# DENet: A Universal Network for Counting Crowd with Varying Densities and Scales

Lei Liu, Jie Jiang, Wenjing Jia, *Member, IEEE*, Saeed Amirgholipour, Yi Wang, Michelle Zeibots, Xiangjian He, *Senior Member, IEEE*

***Abstract*—Counting people or objects with significantly varying scales and densities has attracted much interest from the research community and yet it remains an open problem. In this paper, we propose a simple but efficient and effective network, named DENet, which is composed of two components, *i.e.*, a detection network (DNet) and an encoder-decoder estimation network (ENet). We first run the DNet on the input image to detect and count individuals who can be segmented clearly. Then, the ENet is utilized to estimate the density maps of the remaining areas, typically with low resolution and high densities where individuals cannot be detected. For this purpose, we propose a modified Xception network as the encoder for feature extraction and a combination of dilated convolution and transposed convolution as the decoder. When evaluated on the ShanghaiTech Part A, UCF and WorldExpo'10 datasets, our DENet has achieved lower Mean Absolute Error (MAE) than those of the state-of-the-art methods.**

***Index Terms*—Crowd counting, Density estimation, Detection**

## I. Introduction

Vision-based techniques for accurately counting or estimating the number of people (or objects) in a crowded scene are desirable techniques in many real world applications including visual surveillance, traffic monitoring and crowd analysis. This is true especially in restricted, public places such as train stations, where incidents, traffic delay and even terrible stampedes have been reported due to overcrowding in these places. However, various real-world situations, such as heavy occlusions, cluttered background, size and shape variations of people, and perspective distortion, have posed great challenges for practical solutions capable of handling such situations. Thus, accurate counting in crowded scenes is still an open and popular research problem.

Existing crowd counting approaches can be classified to three types, *i.e.*, detection-based methods, regression-based methods and CNN-based methods. The detection-based methods [1] [2] [3] [4] [5] [6] can only detect large scale people and cannot handle high density crowd images. The regression-based methods [7] [8] [9] [10] [11] [12] have difficulties in preserving the high-frequency variation in the density map. With the development of CNNs, CNN-based methods [13] [14] [15] [16] [17] [18] [19] have improved the performance of crowd counting dramatically, which mainly focus on improving the scale invariance of feature representation.

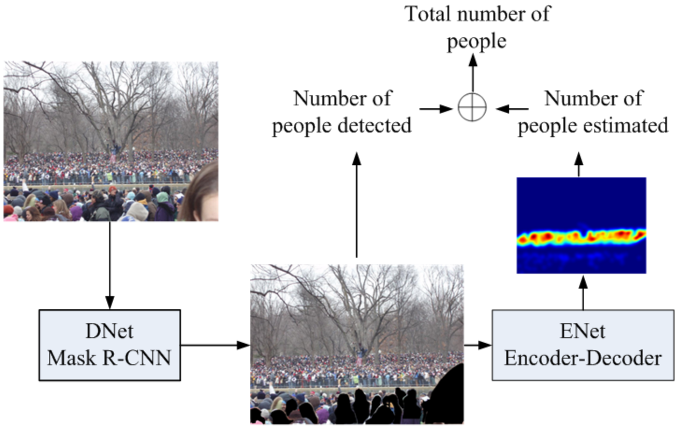

Fig. 1. The architecture of the proposed DENet

We find that most of the CNN based methods concentrate on model design and loss function design. They all try to solve the perspective distortion problem. Can we combine the advantages of the detection based methods and the CNN based methods?

In our work, we propose a simple but an efficient and effective solution. We first run a detection network (the 'DNet') on an input image to detect and count the detected people who can be segmented clearly. These segmented people areas are then removed from the input image. Then, an encoder-decoder estimation network (the 'ENet') is utilized to generate the density map over the crowded areas where individual people cannot be segmented. Furthermore, to combine the advantages of detection and estimation, we propose a novel loss function to train the ENet. The loss is composed of a Euclidean loss and a counting loss. The Euclidean loss is used to generate accurate heat maps and the counting loss is to compare the number of predicted people and the ground truth.

The main contributions of our work are summarized as follows.

- We propose a novel network structure, namely Detection-Estimation Network (simplified as 'DENet') for accurately and efficiently counting crowds of varying densities. The structure improves the multi-scale representation

Lei Liu and Jie Jiang are in School of Instrumentation Science and Opto-Electronics Engineering, Beihang University, China. e-mail: by1417114@buaa.edu.cn; jiangjie@buaa.edu.cn

Xiangjian He, Wenjing Jia and Saeed Amirgholipour are in Global Big Data Technologies Centre, University of Technology Sydney, Australia. email: Xiangjian.He@uts.edu.au; Wenjing.Jia@uts.edu.au; Saeed.AmirgholipourKasmani@student.uts.edu.au; Michelle.E.Zeibots@uts.edu.au

Yi Wang is with the School of Information Science and Engineering, and Key Laboratory for Ubiquitous Network and Service Software of Liaoning Province, Dalian University of Technology. email: dlutwangyi@dlut.edu.cn

Corresponding author: Xiangjian He and Jie Jiang

of the learned network and can produce high-resolution density maps. When applying our DENet structure to some state-of-the-art crowd-counting networks, all estimation accuracies have been improved to some extent, demonstrating the applicability of our core idea.

