# An improved deep multiscale crowd counting network with perspective awareness[*]

**ZHUGE Jingchang (诸葛晶昌)**, DING Ningning (丁宁宁), XING Shujian (邢书剑), and YANG Xinyu (杨新宇)**

*College of Electronic Information and Automation, Civil Aviation University of China, Tianjin 300300, China*

(Received 18 November 2020; Revised 4 February 2021)

Crowd counting is a challenging task, which is partly due to the multiscale variation and perspective distortion of crowd images. To solve these problems, an improved deep multiscale crowd counting network with perspective awareness was proposed. This network contains two branches. One branch uses the improved ResNet50 network to extract multiscale features, and the other extracts perspective information using a perspective-aware network formed by fully convolutional networks. The proposed network structure improves the counting accuracy when the crowd scale changes, and reduce the influence of perspective distortion. To accommodate various crowd scenarios, data-driven approaches are used to fine-tune the trained convolutional neural networks (CNN) model of the target scenes. The extensive experiments on three public datasets demonstrate the validity and reliability of the proposed method.

**Document code:** A **Article ID:** 1673-1905(2021)06-0367-6

**DOI** https://doi.org/10.1007/s11801-021-0184-5

With the development of the economy and society, the law of crowd activity has been attracting increasing attention from researchers. Crowd counting provides a quantifiable digital foundation for crowd activity laws and plays an important role in traffic monitoring and risk control in cities. The development of crowd counting mainly revolves around convolutional neural networks (CNN). Given the further improvement of graphics processing unit, the detection method of crowd counting continuously emerges.

Crowd counting methods are divided into three categories: methods based on feature detection, regression, and density map. The methods based on pedestrian feature detection are implemented by designing a pedestrian detector to detect human body contours or head features and calculate the number of individuals. One example of these methods is the traditional human body feature detection based on histogram of oriented gradients[1]. However, the posture of the human body in crowd images is complex and variable. Regression-based approaches learn manual features from the input images and subsequently use a machine learning model to map the relationship between the features and the counts. The regression model is established through methods that directly map the relationship between the crowd image and the count, such as support vector machine[2] and deep learning. However, some effective manual features are difficult to capture， such challenge reduces the accuracy and robustness of regression-based methods.

Methods based on density map consider the crowd counting problem as estimating a continuous density function that integrates feature maps to determine the number of persons. These approaches are the current mainstream method for crowd counting. Compared with the feature detection and regression-based methods, density maps not only provide information about the number of individuals, but also indicate the distribution of pedestrians. This advantage allows the model to effectively fit the original image. The density map of crowd images can be obtained using CNNs.

The problem of the scale change of pedestrian targets is one of the main factors affecting the performance of crowd counting. To solve the problem of multiscale change in crowd images, Zhang et al[3] proposed a multi-column CNN (MCNN) that uses three different columns of branches to determine different crowd densities and extract parallel scale variation information. Sam et al[4] introduced the switching CNN, which uses a classifier to explicitly select one of the three branches for a given input image in accordance with its crowd density level. Cao et al[5] proposed an encoder-decoder network called scale aggregation network (SANet), which uses modules similar to the Inception architecture to extract multiscale features and improves the quality of the density maps through deconvolution. Li et al[6] presented a single-column CNN method based on dilated convolution,

[*] This work has been supported by the Fundamental Research Funds for the Central Universities from the Civil Aviation University of China (Nos.3122019047 and 3122019054).

[**] E-mail: 12315414@qq.com

i.e., congested scene recognition net (CSRNet). The front end uses a VGG-16 network that removes the fully connected layers to extract features, whereas the back end utilizes dilated convolution to analyze highly crowded scenes. In order to reduce the influence of perspective distortion in crowd images, perspective information is used in ground truth or body part mapping to normalize the proportion of pedestrians. Idrees et al[7] used locally consistent proportional prior maps to detect and calculate densely populated people. Arteta et al[8] utilized depth maps to predict the size of field objects and count them. With the deepening of research methods[9,10], the accuracy of population counting has been continuously improved. However, the problems caused by scale changes and perspective distortion have not been effectively solved.

The multi-column CNN structure addresses the problem of scale change within a certain range. However, integrating and utilizing the multiscale features learned by different CNN structures without losing information while reducing the influence of perspective distortion to generate high quality density maps remains a challenge. Therefore, an improved deep multiscale CNN with perspective awareness is proposed in this study.

