Crowd counting with segmentation attention convolutional neural network

Jiwei Chen1,2 | Zengfu Wang1,2

1 Institute of Intelligent Machines, Hefei Institutes of Physical Science, Chinese Academy of Sciences, Hefei, Anhui, People’s Republic of China
2 Department of Automation, University of Science and Technology of China, Hefei, Anhui People’s Republic of China

Abstract
Deep learning occupies an undisputed dominance in crowd counting. This paper proposes a novel convolutional neural network architecture called SegCrowdNet. Despite the complex background in crowd scenes, the proposed SegCrowdNet still adaptively highlights the human head region and suppresses the non-head region by segmentation. With the guidance of an attention mechanism, the proposed SegCrowdNet pays more attention to the human head region and automatically encodes the highly refined density map. The crowd count can be obtained by integrating the density map. To adapt the variation of crowd counts, SegCrowdNet intelligently classifies the crowd count of each image into several groups. In addition, the multi-scale features are learned and extracted in the proposed SegCrowdNet to overcome the scale variations of the crowd. To verify the effectiveness of this proposed method, extensive experiments are conducted on four challenging datasets. The results demonstrate that the proposed SegCrowdNet achieves excellent performance compared with the state-of-the-art methods.

1 INTRODUCTION

Recently, crowd counting has become a research hotspot owing to its wide applications including public safety management [1], crowd analysis [2] and urban planning [3]. Besides, the methods of crowd counting have important references in vehicle counting, cell counting, and other objects counting [4]. However, due to the crowd’s complexity such as severe occlusion, high diversity, scale variation, view-point variation, and non-uniform distribution, the accuracy of crowd counting still has significant room for improvement.

Lots of efforts have been done in crowd counting. We generally classify these methods into detection-based methods and regression-based methods. The detection-based methods [5–8] obtain the crowd count by counting the number of positive detections. The regression-based methods [6, 9–11] utilise the mapping between extracted features and cropped patches to regress the crowd count. However, the aforementioned methods are complex, because their features need to be extracted by extremely numerous and cumbersome designs. In recent years, deep learning has obtained breakthroughs in many fields [12–14]. Many CNN-based methods [15–17] have been proposed in crowd counting. They learn the non-linear mapping between training images and crowd counts. Researchers employ these CNN models to generate a density map that records the count and spatial information of the crowd at each pixel location, then the density map is integrated to get the crowd count.

Although deep learning has achieved great improvement in crowd counting, we find that the background is extremely complex in the crowd scene, and we believe that the accuracy of crowd counting can be further improved with the segmentation attention mechanism which highlights the foreground, suppresses the background and makes the proposed SegCrowdNet pay more attention to the foreground. Some example images are shown in Figure 1. Let’s carefully distinguish the foreground and background in crowd counting. We obtain the crowd counting result by counting the number of the human head. The human head belongs to the foreground. Everything apart from the human head belongs to the background. Therefore, the background is extremely complicated. In Figure 1, we can also observe that the crowd count changes dramatically...
across images, which is the second challenging problem. To address these two challenging problems, firstly, we predict a segmentation map to adaptively highlight the human head region and suppress the non-head region. With the guidance of the segmentation results, the adaptive attention weights are used in the estimation of the density map to guide our network called SegCrowdNet to pay more attention to the human head region and generate a highly refined density map. Integrating the refined density map can get the crowd counting result. Unfortunately, the present datasets don’t provide the ground truth of segmentation. We propose a simple but effective method in which the ones template is pasted on a binary map to encode the ground truth of segmentation.

2 RELATED WORK

The researches [19–25] about crowd counting are too rich to elaborate all of them. Next, we briefly review some of them.

2.1 Counting by detection

In early researches, many kinds of detection frameworks are proposed in crowd counting. The pivotal features from human body are extracted by the well-trained classifiers such as HOG [5], Random Forest [6], and Haar wavelets [7], then the classifier outputs the positive samples. The total number of positive samples represents the crowd count. [26–28] complete the crowd counting task by using the aforementioned method, and [22] leverages adaptive thresholds to binarize the image to detect the crowd. They can get a good result in the sparse crowd scene. However, when the crowd density becomes high, some persons are too small to be detected.