- To further improve the generality of the estimation network for different scales, we propose a novel estimation network ('ENet'), which uses a modified Xception architecture as the encoder and combines the dilated convolution and transposed convolution as its decoder.
- We propose a new loss function for training the two networks jointly. The function combines a Euclidean loss and a newly proposed counting loss. The Euclidean loss measures the estimation error at the pixel level, and the counting loss measures the counting error of people over the whole image.
- Extensive experiments on several challenging benchmark datasets are conducted to demonstrate the superior performance of our approach over the state-of-the-art solutions.

## II. Related Work

Early works addressing the crowd counting problem mainly follow the counting-by-detection framework [20], which uses body or part-based detectors to detect individual people in crowd images. These methods require well-trained classifiers to extract low-level features (*e.g.*, Haar wavelets [21] and HOG-histogram oriented gradients [5]) from a whole human body. Recent approaches seeking an end-to-end solution using CNN-based object detectors such as YOLO3 [1], SSD [2], Fast R-CNN and Faster R-CNN [3] have greatly improved the detection accuracy. Mask R-CNN [22], proposed by He, can not only detect objects, but also segment them from the background. Although detection-based crowd counting methods are successful for dealing with scenes with low crowd density, when it comes to highly congested environments where only parts of the whole objects are visible, the performance of these detection-based approaches, affected by the size of the targets and occlusions, always degrades significantly. This poses great challenges to object detectors.

The feature-regression-based approaches, as proposed in [13], [23], [14], [7], [24], aim to obtain the density function of an image containing people and then calculate the total count by integrating the density function over the whole image space. More features, such as foreground and texture features, have been used for generating low-level information [25]. Following similar approaches, Idrees *et al.* [10] proposed a model to extract features by employing Fourier analysis and SIFT (Scale-Invariant Feature Transform) [26] interest-point based counting. They have demonstrated a countable solution for handling highly crowded scenes. Recently, CNN-based approaches have shown a remarkable success for crowd counting because of their excellent representation learning ability. Zhang *et al.* [13] designed a multi-column CNN (MCNN) to tackle the large scale variation in crowd scenes. With a similar idea, Onoro and Sastre [7] proposed a scale-aware network, called Hydra, to extract features at different scales. Very recently, inspired by MCNN, Sam *et al.* [14] presented the Switch-CNN, which trained a classifier to select the optimal regressor from multiple independent regressors for specific input patches. Sindagi *et al.* [15] proposed to consider the global and local contextual information by using four modules. They used a combination of adversarial loss and pixel-wise Euclidean loss to improve the accuracy of the density map.

Liu *et al.* proposed DecideNet [27], which adaptively adopted detection and regression based count estimations under the guidance of an attention mechanism. Li *et al.* [16] proposed CSRNet by using VGG-16 [28] to extract feature and dilated convolution layers, to generate the density map. Cao proposed SANet [17] by combining the Euclidean loss and counting loss, and used a set of transposed convolutions to create high-resolution density maps.

The most recent works, *e.g.*, [13] [14] [15], have attempted to address the scale variation issue with multi-scale architectures. They used CNNs with different field sizes to extract features adaptive to the large variation in people sizes, and have achieved significant improvements. Since the high-resolution density maps contain finer details, we believe that it is of great value to develop crowd density estimation techniques that can produce high-resolution and high-quality density maps. However, there exist the following limitations in existing crowd counting works dealing with varying crowdedness. Most of the work has focused on either density estimation or people detection. Although some of the recent works have attempted to develop an adaptive network by combining density estimation and people detection, they are only suitable for low density scenes and are very inflexible. For example, DecideNet [27] introduced a ratio that needs to be retrained each time when the ratio of high-density areas to low-density areas changes. This inflexibility poses practical issues for applications. Some recently published works, *e.g.*, [18] [19] [29] [30], proposed new loss functions. Moreover, in existing CNN-based methods, the input of an estimation network is typically the whole image, where people of different scales are all annotated as dots of the same scale in the heat map. However, it is difficult to learn a network that can produce similar performance with inputs of different scales.

## III. Proposed algorithm

To address the varying scale issues, we follow the two points discussed above and propose a novel encoder-decoder network, named Detection-Estimation Network (DENet), which architecture is shown in Fig. 1. Inspired by the success of Mask R-CNN [22] on object detection, we first adopt Mask R-CNN to detect and segment people who can be clearly differentiated from the crowd, and then we propose a novel estimation network to estimate the density map for the areas where individuals cannot be segmented due to high crowdedness. We modify Xception [31] to be the encoder of the estimation network so as to improve the representation ability and scale diversity of features. The decoder is composed of a set of dilated convolutions [32] and transposed convolutions. It is used to generate high-resolution and high-quality density maps, of which the sizes can be exactly the same as that of the