One branch adopts a network structure similar to that of ResNet50[11] to remap the multiscale features of the images. Four convolutional blocks are used in the middle to extract different scale features. Each convolutional layer is followed by a batch normalization layer and a rectified linear unit (ReLU) activation function. The setting of the residual block allows the input to skip the convolution and be added directly before the final activation function, so that it can extract accurate multiscale visual information and train an effective deep neural network. Another branch introduces the perspective awareness network (PAN) with a fully convolutional network (FCN) structure containing five convolutional layers. After obtaining the perspective map, the local pattern consistency between this and the ground truth map is measured using the DSSIM loss. The Euclidean distance of each pixel between the ground truth map and the density map generated by the network is used as a loss function to optimize the predicted density map while reducing the difficulty of network training and improving the accuracy of crowd counting. Finally, the outputs of the two branches are fused through fully connected layers to obtain the estimated density map.

In this study, an improved deep multiscale crowd counting network with perspective awareness is proposed. This paper also adopts a patch-wise strategy that utilizes an image patch X as the input of the networks. Unlike the traditional neural network-based density estimation methods, the mapping from scene patches to density maps has singly one branch, whereas the multitask learning framework uses another network branch for perspective awareness. The outputs of the two branches are concatenated, and the number of pedestrians is obtained through integration after mapping using the fully connected neural networks.

The architecture of the proposed network is shown in Fig.1. Information transmission in the traditional CNN requires layer-by-layer connection, which means that the information input of any layer in the network is a reflection of the information output from a previous layer. But such information transmission will lead to loss of information more or less. ResNet network solves this problem to some extent[11]. It safeguards the integrity of information by directly transmitting the input information through cross-layer connection. And the entire network only needs to learn the part that differs between input and output, which makes it easier and more simple to learn. As shown in Fig.2, cross-layer connectivity is a direct connection from a lower level of the network to a higher level of the network, introducing additional connections between the original input and output channels. Thus, multiple scales of receptive fields in the network will be generated as a result as multiple connections different in network depths between inputs and outputs play a role. Cross-layer connectivity can be approximately considered as the multilayer CNN model, different only in that cross-layer connectivity enables the extraction of multiscale features through low-level feature multiplexing.