2.2 Counting by regression

In the highly congested crowd scene, the regression-based methods are usually chosen. They have two important stages. The low-level features are first extracted from the input image, then the crowd count is regressed according to the low-level features. To record the space information of the crowd, Lempitsky et al. [10] proposed a good method that could learn a linear mapping between local area features and corresponding density maps. Based on Lempitsky’s work, Pham et al. [6] proposed a more applicable method. The linear mapping was replaced with a non-linear mapping based on random forest regression. In [24], SVM is proposed to map the features extracted by AdaBoost [29] to the crowd counting result.

2.3 Counting by CNN

In recent years, CNN has achieved great success in the computer vision task including crowd counting. Boominathan et al. [30] combined the deep and shallow CNN to extract large scale features to overcome the scale variation. Zhang et al. [15] proposed a three-branch convolutional neural network to encode the density map with different scales. Based on Zhang’s work, Sam et al. [16] designed a classifier to select the most appropriate branch to generate the density map according to the density variation in each image. Similarly, Hossain et al. [31] proposed a scale-aware method to make each branch in their model focus on a particular scale. Mahmood et al. [23] designed an Encoder–Decoder CNN to learn the head heatmap where experiments are conducted on ShanghaiTech Part_A dataset [15], ShanghaiTech Part_B dataset [15], UCF_CC_50 dataset [11] and WorldExpo’10 dataset [18]. The results demonstrate that our proposed method outperforms many state-of-the-art methods.
human heads can be segmented. Shen et al. [32] introduced the scale-consistency model to alleviate the blurry effects in the estimated density map. Sam et al. [33] designed the TDF-CNN which contains the bottom-up and top-down CNNs to obtain initial accurate predictions. Marsden et al. [4] proposed a CNN model that can count the crowd, wildlife, vehicles, and cells. Wang et al. [34] designed a collector and labeler to generate the crowd data with corresponding labels to reduce the overfitting caused by limited training crowd data. Xu et al. [21] proposed a segmentation branch based on the probability to guide their framework to encode the final density map. Chen et al. [25] proposed a crowd attention network to avoid the enormous misjudgements caused by the scene complexity. This paper proposes a novel segmentation attention mechanism to guide the end-to-end architecture called SegCrowdNet to pay more attention to the human head region. Our segmentation is based on the dice coefficient and we employ the CNN and ground-truth density map to fine tune the final density map. Moreover, our proposed SegCrowdNet can adapt the variation of crowd counts by a classification task.

3 METHODOLOGY

3.1 SegCrowdNet architecture

An overview of the proposed SegCrowdNet can be seen in Figure 2. The backbone is designed to extract multi-scale features. The proposed SegCrowdNet contains the classification task, the segmentation task and the density estimation task.

The configurations of the backbone are shown in Figure 3. To extract multi-scale features to overcome the scale variations of the crowd, firstly, four different receptive fields are inserted into our system at the beginning of this backbone. Each of them has 16 filters. The input image is mapped at different scales by them synchronously, then the results are fed to the following modules. Secondly, inside this module, we design several 2x2 max-pooling layers to extract multi-scale features. Thirdly, we fuse the feature maps with complementary information from different depths of convolutional layers. Every convolutional layer is followed by ReLU. In order to increase the depth of the network to enhance its learning ability without introducing too many parameters, the parameters of the shared module are shared to alleviate overfitting caused by excessive parameters. Since the dilated convolution [35] can increase the receptive field with fewer parameters, the dilated convolution is widely used in our SegCrowdNet.

In Table 1, it can be observed that the crowd count changes greatly across images in the “Range” column. For example, the UCF_CC_50 dataset only contains 50 images. However, the crowd counts range from 94 to 4,543. So the classification task is employed. The classifier can automatically learn the crowd count distribution to adapt the variation of crowd counts. Inspired by [36], the crowd count of each image in the crowd dataset is quantised into several groups. As shown in Fig 2, the fully connected (FC) layers are connected at the end of the backbone. The two fully connected layers followed by PReLU separately have 64 neurons and 5 neurons. The 5 neurons represent five count groups. To avoid the distortion of images and maintain the original distribution of the crowd, we do not resize the
The proposed architecture of the backbone. The convolutional layers’ parameters are represented as “Conv(kernel size)-(number of filters)-(dilation rate)”. The kernel size of the max pool is $2 \times 2$ with stride 2.