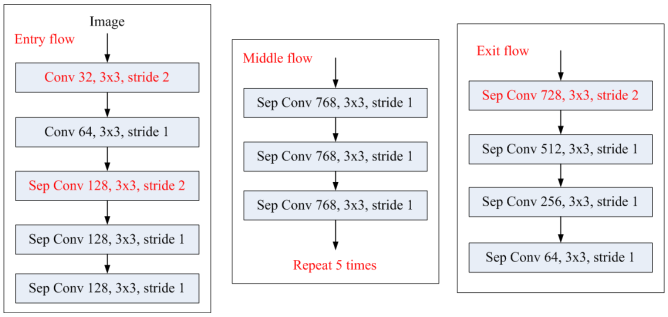
Image
Entry flow
Conv 32, 3x3, stride 2
Conv 64, 3x3, stride 1
Sep Conv 128, 3x3, stride 2
Sep Conv 128, 3x3, stride 1
Sep Conv 128, 3x3, stride 1
Middle flow
Sep Conv 768, 3x3, stride 1
Sep Conv 768, 3x3, stride 1
Sep Conv 768, 3x3, stride 1
Repeat 5 times
Exit flow
Sep Conv 728, 3x3, stride 2
Sep Conv 512, 3x3, stride 1
Sep Conv 256, 3x3, stride 1
Sep Conv 64, 3x3, stride 1

Fig. 2. The modified Xception in our DNet

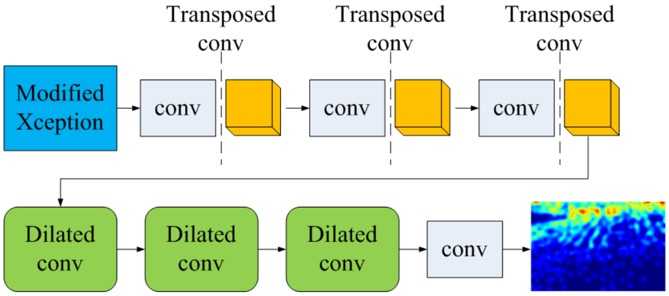

Fig. 3. The architecture of the DME in our DNet

input images. This section presents the details of our proposed DENet. Moreover, we propose a new loss function.

### A. DNet

Mask R-CNN proposed in [22] has further advanced Faster R-CNN [3] by adding a branch for predicting an object mask in parallel with the existing branch for bounding box recognition. Thus, Mask R-CNN can not only detect objects but also segment them from the input image. We adopt Mask R-CNN as our DNet in our work and we only retain large people segmented by DNet, because the smaller the size of the people, the greater the number of miss detections and false detections.

### B. ENet

As shown in Fig. 1, our estimation network consists of two components, *i.e.*, feature map encoder (FME) and density map estimator (DME). We adopt a modified Xception as the FME to extract features, and a set of dilated convolutions and transposed convolutions as DME to create high-resolution and high-quality density maps.

FME: Following the similar idea in [31], we modify Xception to form an FME of the estimation net because it has been widely used as the encoder for feature extraction. Moreover, based on our observation, the performance of the network for our counting task is not sensitive to the number of network parameters and using fewer parameters does not degrade the counting accuracy significantly. Thus, in order to reduce the computation complexity, we have removed four blocks from the original Xception architecture, and have achieved similar accuracy. The architecture of the modified Xception is shown in Fig. 2.

For the DME, while density-estimation-based approaches take the spatial information into account, the outputs of these works are mostly of low-resolution due to several pooling layers, and hence cause the loss of detail compared with the ground truth. Inspired by the approaches of CSRNet [16] and SANet [17], the dilated convolution can keep more details than the traditional convolution, and the transposed convolution can alleviate the loss of information. Therefore, in our work, we deploy a combination of dilated convolutions and transposed convolutions as the decoder for the ENet to create high-resolution and high-quality density maps.

Fig. 3 shows the architecture of the DME of our ENet. Three pairs of convolution layer and transposed convolutional layer are added after FME. The size of the filters in convolution layer ranges from 7 × 7 to 3 × 3, and each transposed convolutional layer doubles the size of the output of the previous layer. A set of dilated convolutional layers are added after the last transposed convolutional layer to keep the details of feature maps. Then, a $1 \times 1$ convolution layer is added after the last dilated convolutional layer to generate density maps. ENet focuses on the features of high density parts in the input image, and hence facilitates the feature learning in the model. The size of density maps generated by ENet are the same as the size of the input image.

### C. Loss function

Since it is unavoidable that the DNet may miss detecting some targets, even with a highly accurate ENet, the total number of the estimated people will not be correct. To solve this problem, in our DENet, in order to learn a model that can achieve more accurate density maps, we propose to integrate the Euclidean loss with a new counting loss. Euclidean loss is used to measure the pixel-level density map error. Counting error is the difference between the ground-truth number of people and the sum of the number of people obtained by detection and the number of people obtained by density estimation. We propose to consider counting loss to measure counting error. Mathematically, this can be illustrated as below.

Let $L_E$ denote the Euclidean loss, which measures the estimation error at pixel level and is defined as:

$$L_E = \frac{1}{N}||F(X) - Y(X)||_2^2, \tag{1}$$

where $N$ is the number of pixels in the density map, $X$ is the input image, $Y(X)$ is the corresponding ground-truth density map, and $F(X)$ denotes the estimated density map. Following the approach proposed by [33], we use $L_C$ to measure the difference between the ground truth number of people and the sum of the numbers of people detected and estimated. $L_C$, which represents the total computing loss, is defined as:

$$L_C = ||\frac{(N_{GT} - N_D - N_E)}{(N_{GT} - N_D + 1)}||^2, \tag{2}$$

where $N_{GT}$ is the ground-truth of the total number of people in an input image, and $N_D$ and $N_E$ are the numbers of people obtained by the detection and estimation, respectively. In case DNet detects all the people, we add 1 in the denominator of this equation to avoid the denominator becoming zero. By weighting the above two loss functions equally, we define the final loss function as:

$$Loss = L_E + L_C, \tag{3}$$

## IV. Implementation Details

In this section, we provide more details of training the proposed DENet, generating ground truth and performance evaluation metrics.