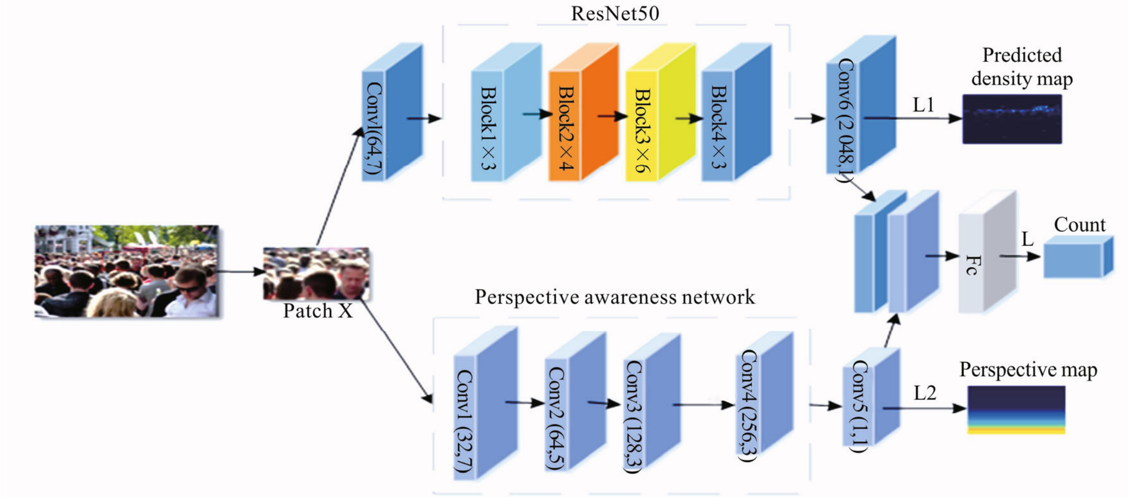

**Fig.1 Architecture of the proposed network**

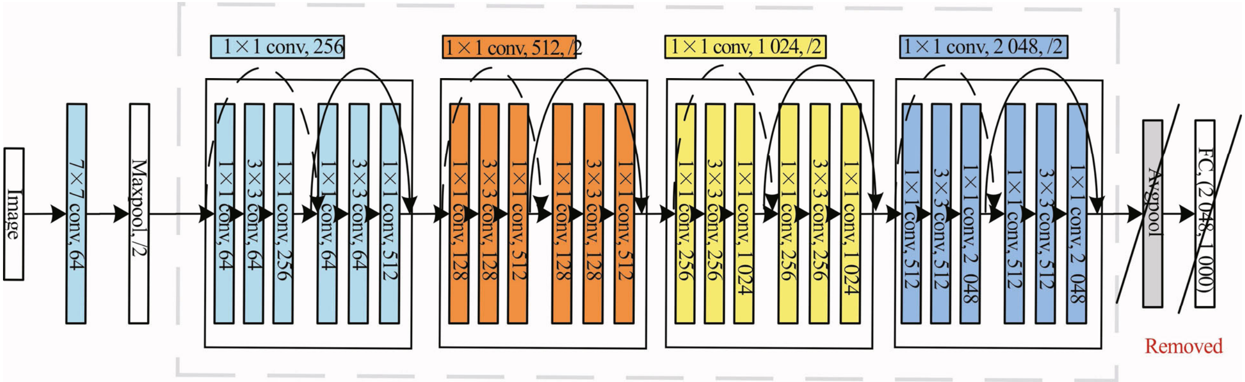

**Fig.2 The improved ResNet50 network**

This study improved the ResNet50 network to introduce beneficial modifications to the original one, which is used for crowd counting. The final average pooling layer and fully connected layer are removed, but all convolutional layers are retained. The input size of the original Resnet50 network is 224×224, and the output size of the final convolution is 7×7, that is, the feature map generated by the last convolutional layer has a spatial resolution of approximately 1/32 of the input image. This is achieved using the CNN with a first convolution kernel of 7×7 (stride of 2), a max pooling layer of 3×3 (stride of 2), and four different convolutional blocks. The backbone of Resnet50 is made up of four multiscale units, each with two additional cross-layer connections, generating three receptive fields different in size. By cascading the multiscale units in Resnet50, the number of receptive fields with different sizes can be greatly increased. To ensure the spatial resolution of the crucial output density maps in crowd counting, an upsampling layer starter module is added from the last of the original ResNet50 network. Therefore, when the input size is $2^n$, the output of the modified network is exactly 1/8 spatial resolution of the input image. The pretrained weights can be directly loaded and used because the number of parameters of the network structure did not change.

Unlike in existing works, the perspective map in the present study is used to extract more salient features for density mapping estimation under the multitask learning framework. The PAN is then added to guide the generation of density maps, which comprises conv1—conv5. Each convolutional layer is followed by a ReLU activation function, which is populated to make it identical to the last layer. The network structure is described as conv1(32,7)-conv2(64,5)-pool3(2)-conv3(128,3)-conv4 (256,3)-conv5(1,1), where ‘conv’ represents a convolution layer, and ‘pool’ represents a max-pooling layer. Numbers in the parentheses are respectively number of channels and kernel size. The regression perspective after conv5(1,1) has exactly 1/16 resolution of the input image. The input image is further upsampled to 1/8 resolution to obtain the final perspective view. The ground truth perspective map is then downsampled to match the size of the predicted density map.

The perspective map plays an important role in the generation of density maps. The ground truth perspective value of $p^g$ each mapped pixel is defined as the number of pixels that represent one meter at that location in the real scene. Therefore, the size of the object observed in the image is related to the perspective value. The traditional method for calculating the value of perspective $p_j^g$ is to interpret the sample perspective values in terms of human height, that is,

$$y_h = \frac{f(C-H)}{z}, \tag{1}$$

where $y_h$ is the observed position of the human head in the image plane, $h$ is the height of the observed person, $z$ is the optical depth of the observed person, and $p^g$ is calculated as

$$p^g = \frac{h}{H} = \frac{1}{C-H} y_h . \tag{2}$$

To generate a perspective view of the crowd image, the general approximate is set to the average height of 1.75 m. Given that $C$ is fixed for each image, $p^g$ becomes a linear function of $y_h$ and remains unchanged in each row. To estimated image $C$, the heights of several pedestrians at different positions in each image are manually marked, linear regression method is then used to fit Eq.(2) to generate the entire ground truth perspective map. The ground truth perspective map generated using this method is displayed in Fig.3.