**TABLE 1** Statistics of the training datasets: Num represents the number of images; Range represents the range of crowd count in each image; Average represents the average crowd count; Total represents the total number of labelled people

| Dataset            | Resolution         | Color   | Num  | Range    | Average | Total    |
|--------------------|--------------------|---------|------|----------|---------|----------|
| ShanghaiTech Part_A| different          | RGB,Grey| 482  | [33, 3139]| 501.4   | 241,677  |
| ShanghaiTech Part_B| 768 x 1024         | RGB     | 716  | [9, 578] | 123.6   | 88,488   |
| UCF_CC_50          | different          | Grey    | 50   | [94, 4543]| 1279.5  | 63,974   |
| WorldExpo'10       | 576 x 720          | RGB     | 3980 | [1, 253] | 50.2    | 199,923  |

input image. The Spatial Pyramid Pooling (SPP) [37] layers are placed between convolutional layers and FC layers. Arbitrarily sized feature map extracted from the input image can be fed to the SPP layer, and the SPP layer produces fixed size outputs to feed the FC layers. In the classification task, the crowd counts of each dataset are classified into five count groups to adapt the variation of crowd counts. We choose to minimise the cross-entropy loss to optimise this process.

Two other tasks are also shown in Figure 2. In the segmentation task, the segmentation map which adaptively emphasises the head region and suppresses the non-head region is encoded with the supervision of ground-truth segmentation. In the density estimation task, the estimated density map (outlined in blue) and the final estimated density map are predicted with the supervision of ground-truth density map. Integrating the final estimated density map can get the crowd count. As shown in Figure 2, we add the segmentation map to the estimated density map (outlined in blue) and the results are fed to the following convolution layers and ReLU layers to automatically encode the final estimated density map. With the guidance of the segmentation map where the human head regions have higher weights, more attention is paid to the human head region in the density estimation task. In the segmentation task, a novel loss based on the dice coefficient is employed. The novel loss and the ground truth of segmentation will be elaborated in Sec 3.2. In the density estimation task, the Euclidean distance loss is employed to optimise the estimated density map and the final estimated density map.

### 3.2 Model learning

#### 3.2.1 Ground truth generation

The ground-truth density map is extremely important for the supervised methods. For any training image, the 2D point $p$
Model optimisation

The proposed SegCrowdNet is a multi-task model including classification task, segmentation task, and density estimation task. The multi-task learning [38] can serve as a regularisation to alleviate overfitting, which requires our proposed SegCrowdNet to consider every task synchronously, not only one of them. The proposed SegCrowdNet is optimised by minimising four loss functions in a synergistic manner, including an intermediate supervision loss. In the density estimation task, the Euclidean loss is utilised to optimise the estimated density map (outlined in blue) and the final estimated density map,

\[
L_{\text{den}} = \frac{1}{2U} \sum_{i=1}^{U} \| \hat{D}_i - D_i \|_2^2
\]

where \( \hat{D}_i \) represents the estimated density in the intermediate supervision process. \( D_i \) represents the final estimated density. \( D_i \) represents the ground-truth density. \( U \) represents the number of pixels in the ground-truth density map.

In the segmentation task, inspired by [39, 40], we introduce a novel loss in crowd counting which is based on the dice coefficient. The loss is optimised to predict the segmentation map for the human head. The dice coefficient \( D(\hat{s}_i, s_i) \) is between 0 and 1. In the process of optimisation, \( D(\hat{s}_i, s_i) \) is expected to maximise, and the loss \( L_{\text{seg}} \) is expected to minimise. They are formulated as:

\[
D(\hat{s}_i, s_i) = \frac{2\sum_{j} \hat{s}_i \cdot s_j}{\sum_{j} \hat{s}_i^2 + \sum_{j} s_j^2},
\]

\[
L_{\text{seg}} = 1 - D(\hat{s}_i, s_i),
\]

where \( U \) represents the total number of pixels in the ground truth segmentation map. \( s_i \) represents the \( i^{th} \) value in the ground truth segmentation map. \( \hat{s}_i \) represents the \( i^{th} \) value in the predicted segmentation map.