### A. Training DENet

The weights in DNet are from a well-trained Mask R-CNN. For the ENet, all of the convolution layers are initialized with zero. During the training of ENet, Adam is used as the optimizer. During the training stage, we use the whole images to train the network, where each training image is augmented (flipping horizontally and vertically, etc) four times for training.

The implementation of our approach is based on the PyTorch framework [34]. The hardware test-bed is GPU: P5000, CPU: Xeon e5, RAM: 16G. We train every dataset for 400 epochs, the learning rate is 1-e6, and decreases as the iteration increases, batch size is 4. The training time for each dataset is different: Shanghai Tech PartA: 20h; Shanghai Tech PartB: 26h; UCSD: 14h; UCF: 30h; WorldExpo'10: 96h.

### B. Ground truth generation

Ground truth generation is similar to that in existing works [13], [16], [17], where annotations for crowd images are dots at the center of pedestrians' heads, and the ground truth density functions are generated at each of the dots. We generate the ground truth by blurring the dot map with a Gaussian kernel (which is normalized to 1). The density map is defined as:

$$F(x) = \sum_{i=1}^{N} \sigma(x - x_i) * G_{\sigma_i}(x). \tag{4}$$

To generate the density map, the dot map $\sigma(x - x_i)$ is convolved with a Gaussian kernel of a standard deviation $\sigma_i$, where $x_i$ is the position of the pixel at the $i^{th}$ dot in the image. In our experiments, we follow the configuration in [16], where different datasets use Gaussian kernels of different values of $\sigma$.

Note that all the Gaussian functions are summed and normalized, so that the total object count is preserved even when there is overlapping between targets.

### C. Evaluation metrics

In previous work, for crowd density estimation, two metrics have been widely used to measure the counting error, *i.e.*, Mean Absolute Error (MAE) and Mean Squared Error (MSE), defined as:

$$MAE = \frac{1}{N} | \sum_{i=1}^{N} |C_i - C_i^{GT}| \tag{5}$$

TABLE I
Statistics of different datasets

| Dataset | Images | Annotations | Average Count |
|---|---|---|---|
| ShanghaiTech [13] | 1198 | 33,0165 | 501 |
| UCF [10] | 50 | 63,974 | 1279 |
| UCSD [8] | 2000 | 49,885 | 29 |
| World expo [35] | 3980 | 199,923 | 56 |
| UCF-QNRF [36] | 1535 | 1,251,642 | 815 |

and

$$MSE = \sqrt{\frac{1}{N} | \sum_{i=1}^{N} |C_i - C_i^{GT}|^2}, \tag{6}$$

where $N$ is the number of images in the test set and $C_i^{GT}$ is the ground truth count of people in the $i^{th}$ test image. $C_i$ is the estimated number counted in the $i^{th}$ test image. This is defined as:

$$C_i = N_E + N_D, \tag{7}$$

where $N_D$ and $N_E$ are the numbers of people obtained by detection and estimation, respectively. Roughly speaking, the lower the MAE and MSE, the better accuracy the estimation.

## V. Experiments

We evaluate the performance of our approach on five public datasets, *i.e.*, ShanghaiTech, UCF, UCSD, WorldExpo'10, and UCF-QNRF [10], [8], [13], [35], [36] and compare with state-of-the-art methods [16], [17]. The details of each dataset are shown in Table I. In this section, we first present comparative experimental results on the above benchmark datasets. Then an ablation study conducted on the ShanghaiTech dataset is included to analyze the effect of applying our baseline idea of combining detection and estimation.

### A. Results on the ShanghaiTech dataset

The ShanghaiTech crowd counting dataset was firstly introduced by Zhang *et al.* [13]. This includes 1198 annotated images and in total 330,165 annotated people. This dataset is composed of two parts A and B, containing 482 and 716 images respectively. Images in Part A were downloaded from Internet and images in Part B were captured from streets in Shanghai. Moreover, images in Part A have more people in each image than that of the Part B images. In our experiments, 300 images of Part A were used for training and the other 182 images were used for testing. Similarly, for Park B, 400 images were used for training and the other 316 images were used for testing. The result of our method and other recent works are compared in Table II. As shown in the table, our method has achieved the lowest MAE (*i.e.*, the best accuracy) in ShanghaiTech Part A and achieved the second lowest MAE on ShanghaiTech Part B. Some sample images together with their results can be found in Fig. 4 on ShanghaiTech Part A and Fig. 5 on ShanghaiTech Part B.