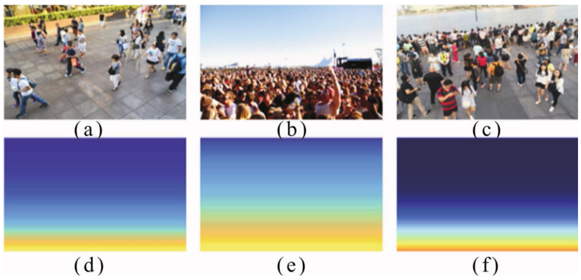

**Fig.3 (a)–(c) Original images and (d)–(f) the corresponding ground truth perspective maps**

Two loss functions are used to optimize the network.

The first is the Euclidean loss function (L1), which is applied to the density map estimation. This function is used to estimate the difference between the ground truth and estimated density maps in the crowd count estimation network module.

$$L(\Theta)=\frac{1}{2N}\sum_{i=1}^{N}\left\|F(X_i;\Theta)-F_i\right\|_2^2 , \tag{3}$$

where $\Theta$ represents the parameter model, $F(X_i;\ \Theta)$ denotes the final output of the ResNet50 network, $X_i$ is the $i$th input image, and $F_i$ is the ground truth.

The second loss function $L^{\text{DSSIM}}$ (L2) is applied to the PAN. The DSSIM loss estimates the local pattern consistency between the perspective and ground truth maps. This loss is derived from the structural similarity (SSIM).

$$L^{\text{DSSIM}}=\frac{1}{N}\sum_{i=1}^{N}(1-\frac{1}{M}\sum_{j}SSIM(j)) , \tag{4}$$

$$SSIM_i=\frac{(2U_{E_i}U_{G_i}+\mathrm{C}_1)}{U_{E_i}^{\ 2}+U_{G_i}^{\ 2}+\mathrm{C}_1}\bullet\frac{2\delta_{E_iG_i}+\mathrm{C}_2}{\delta_{E_i}^{\ 2}+\delta_{G_i}^{\ 2}+\mathrm{C}_2} , \tag{5}$$

where $E$ and $G$ denote the perspective and ground truth maps, respectively. The mean value $(U_{E_i},U_{G_i})$ and standard deviation $(\delta_{E_i},\delta_{G_i},\delta_{E_iG_i})$ are calculated using a Gaussian filter, and constants $\mathrm{C}_1$ and $\mathrm{C}_2$ are included to prevent the denominator from being zero. Therefore, the overall loss for proposed network structure is expressed as follows:

$$L=L_1+\alpha L_2, \tag{6}$$

where the coefficient $\alpha$ is a hyper parameter that balances the relative weight of losses.

Extensive experiments were conducted on published datasets to evaluate the performance of the proposed network. Moreover, a comparative analysis with other existing advanced methods was conducted.

The ShanghaiTech dataset consists of two parts: Part_A and Part_B. The UCF-CC-50 dataset contains 50 images with minimum and maximum counts of 94 and 4 534, respectively. The UCF-QNRF dataset contains 1 535 high-resolution images, among which 1 201 images are used for training and 334 images are used for testing. The details of these datasets are summarized in Tab.1.

**Tab.1 Comparison of dataset statistics**

| Dataset | Number of images | Average count | Number of annotation |
|---|---|---|---|
| ShanghaiTech(A) | 482 | 501 | 241 677 |
| ShanghaiTech (B) | 716 | 123 | 88 488 |
| UCF_CC_50 | 50 | 1 279 | 63 974 |
| UCF-QNRF | 1 535 | 815 | 1 251 642 |

Two metrics are used to evaluate the performance of the different models in the experiments, namely, mean absolute error (*MAE*) and mean square error (*MSE*). *MAE* and *MSE* are used to measure the accuracy and robustness of the model prediction, respectively.

$$MAE=\frac{1}{N}\sum_{i=1}^{N}\left|y_i-\hat{y}_i\right| , \tag{7}$$

$$MSE=\sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i-\hat{y}_i)^2} . \tag{8}$$

The model is implemented on the basis of the Pytorch network framework, the network is trained with the Adam optimizer. The initial learning rate is set to $1\times10^{-5}$, with a decay of 0.1 after every 50 epochs. The total number of training epochs is set to 400 to facilitate an effective model convergence.