In the classification task, the crowd counts are quantised into five groups. For example, if the crowd counts of a dataset range from 1 to 500 and they are quantised into five classes, the images with populations between 1 and 100 belong to the first class, and the images with populations between 401 and 500 belong to the fifth class. The cross-entropy loss function is utilised,

\[
L_{\text{cla}} = - \frac{1}{M} \sum_{m=1}^{M} \sum_{i=1}^{K} \left( y_i^m \log(\hat{y}_i^m) \right),
\]

where \( M \) represents the total number of training samples. \( K \) represents the total number of classes. \( y_i^m \) represents the
ground-truth class, \( y^* \) represents the output of classification. The final weighted loss is given by Equation (7). Empirically, we set \( \lambda_1 \) as 0.01.

\[
L_{fin} = L_{den} + L_{int} + L_{seg} + \lambda_1 L_{cla}
\]  

(7)

4 | EXPERIMENTS

4.1 | Experiment setting

We conduct extensive experiments on four challenging datasets. The statistics of the four challenging datasets are summarised in Table 1. Some example frames from these four datasets are shown in Figure 6. In the process of creating training data, nine patches are cropped from each original image randomly, and the size of each patch is 1/4 of the original image. Horizontal flipping and noise addition are also used in the training data. We utilise Adam optimisation with a fixed constant momentum of 0.9 to train our model, and we set the learning rate as 1e-6.

4.2 | Evaluation metric

The mean absolute error (MAE) and the mean squared error (MSE) are widely used to measure the crowd counting error on different methods [15, 16, 33, 34]. They are defined as follows:

\[
MAE = \frac{1}{V'} \sum_{i=1}^{V'} |z_i - \hat{z}_i|,
\]  

(8)

\[
MSE = \sqrt{\frac{1}{V'} \sum_{i=1}^{V'} (z_i - \hat{z}_i)^2},
\]  

(9)

where \( z_i \) represents the ground-truth crowd count. \( \hat{z}_i \) represents the estimated crowd count. \( V' \) represents the number of test images.

4.3 | Ablation study using ShanghaiTech part_A

To verify the effectiveness of each task in the proposed SegCrowdNet, we perform ablation studies on ShanghaiTech Part_A [15] which is a large scale and high-density dataset, the dataset contains 482 images with 241,667 annotated persons. In other datasets, similar results can be observed.

To investigate the ability of this classification task where the crowd count of each image is classified, we conduct a set of comparative experiments. In one experiment, the classification task can function normally. In the second experiment, the classification task is removed. The estimation errors are given in Table 3. We can observe that the performance with Cla-task is better, with MAE/MSE 12.4/19.9 lower than that without Cla-task. We think that the proposed SegCrowdNet can learn the crowd count distribution to adapt the variation of crowd counts, which contributes to reducing the crowd counting errors. In the proposed SegCrowdNet, we classify the crowd counts of images in each dataset into five categories according to extensive experiments. From Table 4, we find that our proposed SegCrowdNet with five categories perform best. We think that the crowd counting data are limited and five categories are most suitable. Last but not least, we can further observe that adopting the classification task with different categories in Table 4, the MAE is further alleviated compared with the performance without Cla-task in Table 3, which further indicates that the Cla-task can assist in reducing the estimation error of crowd counting.

The proposed segmentation attention mechanism is the most important idea in this paper. It is included in the segmentation task. To demonstrate the effectiveness of the segmentation task, we also conduct a set of comparative experiments. In one experiment, the segmentation task can function normally. In the second experiment, the segmentation task is removed from our proposed SegCrowdNet and the segmentation attention mechanism is subsequently removed. The qualitative results are illustrated in Table 5. We can observe that the SegCrowdNet greatly reduces the error of crowd counting with the guidance of segmentation results, with MAE/MSE 17.5/24.5 lower than that without Seg-task, which reveals the power of the segmentation attention mechanism. In Figure 5, we demonstrate the estimated crowd count of each image in the comparative experiment. It can be observed that the blue circle (predicted crowd counts with the guidance of the Seg-task) is closer to the red triangle (actual crowd counts) than the green pentagram (predicted crowd counts without Seg-task) in
most images, which indicates that the segmentation attention mechanism plays a vital role in alleviating the errors of crowd counting.