TABLE II
COMPARISON WITH STATE-OF-THE-ART METHODS ON THE SHANGHAITECH DATASET

|  | partA |  | partB |  |
|---|---|---|---|---|
| Methods | MAE | MSE | MAE | MSE |
| Zhang *et al* [35] | 181.8 | 277.7 | 32.0 | 49.8 |
| MCNN [13] | 110.2 | 173.2 | 26.4 | 41.3 |
| Huang *et al* [37] | - | - | 20.2 | 35.6 |
| Switch-CNN [14] | 90.4 | 135.0 | 21.6 | 30.1 |
| CP-CNN [15] | 73.6 | 106.4 | 20.1 | 30.1 |
| CSRNet [16] | 68.2 | 115.0 | 10.6 | 16.0 |
| SANet [17] | 67.0 | 104.5 | **8.4** | **13.6** |
| ic-CNN [38] | 68.5 | 116.2 | 10.7 | 16.0 |
| DENet (ours) | **65.5** | **101.2** | 9.6 | 15.4 |

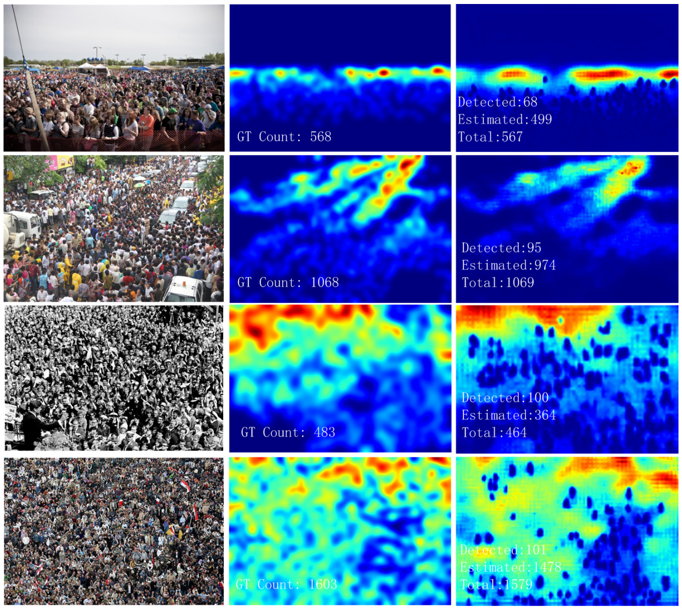

Fig. 4. Visualization of the estimated density maps of ShanghaiTech Part A using our approach.

### B. Results on UCF CC 50 dataset

The UCF CC 50 dataset includes 50 annotated images with 63,974 people in total. This dataset is downloaded from Internet with different perspective and resolutions [10]. The number of annotated people in each image ranges from 94 to 4543, and the average number of people per image is 1,280. 5-fold cross-validation is used to evaluate the performance, which follows the standard setting in [10]. The comparison results are presented in Table III and the visual qualities of generated density maps can be found in Fig. 6.

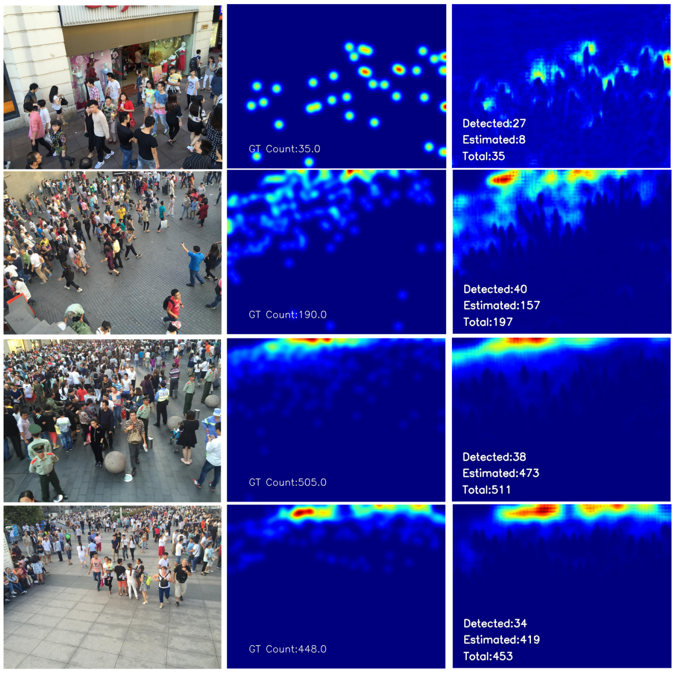

Fig. 5. Visualization of the estimated density maps of ShanghaiTech Part B using our approach.

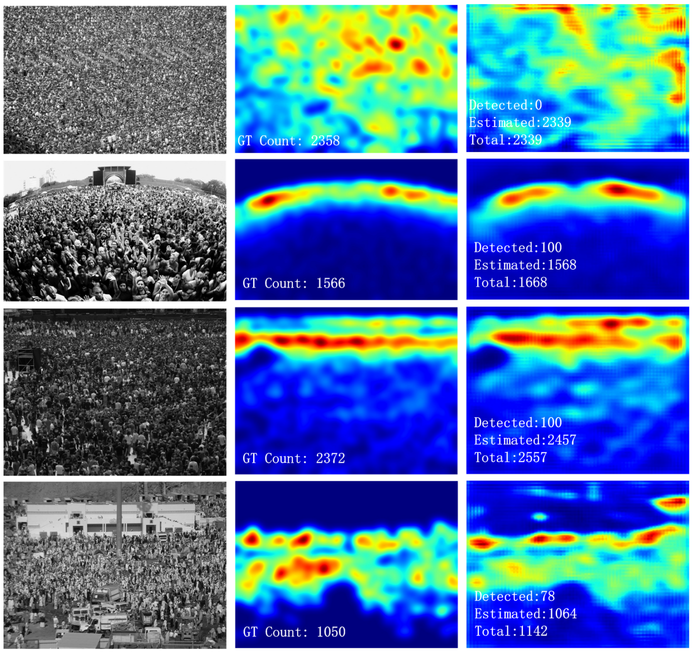

Fig. 6. Visualization of the estimated density maps of UCF dataset using our approach.