Tab.2 shows that the proposed method achieves an *MAE* and *MSE* of 65.3 and 108.4 for the relatively dense Part_A, and the best *MAE* of 8.1 and the second best *MSE* of 13.3 for the more realistic Part_B. The performance is improved compared to other crowd counting algorithms. Fig.4 shows the true and estimated density maps for the test images in the ShanghaiTech dataset. The visualization results are closer to each other as can be seen by comparison, which further demonstrates the effectiveness of the proposed method.

**Tab.2 Comparison on ShanghaiTech dataset**

| Method | ShanghaiTech Part_A | | ShanghaiTech Part_B | |
|---|---|---|---|---|
| | *MAE* | *MSE* | *MAE* | *MSE* |
| MCNN [3] | 110.2 | 173.2 | 26.4 | 41.3 |
| Switching CNN[4] | 90.4 | 135.0 | 21.6 | 33.4 |
| MSCNN[12] | 83.8 | 127.4 | 17.7 | 30.2 |
| CSRNet[6] | 68.2 | 115.0 | 10.6 | 16.0 |
| SANet[5] | 67.0 | 104.5 | 8.4 | 13.6 |
| ESRN[17] | 64.1 | **104.1** | 8.3 | 12.8 |
| DADNet[13] | **63.7** | 107.4 | 9.4 | 15.1 |
| TEDNet[14] | 64.2 | 109.1 | 8.2 | **12.8** |
| **Proposed method** | 65.3 | 108.4 | **8.1** | 13.3 |

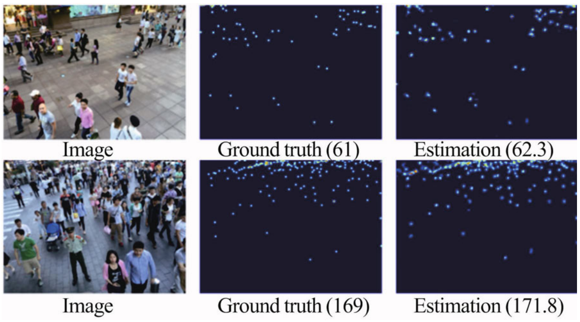

**Fig.4 Results on ShanghaiTech dataset, where the number inside the parenthesis indicates the count**

Compared with other datasets, the UCF-QNRF dataset has a large variation in crowd scale. The experimental results on the UCF-QNRF dataset are presented in Tab.3. The proposed method obtains the best *MAE* and a competitive *MSE*. The *MAE* value of the proposed method decreases by 2.5% compared to TEDNet, which has the advantage of dealing with the scale change problem. Compared with other advanced methods, the generated model can count people in different densities. The details of the findings are depicted in Fig.5.

**Tab.3 Comparison on UCF-QNRF dataset**

| Method | *MAE* | *MSE* |
|---|---|---|
| MCNN[3] | 277.0 | 426.0 |
| Switching CNN[4] | 228.0 | 445.0 |
| PCCNet[15] | 148.7 | 247.3 |
| CSRNet[6] | 122.1 | 192.9 |
| DADNet[13] | 113.2 | 189.4 |
| TEDNet[14] | 113.0 | **188.0** |
| **Proposed method** | **110.1** | 197.8 |

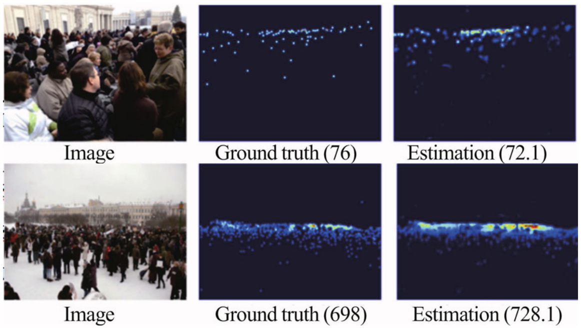

**Fig.5 Results on UCF-QNRF dataset, where the number inside the parenthesis indicates the count**

Tab.4 shows the comparison of the performances of several methods trained on UCF_CC_50 dataset. The proposed method achieves the best MAE (245.3) and MSE (318.8) among the compared approaches. Compared to TEDNet, the ResNet network is simple and the network depth is deep enough. The inclusion of cross-layer connectivity network branches also ensures the integrity of the information. The details of the results are shown in Fig.6.

