learning. In *European conference on computer vision* (pp. 615–629). Springer.

Parmar, N., Vaswani, A., Uszkoreit, J., Kaiser, L., Shazeer, N., Ku, A., & Tran, D. (2018). Image transformer. In *International Conference on Machine Learning* (pp. 4055–4064). PMLR.

Rodriguez-Vazquez, J., Alvarez-Fernandez, A., Molina, M., & Campoy, P. (2022). Zenithal isotropic object counting by localization using adversarial training. *Neural Networks*, *145*, 155–163.

Rong, L., & Li, C. (2021). Coarse-and fine-grained attention network with background-aware loss for crowd density map estimation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* (pp. 3675–3684).

Ryan, D., Denman, S., Fookes, C., & Sridharan, S. (2009). Crowd counting using multiple local features. In *Digital Image Computing: Techniques and Applications* (pp. 81–88). IEEE.

Sam, D. B., Peri, S. V., Sundararaman, M. N., Kamath, A., & Radhakrishnan, V. B. (2020). Locate, size and count: Accurately resolving people in dense crowds via detection. *IEEE transactions on pattern analysis and machine intelligence*, *PP*, 1–1.

Shang, C., Ai, H., & Bai, B. (2016). End-to-end crowd counting via joint learning local and global count. In *IEEE International Conference on Image Processing (ICIP)* (pp. 1215–1219). IEEE.

Sindagi, V., Yasarla, R., & Patel, V. M. (2020). Jhu-crowd++: Large-scale crowd counting dataset and a benchmark method. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (pp. 1–1).

Sindagi, V. A., & Patel, V. M. (2017). Cnn-based cascaded multi-task learning of high-level prior and density estimation for crowd counting. In *14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)* (pp. 1–6). IEEE.

Sindagi, V. A., & Patel, V. M. (2019). Multi-level bottom-top and top-bottom feature fusion for crowd counting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 1002–1012).

Szeskin, A., Yehuda, R., Shmueli, O., Levy, J., & Joskowicz, L. (2021). A column-based deep learning method for the detection and quantification of atrophy associated with amd in oct scans. *Medical Image Analysis*, (p. 102130).

Thanasutives, P., Fukui, K.-i., Numao, M., & Kijsirikul, B. (2021). Encoder-decoder based convolutional neural networks with multi-scale-aware modules for crowd counting. In *25th International Conference on Pattern Recognition (ICPR)* (pp. 2382–2389). IEEE.

Viola, P., & Jones, M. J. (2004). Robust real-time face detection. *International journal of computer vision*, *57*, 137–154.

Wang, B., Liu, H., Samaras, D., & Hoai, M. (2020a). Distribution matching for crowd counting. In *Advances in Neural Information Processing Systems*.

Wang, Q., Gao, J., Lin, W., & Li, X. (2020b). Nwpu-crowd: A large-scale benchmark for crowd counting and localization. *IEEE transactions on pattern analysis and machine intelligence*, *43*, 2141–2149.

Wang, Q., Gao, J., Lin, W., & Yuan, Y. (2019). Learning from synthetic data for crowd counting in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 8198–8207).

Wang, X., Lv, R., Zhao, Y., Yang, T., & Ruan, Q. (2020c). Multi-scale context aggregation network with attention-guided for crowd counting. In *15th IEEE International Conference on Signal Processing (ICSP)* (pp. 240–245). IEEE volume 1.

Wu, B., & Nevatia, R. (2005). Detection of multiple, partially occluded humans in a single image by bayesian combination of edgelet part detectors. In *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1* (pp. 90–97). IEEE volume 1.

Yang, Y., Li, G., Wu, Z., Su, L., Huang, Q., & Sebe, N. (2020). Reverse perspective network for perspective-aware object counting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 4374–4383).

Zhang, A., Shen, J., Xiao, Z., Zhu, F., Zhen, X., Cao, X., & Shao, L. (2019a). Relational attention network for crowd counting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 6788–6797).

Zhang, A., Yue, L., Shen, J., Zhu, F., Zhen, X., Cao, X., & Shao, L. (2019b). Attentional neural fields for crowd counting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 5714–5723).

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

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

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