### C. Results on UCSD dataset

The UCSD dataset [8] has 2000 images taken from surveillance cameras with total 49,885 annotated people. The number of people in one image varies from 11 to 44 and the size of individual people in the image are similar. Following [8], the training set contains 800 images and the rest of the 1200 images are used for testing. Most people can be detected by DNet, and the results of UCSD dataset are shown in Table IV. The proposed algorithm outperforms existing methods except for the SANet [17] and Huang *et al* [37] in the MAE category. We provide more results in Table IV and Fig. 7.

### D. Results on WorldExpo'10 dataset

The WorldExpo'10 dataset [35] includes 3,980 annotated images from 1,132 video sequences taken from 108 surveillance cameras, which have 199,923 annotated people. This dataset consists of 3,380 images for training, and has five subsets each containing 120 images, for testing with different

TABLE III
COMPARISON WITH STATE-OF-THE-ART METHODS ON THE UCF DATASET

| Methods | MAE | MSE |
|---|---|---|
| Zhang *et al* [35] | 467.0 | 498.5 |
| MCNN [13] | 377.6 | 509.1 |
| Huang *et al* [37] | 409.5 | 563.7 |
| Hydra-2s [7] | 333.7 | 425.3 |
| Switch-CNN [14] | 318.1 | 439.2 |
| CP-CNN [15] | 295.8 | 320.9 |
| CSRNet [16] | 266.1 | 397.5 |
| SANet [17] | 258.4 | **334.9** |
| ic-CNN [38] | 260.9 | 365.5 |
| DENet (ours) | **241.9** | 345.4 |

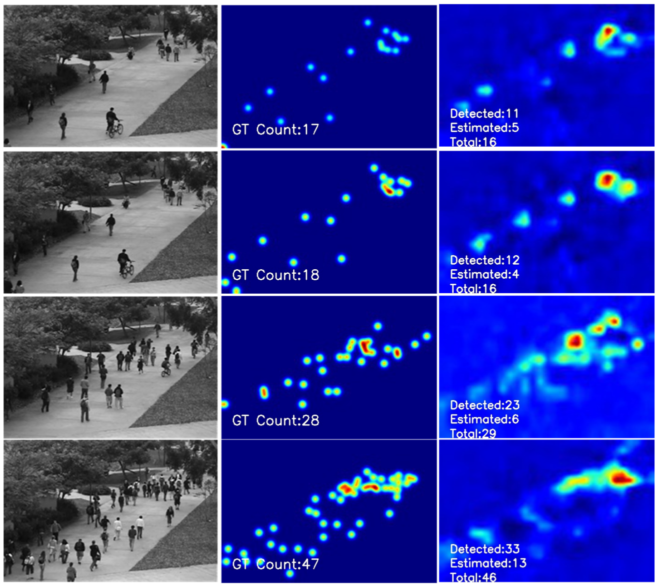

Fig. 7. Visualization of the estimated density maps of UCSD using our approach.

TABLE IV
COMPARISON WITH STATE-OF-THE-ART METHODS ON THE UCSD DATASET

| Methods | MAE | MSE |
|---|---|---|
| Zhang *et al* [35] | 1.60 | 3.31 |
| MCNN [13] | 1.07 | 1.35 |
| Huang *et al* [37] | **1.00** | 1.40 |
| CCNN [7] | 1.51 | - |
| Switch-CNN [14] | 1.62 | 2.10 |
| CSRNet [16] | 1.16 | 1.47 |
| SANet [17] | 1.02 | **1.29** |
| DENet (ours) | 1.05 | 1.31 |

scenes. We train our model following the instructions given in Section IV-A. Results are shown in Table V and Fig. 8. The proposed DENet delivers the best MAE in Scene1, Scene2 and Scene3, and it achieves the best accuracy on average.

### E. Results on UCF-QNRF dataset

The UCF-QNRF dataset was released by [36]. It contains 1535 high resolution images with 199,923 annotated people. The number of people in one image varies from 815 to 12865. The training dataset and testing sets contains 1,201 images and 334 images, respectively. The results for UCF-QNRF dataset are shown in Table VI. The proposed algorithm outperforms existing methods in the MAE category.