The initial set of network was trained with 200 epochs, and after gradually increasing to 400 epochs, the network gradually converges. Despite spending a long time training the model, the proposed method demonstrates a better feature extraction performance in scenes with extremely dense crowds. In summary, the above results show that the proposed method not only achieves excellent performance in scenes with large variation of crowd target scales, but also still achieves low counting errors in scenes with extremely dense crowds.

To more effectively verify the effectiveness of the proposed method when the scale changes. 1) The effectiveness of crowd scale variation between different images. 100 crowd images with large scale variation were selected from the above three datasets for validation in this paper. And compared with the published PCCNet, CSRNet and DADNet, the prediction accuracy in the above datasets were 81.7%, 85.1% and 86.1% respectively. But the proposed method has the best prediction accuracy of 86.4%. 2) The effectiveness of scale changes under the same crowd image. An image was randomly selected from the above and the visualization results as shown in Fig.7. After enlarging and shrinking the red box area of image A, images B and C are obtained. The proposed method still obtains more accurate counts, and has good adaptability to different scales of crowd images. In conclusion, the proposed method improves the counting accuracy when the population scale changes.

**Tab.4 Comparison on UCF_CC_50 dataset**

| Method | *MAE* | *MSE* |
|---|---|---|
| MCNN[3] | 377.6 | 509.1 |
| IG-CNN[16] | 291.4 | 349.4 |
| DADNet[13] | 285.5 | 389.7 |
| CSRNet[6] | 266.1 | 397.5 |
| SANet[5] | 258.4 | 334.9 |
| TEDNet[14] | 249.4 | 354.5 |
| **Proposed method** | **245.3** | **318.8** |

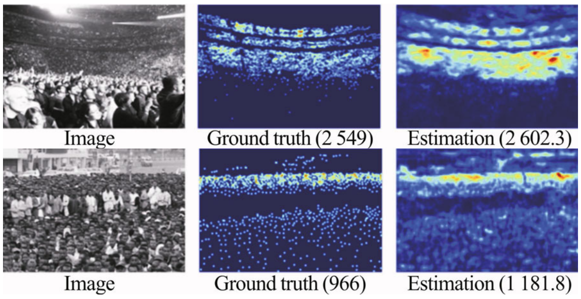

**Fig.6 Results on UCF-CC-50 dataset, where the number inside the parenthesis indicates the count**

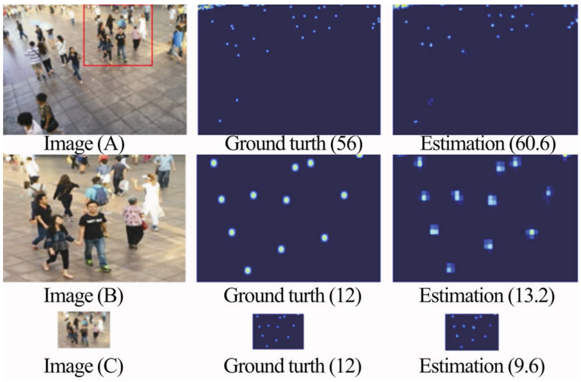

**Fig.7 Results on scale changes, where the number inside the parenthesis indicates the count**

The ResNet50 network exhibits positive results in crowd counting. Tab.5 presents the detailed result of the ablation study. The findings indicate that the addition of PAN improves the counting accuracy. The perspective map strengthens the connection among local pixels while making the details of the generated density map easier to learn for the network. This phenomenon results in a density map with improved quality (Fig.8). The density map generated by the ResNet50+PAN network architecture reflects the crowd distribution better than that generated by ResNet50 alone. The results show that PAN can effectively eliminate perspective distortion and obtain more accurate scene expression capabilities.