We visualise the effectiveness of the Seg-task in Figure 7. In the third column, it can be observed that the human head regions are highlighted and the non-head regions are suppressed excellently in the estimated segmentation map, which is important to guide our proposed SegCrowdNet to focus on the human head region. In the fourth and fifth columns, we can observe that with the guidance of the segmentation results, the proposed SegCrowdNet pays more attention to the human head region and the crowd counts are predicted beautifully. Similar results can be found in Figure 8. From these results, we can see that every task in the proposed SegCrowdNet is necessary. They work collaboratively to better accomplish the task of crowd counting.

4.4 ShanghaiTech dataset

The ShanghaiTech dataset [15] is highly challenging because of its large-scale crowds. It contains 1,198 annotated images with a total number of 330,165 persons. So far the dataset contains the richest annotated crowd information and it is composed of two parts: Part_A with 300 training images and 182 test images, Part_B with 400 training images and 316 test images. Part_A consists of the high-density crowd images downloaded from the internet. Part_B consists of the relatively low-density crowd images collected from Shanghai streets.

As shown in Table 6, we compare our proposed method with the other fourteen state-of-the-art methods. In the LBP+RR method, the LBP features were extracted by hand to regress the crowd count. Owing to the scale variation in the crowd, the work of MCNN [15], CAFN [41], TDF-CNN [33], Switch-CNN [16], and CP-CNN [17] employed multi-column CNNs to extract the multi-scale features to cope with the scale variation in the crowd. Instead of using the multi-column CNN, in [31], Hossain et al. proposed the scale-aware attention network to encode the multi-scale density maps. In [42], IG-CNN adapted the multi-scale crowd scenes by increasing its capacity. The work of Cascaded-MTL [36], Marsden et al. [4], DecideNet+R3 [43], and DDCN [44] introduced the multi-task learning to assist in reducing the count estimation errors. The work of SE Cycle GAN [34] designed the data collector and labeler to generate the crowd data with the corresponding labels to alleviate the overfitting caused by limited data for training. In [45], the deep negative correlation learning was utilised to extract the general
FIGURE 7 Several visualisation results of the ShanghaiTech Part_A. Top Column: Some input images from the ShanghaiTech Part_A dataset. Second Column: The ground-truth segmentation. Third Column: The predicted segmentation map. Fourth Column: The ground-truth density map. Fifth Column: The final estimated density map. GT: The crowd count of the ground-truth density map. Est: The crowd count of the estimated density map.

TABLE 6 Estimation errors on the ShanghaiTech dataset

| Method           | Part_A    |         | Part_B    |         |
|------------------|-----------|---------|-----------|---------|
|                  | MAE       | MSE     | MAE       | MSE     |
| LBR + RR         | 303.2     | 371.0   | 59.1      | 81.7    |
| SE Cycle GAN [34]| 123.4     | 193.4   | 19.9      | 28.3    |
| MCNN [15]        | 110.2     | 173.5   | 26.4      | 41.3    |
| Cascaded-MTL [36]| 101.3     | 152.4   | 20.0      | 31.1    |
| CAFN [41]        | 100.8     | 152.3   | 21.5      | 33.4    |
| TDF-CNN [33]     | 97.5      | 145.1   | 20.7      | 32.8    |
| Switch-CNN [16]  | 90.4      | 135.0   | 21.6      | 33.4    |
| Marsden et al. [4]| 85.7     | 131.1   | 17.7      | 28.6    |
| DecideNet+R3 [43]| —1       | —       | 20.8      | 29.4    |
| CP-CNN [17]      | 73.6      | 106.4   | 20.1      | 30.1    |
| D-ConvNet-V1 [45]| 73.5      | 112.3   | 18.7      | 26.0    |
| Hossain et al. [31]| —      | —       | 16.9      | 28.4    |
| IG-CNN [42]      | 72.5      | 118.2   | 13.6      | 21.2    |
| DDCN [44]        | 71.5      | 110.4   | 13.8      | 20.1    |
| OURS             | 68.3      | 104.1   | 12.1      | 19.3    |

"—" represents that the result of this dataset is not reported in that paper.

feature representation. The estimation errors are illustrated in Table 6. It can be observed that, firstly, the performance of deep learning is far better than the hand-crafted features. Secondly, lots of efforts are done to extract the multi-scale features of the crowd, which is very important to reduce the estimation errors to some extent. Finally, our proposed SegCrowdNet achieves the best performance in MAE and MSE on both datasets.