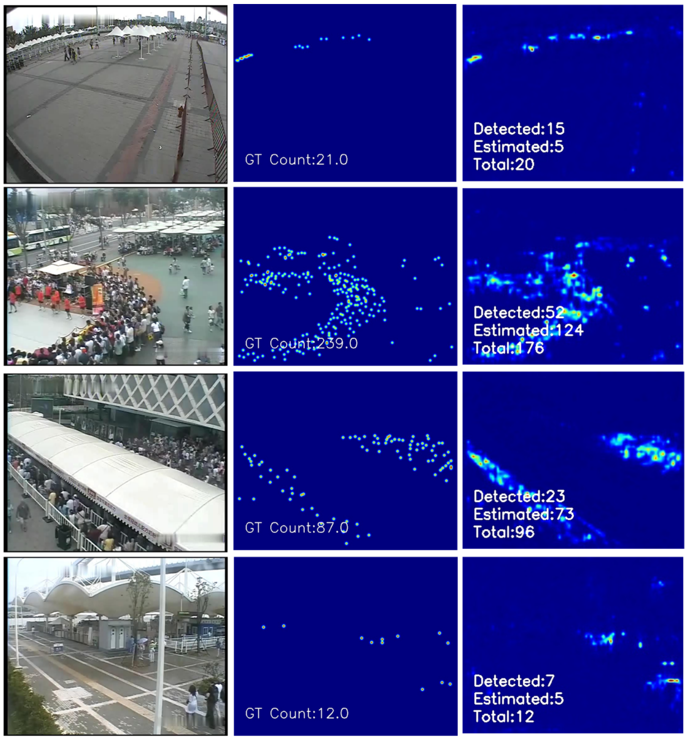

Fig. 8. Visualization of the estimated density maps of WorldExpo'10 for using our approach.

### F. Ablation studies

To demonstrate the effectiveness of our DENet idea, we implemented the MCNN and CSRNet to train them with our baseline idea of combining detection with the Mask R-CNN and their density-estimation networks. Based on the MCNN and CSRNet models, several ablation studies are conducted on the ShanghaiTech Part A and ShanghaiTech Part B datasets. The evaluation results are reported in Table VII. As shown in this table, all results are improved over those originally reported in [13] and [16]. Thus, we can conclude that most of the crowd counting estimation networks can be improved with our baseline idea.

## VI. CONCLUSION

In this paper, we have proposed a novel encoder-decoder architecture, which is called DENet, for accurate crowd counting. We used the Mask R-CNN to detect and segment people who can be clearly segmented, and have proposed a novel estimation network to create density maps. The numbers of detections and estimations are added to calculate the total number of people. By taking advantage of the dilated convolutional layers and transposed layer, our estimation network can create high-quality density maps without losing resolution. We have proposed a new loss function that combines counting loss and Euclidean loss to train our estimation network. Experiments have shown that our method has achieved better performance

TABLE V
COMPARISON WITH STATE-OF-THE-ART METHODS ON WORLDEXPO'10 DATASET

| Method | Scene1 | Scene2 | Scene3 | Scene4 | Scene5 | Average |
|---|---|---|---|---|---|---|
| Zhang *et al* [35] | 9.8 | 14.1 | 14.3 | 22.2 | 3.7 | 12.9 |
| MCNN [13] | 3.4 | 20.6 | 12.9 | 13.0 | 8.1 | 11.6 |
| Huang *et al* [37] | 4.1 | 21.7 | 11.9 | 11.0 | 3.5 | 10.5 |
| Switch-CNN [14] | 4.4 | 15.7 | 10.0 | 10.4 | 5.8 | 8.9 |
| CP-CNN [15] | 2.9 | 14.7 | 10.5 | 10.4 | 5.8 | 8.9 |
| CSRNet [16] | 2.9 | 11.5 | 8.6 | 16.6 | **3.4** | 8.6 |
| SANet [17] | 2.8 | 14.0 | 10.2 | **12.5** | 3.5 | 8.6 |
| ic-CNN [38] | 17.0 | 12.3 | 9.2 | 8.1 | 4.7 | 10.3 |
| DENet (ours) | **2.8** | **10.7** | **8.6** | 15.2 | 3.5 | **8.2** |

TABLE VI
COMPARISON WITH STATE-OF-THE-ART METHODS ON THE UCF-QNRF DATASET

| Methods | MAE | MSE |
|---|---|---|
| MCNN [13] | 277 | 426 |
| Switch-CNN [14] | 228 | 445 |
| [36] | 132 | 191 |
| CSRNet [16] | 129 | 209 |
| DENet (ours) | **121** | **205** |

TABLE VII
COMPARISON OF THE ESTIMATION ERROR OF TWO DIFFERENT NETWORK CONFIGURATIONS, *i.e.*, MCNN [13] AND CSRNET [16], COMBINING WITH OUR IDEA, *i.e.*, MASK R-CNN + MCNN AND MASK R-CNN + CSRNET.

| | partA | | partB | |
|---|---|---|---|---|
| Methods | MAE | MSE | MAE | MSE |
| MCNN [13] | 110.2 | 173.2 | 26.4 | 41.3 |
| Mask R-CNN + MCNN | **105.6** | **164.1** | **23.2** | **37.5** |
| CSRNet [16] | 68.2 | 115.0 | 10.6 | 16.0 |
| Mask R-CNN + CSRNet | **67.5** | **112.1** | **10.1** | **15.5** |

on some major crowd counting datasets compared to the state-of-the-art methods.

## ACKNOWLEDGMENT

This work was partly supported by National Natural Science Fund of China (No. 61725501 and No. 61976037), Specialized Research Fund for the Doctoral Program of Higher Education of China (No. 20121102110032).

## REFERENCES

[1] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.

[2] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *European conference on computer vision*. Springer, 2016, pp. 21–37.

[3] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1440–1448.

[4] G. J. Brostow and R. Cipolla, "Unsupervised bayesian detection of independent motion in crowds," in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, vol. 1. IEEE, 2006, pp. 594–601.

[5] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 1. IEEE, 2005, pp. 886–893.

[6] P. Dollár, B. Babenko, S. Belongie, P. Perona, and Z. Tu, "Multiple component learning for object detection," in *European conference on computer vision*. Springer, 2008, pp. 211–224.