**Tab.5 Ablation study on ShanghaiTech dataset**

| Method | Shanghaitech Part_A | | Shanghaitech Part_B | |
|---|---|---|---|---|
| | *MAE* | *MSE* | *MAE* | *MSE* |
| ResNet50 (only) | 67.3 | 112.6 | 10.5 | 15.5 |
| ResNet50+PAN | **65.3** | **108.4** | **8.1** | **13.3** |

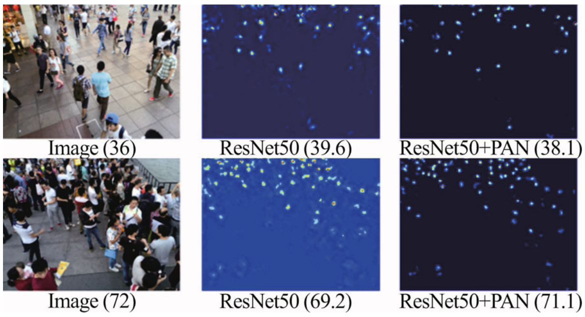

Image (36) ResNet50 (39.6) ResNet50+PAN (38.1)

Image (72) ResNet50 (69.2) ResNet50+PAN (71.1)

**Fig.8 Results on ablation study, where the number inside the parenthesis indicates the count**

In this study, the effectiveness of ResNet50 network in crowd counting is investigated. An improved deep multiscale network structure with perspective awareness is proposed that uses Resnet50 network as the backbone. Using the perspective map to guide the density map generation yields a high-quality density map and improved crowd counting accuracy. The experimental results of three commonly used datasets demonstrate the excellent performance of the proposed network architecture, which can be ascribed to the combination of perspective awareness and the improved ResNet50 network. In the future, the real-time performance of the model will be further improved and actual scenario application may come true.

## References

[1] Dalai N. and Triggs B., Histograms of Oriented Gradients for Human Detection, IEEE Computer Vision and Pattern Recognitio, 886 (2005).

[2] Rabaud V. and Belongie S., Counting Crowded Moving Objects, IEEE Computer Vision and Pattern Recognition, 705 (2006).

[3] Zhang Y., Zhou D., Chen S., Gao S. and Ma Y., Single-Image Crowd Counting via Multi-Column Convolutional Neural Network, IEEE Computer Vision and Pattern Recognition, 589 (2016).

[4] Sam D. B., Surya S. and Babu R.V., Switching Convolutional Neural Network for Crowd Counting, IEEE Computer Vision and Pattern Recognition, 4031 (2017).

[5] Cao X., Wang Z., Zhao Y. and Su F., Scale Aggregation Network for Accurate and Efficient Crowd Counting, European Conference on Computer Vision, 757 (2018).

[6] Li Y., Zhang X. and Chen D., CSRNet: Dilated Convolutional Neural Networks for Understanding the Highly Congested Scenes, IEEE/CVF Conference on Computer Vision and Pattern Recognition, 1091 (2018).

[7] Idrees H., Soomro K. and Shah M., Detecting Humans in Dense Crowds Using Locally-Consistent Scale Prior and Global Occlusion Reasoning, IEEE Transactions on Pattern Analysis and Machine Intelligence, 1986 (2015).

[8] Arteta C., Lempitsky V. and Zisserman A., Counting in the Wild, European Conference on Computer Vision, 483 (2016).

[9] Chen J., Su W. and Wang Z., Neurocomputing **382**, 210 (2020).

[10] Wang Q., Gao J., Lin W. and Li X., NWPU-Crowd: A Large-Scale Benchmark for Crowd Counting and Localization, arXiv:2001.03360, 2020.

[11] He K., Zhang X. and Ren S., Deep Residual Learning for Image Recognition, IEEE Conference on Computer Vision and Pattern Recognition, 770 (2016).

[12] Zeng L., Xu X., Cai B., Qiu S. and Zhang T., Multi-Scale Convolutional Neural Networks for Crowd Counting, IEEE International Conference on Image Processing, 465 (2017).

[13] Guo D., Li K., Zha Z. and Wang M., DadNet: Dilated-attention-Deformable Convnet for Crowd Counting, 27th ACM International Conference on Multimedia, 1823 (2019).

[14] Jiang X., Xiao Z., Zhang B., Zhen X., Cao X., David D. and Shao L., Crowd Counting and Density Estimation by Trellis Encoder-Decoder Networks, IEEE Computer Vision and Pattern Recognition, 6126 (2019).

[15] Gao J., Wang Q. and Li X., PCC Net: Perspective Crowd Counting via Spatial Convolutional Network, arXiv:1905.10085, 2019.

[16] Sam D. B., Sajjan N. N. and Babu R. V., Divide and Grow: Capturing Huge Diversity in Crowd Images with Incrementally Growing CNN, IEEE Computer Vision and Pattern Recognition, 3618 (2018).

[17] Liu C., Duan Y., Du J. and Xu T., IEEE Access **8**, 48352 (2020).