4.5 UCF_CC_50 dataset

The UCF_CC_50 dataset [11] contains 50 crowd images collected from the internet. The number of people in each image ranges from 94 to 4,543. There are only 50 images in this dataset. We follow the standard protocol discussed in [11]. The 5-fold cross-validation is utilised to evaluate the performance on this dataset.

In Table 8, we compare our method with other recent state-of-the-art methods. The hand-crafted features were extracted in [11, 46]. The methods which have been compared on the ShanghaiTech are still compared in this dataset. The detailed comparison results are shown in Table 8. In the same way, it can
be observed that all of the CNN-based methods outperform the methods of traditional feature extraction [11, 46] significantly. The proposed SegCrowdNet leads the best performance in MAE. More specifically, its MAE is 38.0 lower than the second best method. However, the MSE which indicates the robustness of a method is not the best. We think that the root of causing this result lies in the limited data, as there are only 50 images in this dataset.

### 4.6 WorldExpo’10 dataset

The WorldExpo’10 dataset [18] contains 1,132 video sequences that are captured from 108 scenarios and it consists of 3,982 annotated frames with a size of 576×720. The frames with the region of interest (ROI) are divided into a training set with 3380 frames and a test set with 600 frames. The test set consists of 5 five different scenes (S1-S5) and each scene contains 120 frames. For fair comparisons, we followed the methods in [18] to generate the density map.

The results of recent state-of-the-art methods are summarised in Table 7. The MAE is utilised to evaluate these methods. We can observe that the CNN-based methods are still superior to traditional methods. Our proposed SegCrowdNet obtains the best result on the S2, S5, and Average. While it achieves comparable performance on the other three scenes. By reviewing the test images of these three scenes, we find that some people are majorly gathered in these three scenes, which is challenging for our proposed SegCrowdNet to further improve their accuracy.

TABLE 7 Estimation errors on the WorldExpo’10 dataset

| Method       | S1  | S2  | S3  | S4  | S5  | Average |
|--------------|-----|-----|-----|-----|-----|---------|
| LBP+RR       | 13.6| 58.9| 37.1| 21.8| 23.4| 31.9    |
| SE Cycle GAN [34] | 4.3 | 59.1| 43.7| 17.0| 7.6 | 26.3    |
| MCNN [15]    | 3.4 | 20.6| 12.9| 13.0| 8.1 | 11.6    |
| TDF-CNN [33] | 2.7 | 23.4| 10.7| 17.6| 3.3 | 11.5    |
| IG-CNN [42]  | 2.6 | 16.1| 10.2| 20.2| 7.6 | 11.3    |
| DDCN [44]    | 4.8 | 16.2| 12.4| 10.9| 4.9 | 9.8     |
| Switch-CNN [16] | 4.4 | 15.7| 10.0| 11.0| 5.9 | 9.4     |
| DecideNet [43] | 2.0 | 13.1| 8.9 | 17.4| 4.8 | 9.2     |
| D-ConvNet-V1 [45] | 1.9 | 12.1| 20.7| 8.3 | 2.6 | 9.1     |
| CP-CNN [17]  | 2.9 | 14.7| 10.5| 10.4| 5.8 | 8.9     |
| OURS         | 2.3 | 11.9| 11.8| 11.2| 2.4 | 7.9     |
TABLE 8 Estimation errors on the UCF_CC_50 dataset

| Method              | MAE  | MSE  |
|---------------------|------|------|
| Rodriguez et al. [46]| 655.7| 697.8|
| Lempitsky et al. [11]| 419.5| 541.6|
| MCNN [15]           | 377.6| 509.1|
| SE Cycle GAN [34]   | 373.4| 528.8|
| TDF-CNN [53]        | 354.7| 491.4|
| Cascaded-MTL [36]   | 322.8| 397.9|
| Switch-CNN [16]     | 318.1| 439.2|
| CAFN [41]           | 305.3| 429.4|
| CP-CNN [17]         | 295.8| 320.9|
| IG-CNN [42]         | 291.4| 349.4|
| D-ConvNet-V1 [45]   | 288.4| 404.7|
| DDCN [44]           | 286.2| 479.6|
| Hossain et al. [31] | 271.6| 391.0|
| OURS                | 233.6| 352.6|

5 | CONCLUSIONS

An end-to-end architecture named SegCrowdNet is designed in crowd counting. We propose a novel segmentation attention mechanism to guide our SegCrowdNet to pay more attention to the human head regions. The proposed SegCrowdNet can also automatically adapt the variation of crowd counts by optimising a classification task. Moreover, the proposed novel four-loss optimisation improves the generalisation ability of the SegCrowdNet. We verify our method on four popular crowd counting datasets (ShanghaiTech Part A dataset, ShanghaiTech Part B dataset, UCF_CC_50 dataset, and WorldExpo’10 dataset). Extensive experimental results demonstrate that our proposed method outperforms many state-of-the-art methods.

ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China (No.61472393).

REFERENCES

1. Gao, L., et al.: Crowd counting considering network flow constraints in videos. IET Image Proc. 12(1), 11–19 (2017)
2. Ma, T., et al.: Scene invariant crowd counting using multi-scales head detection in video surveillance. IET Image Proc. 12(12), 2258–2263 (2018)
3. Zhan, B., et al.: Crowd analysis: A survey. Machine Vision and Applications 19(5-6), 345–357 (2008)
4. Marshe, M., et al.: People, penguins and petri dishes: adapting object counting models to new visual domains and object types without forgetting. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, 2018, pp. 8070–8079
5. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 1, pp. 886–893. IEEE, Piscataway, NJ (2005)
6. Pham, V.Q., et al.: Count forest: Co-voting uncertain number of targets using random forest for crowd density estimation. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 3253–3261. IEEE, Piscataway, NJ (2015)
7. Viola, P., Jones, M.J.: Robust real-time face detection. Int. J. Comput. Vision 57(2), 137–154 (2004)
8. Benenson, R., et al.: Ten years of pedestrian detection, what have we learned? In: European Conference on Computer Vision, pp. 613–627. Springer, Berlin (2014)
9. Chan, A.B., Vasconcelos, N.: Bayesian poisson regression for crowd counting. In: 2009 IEEE 12th International Conference on Computer Vision, pp. 545–551. IEEE, Piscataway, NJ (2009)
10. Lempitsky, V., Zisserman, A.: Learning to count objects in images. In: Advances in Neural Information Processing Systems, pp. 1324–1332 (2010)
11. Idrees, H., et al.: Multi-source multi-scale counting in extremely dense crowd images. In: 2013 IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, 2013, pp. 2547–2554 (2013)
12. Walach, E., Wolf, L.: Learning to count with cnn boosting. In: European Conference on Computer Vision, pp. 660–676. Springer, Berlin (2016)
13. Long, J., et al.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, 2015, pp. 3431–3440
14. Ren, S., et al.: Faster R-CNN: Towards real-time object detection with region proposal networks. In: Advances in Neural Information Processing Systems, pp. 91–99. Curran Associates, Red Hook, NY (2015)
15. Zhang, Y., et al.: 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. IEEE, Piscataway, NJ (2016)
16. Sam, D.B., et al.: Switching convolutional neural network for crowd counting. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, vol. 1, p. 6. IEEE, Piscataway, NJ (2017)
17. Sindagi, V.A., Patel, V.M.: Generating high-quality crowd density maps using contextual pyramid CNNs. In: 2017 IEEE International Conference on Computer Vision, pp. 1879–1888. IEEE, Piscataway, NJ (2017)
18. Zhang, C., et al.: Cross-scene crowd counting via deep convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 833–841. IEEE, Piscataway, NJ (2015)
19. Sindagi, V.A., Patel, V.M.: A survey of recent advances in cnn-based single image crowd counting and density estimation. Pattern Recognit. Lett. 107, 3–16 (2018)
20. Li, T., et al.: Crowded scene analysis: A survey. IEEE Trans. Circuits Syst. Video Technol. 25(3), 367–386 (2015)
21. Xu, H., et al.: Crowd counting with segmentation map guidance. In: Chinese Control Conference, Guangzhou, China, 2019, pp. 7716–7721
22. Sajid, M., et al.: Crowd counting using adaptive segmentation in a congruence model. In: IEEE International Conference on Signal and Image Processing, Beijing, 2017, pp. 745–749
23. Shaban, M., et al.: Multi-person head segmentation in low resolution crowd scenes using convolutional encoder-decoder framework. In: Chen, L., Ben Amor, B., Ghorbel, F. (eds) Representations, Analysis and Recognition of Shape and Motion from Imaging Data, RFMI 2017. Communications in Computer and Information Science, vol. 842. Springer, Cham (2019)
24. Siva, P., et al.: Scene invariant crowd counting and counting using scale-normalized histogram of moving gradients (hong). arXiv preprint arXiv:160200386 (2016)
25. Chen Jiwei, Su Wen, Wang Zengfu Crowd counting with crowd attention learned? In: European Conference on Computer Vision, pp. 613–627. Springer, Berlin (2014)
26. Li, M., et al.: Estimating the number of people in crowded scenes by mid-based foreground segmentation and head-shoulder detection. In: 2009 19th International Conference on Pattern Recognition, pp. 1–4. IEEE, Piscataway, NJ (2008)
27. Zeng, C., Ma, H.: Robust head-shoulder detection by pca-based multi-level hog-lbp detector for people counting. In: 2010 20th International Conference on Pattern Recognition, pp. 2069–2072. IEEE, Piscataway, NJ (2010)
28. Lin, Z., Davis, L.S.: Shape-based human detection and segmentation via hierarchical part-template matching. IEEE Trans. Pattern Anal. Mach. Intell. 32(6), 604–618 (2010)
29. Freund, Y., Schapire, R.E.: A decision-theoretic generalization of on-line learning and an application to boosting. J. Comput. Syst. Sci. 55, 119–139 (1999)
30. Boominathan, L., et al.: Crowdnet: A deep convolutional network for dense crowd counting. In: Proceedings of the 24th ACM international conference on Multimedia, pp. 640–644. ACM, New York (2016)
31. Hossain, M., et al.: Crowd counting using scale-aware attention networks. In: 2019 IEEE Winter Conference on Applications of Computer Vision, pp. 1280–1288. IEEE, Piscataway, NJ (2019)
32. Shen, Z., et al.: Crowd counting via adversarial cross-scale consistency pursuit. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, 2018, pp. 5245–5254
33. Sam, D.B., Babu, R.V.: Top-down feedback for crowd counting convolutional neural network. In: Thirty-Second AAAI Conference on Artificial Intelligence. AAAI (2018)
34. Wang, Q., et al.: Learning from synthetic data for crowd counting in the wild. In: 2019 IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, 2019, pp. 8190–8199
35. Yu, F., Koltun, V.: Multi-scale context aggregation by dilated convolutions. arXiv preprint arXiv:151107122, (2015)
36. Sindagi, V.A., Patel, V.M.: Cnn-based cascaded multi-task learning of high-level prior and density estimation for crowd counting. In: 2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance, pp. 1–6. IEEE, Piscataway, NJ (2017)
37. He, K., et al.: Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE Trans. Pattern Anal. Mach. Intell. 37(9), 1904–1916 (2015)
38. Collobert, R., Weston, J.: A unified architecture for natural language processing: Deep neural networks with multitask learning. In: Proceedings of the 25th International Conference on Machine Learning, pp. 160–167. ACM, New York (2008)
39. Milletari, F., et al.: V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: 2016 Fourth International Conference on 3D Vision, pp. 565–571. IEEE, Piscataway, NJ (2016)
40. Zettinig, O., et al.: Multimodal image-guided prostate fusion biopsy based on automatic deformable registration. Int. J.Comput. Assist. Radiol. Surg. 10(12), 1997–2007 (2015)
41. Fan, L., et al.: CAFN: The combination of atrous and fractionally strided convolutional neural networks for understanding the densely crowded scenes. In: Chinese Conference on Pattern Recognition and Computer Vision, pp. 297–307. Springer, New York (2018)
42. Babu,Sam, D., et al.: Divide and grow: Capturing huge diversity in crowd images with incrementally growing CNN. In: Proceedings of the 31st IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, 2018, pp. 3618–3626
43. Liu, J., et al.: Decidenet: Counting varying density crowds through attention guided detection and density estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, 2018, pp. 5197–5206
44. Wang, L., et al.: Removing background interference for crowd counting via de-background detail convolutional network. Neurocomputing 332, 360–371 (2019)
45. Shi, Z., et al.: Crowd counting with deep negative correlation learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, 2018, pp. 5382–5390
46. Rodriguez, M., et al.: Density-aware person detection and tracking in crowds. In: 2011 International Conference on Computer Vision, pp. 2423–2430.IEEE, Piscataway, NJ (2011)