[7] D. Onoro-Rubio and R. J. López-Sastre, "Towards perspective-free object counting with deep learning," in *European Conference on Computer Vision*. Springer, 2016, pp. 615–629.

[8] A. B. Chan, Z.-S. J. Liang, and N. Vasconcelos, "Privacy preserving crowd monitoring: Counting people without people models or tracking," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*. IEEE, 2008, pp. 1–7.

[9] A. B. Chan and N. Vasconcelos, "Counting people with low-level features and bayesian regression," *IEEE Transactions on image processing*, vol. 21, no. 4, pp. 2160–2177, 2011.

[10] H. Idrees, I. Saleemi, C. Seibert, and M. Shah, "Multi-source multi-scale counting in extremely dense crowd images," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2013, pp. 2547–2554.

[11] C. S. Regazzoni and A. Tesei, "Distributed data fusion for real-time crowding estimation," *Signal Processing*, vol. 53, no. 1, pp. 47–63, 1996.

[12] D. Lian, J. Li, J. Zheng, W. Luo, and S. Gao, "Density map regression guided detection network for rgb-d crowd counting and localization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1821–1830.

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

[14] D. B. Sam, S. Surya, and R. V. Babu, "Switching convolutional neural network for crowd counting," 2017, p. 6.

[15] V. A. Sindagi and V. M. Patel, "Generating high-quality crowd density maps using contextual pyramid cnns," in *2017 IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2017, pp. 1879–1888.

[16] Y. Li, X. Zhang, and D. Chen, "Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2018, pp. 1091–1100.

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

[18] Z.-Q. Cheng, J.-X. Li, Q. Dai, X. Wu, and A. G. Hauptmann, "Learning spatial awareness to improve crowd counting," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 6152–6161.

[19] Z. Ma, X. Wei, X. Hong, and Y. Gong, "Bayesian loss for crowd count estimation with point supervision," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 6142–6151.

[20] V. A. Sindagi and V. M. Patel, "A survey of recent advances in cnn-based single image crowd counting and density estimation," *Pattern Recognition Letters*, vol. 107, pp. 3–16, 2018.

[21] P. Viola and M. J. Jones, "Robust real-time face detection," *International journal of computer vision*, pp. 137–154, 2004.

[22] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Computer Vision (ICCV), 2017 IEEE International Conference on*. IEEE, 2017, pp. 2980–2988.

[23] C. Zhang, K. Kang, H. Li, X. Wang, R. Xie, and X. Yang, "Data-driven crowd understanding: A baseline for a large-scale crowd dataset," *IEEE Transactions on Multimedia*, vol. 18, no. 6, pp. 1048–1061, 2016.

[24] L. Zeng, X. Xu, B. Cai, S. Qiu, and T. Zhang, "Multi-scale convolutional neural networks for crowd counting," in *Image Processing (ICIP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 465–469.

[25] A. B. Chan and N. Vasconcelos, "Bayesian poisson regression for crowd counting," in *Computer Vision, 2009 IEEE 12th International Conference on*. IEEE, 2009, pp. 545–551.

[26] D. G. Lowe, "Object recognition from local scale-invariant features," in *Computer vision, 1999. The proceedings of the seventh IEEE international conference on*, vol. 2. IEEE, 1999, pp. 1150–1157.

[27] J. Liu, C. Gao, D. Meng, and A. G. Hauptmann, "Decidenet: counting varying density crowds through attention guided detection and density estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5197–5206.

[28] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[29] Y. Liu, M. Shi, Q. Zhao, and X. Wang, "Point in, box out: Beyond counting persons in crowds," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 6469–6478.

[30] X. Fan, C. Xiang, C. Chen, P. Yang, L. Gong, X. Song, P. Nanda, and X. He, "Buildsensys: Reusing building sensing data for traffic prediction with cross-domain learning," *IEEE Transactions on Mobile Computing*, 2020.

[31] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," *arXiv preprint*, pp. 1610–02 357, 2017.

[32] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," *arXiv preprint arXiv:1511.07122*, 2015.

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

[34] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," 2017.

[35] C. Zhang, H. Li, X. Wang, and X. Yang, "Cross-scene crowd counting via deep convolutional neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 833–841.

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

[37] S. Huang, X. Li, Z. Zhang, F. Wu, S. Gao, R. Ji, and J. Han, "Body structure aware deep crowd counting," *IEEE Transactions on Image Processing*, vol. 27, no. 3, pp. 1049–1059, 2018.

[38] V. Ranjan, H. Le, and M. Hoai, "Iterative crowd counting," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 270–285.

**Lei Liu** received his BSc and MSc degrees in University of Science and Technology Beijing (USTB) from 2006 to 2014. He is currently a PhD student in the School of Instrumentation Science and Opto-electronics Engineering, Beihang University (BUAA), China. He was a visiting student in University of Technology Sydney from 2017 to 2019. His research interest is deep learning and computer vision.

**Jie Jiang** received her Bachelor, Master, and PhD degrees from Tianjin University from 1991 to 2000. She is currently a Professor of the School of Instrumentation Science and Opto-electronics Engineering, Beihang University (BUAA), China. She has authored more than 50 articles and 30 inventions. Her research interests include image processing and machine vision.

**Wenjing Jia** received her PhD degree in Computing Sciences from University of Technology Sydney in 2007. She is currently a Senior Lecturer at the Faculty of Engineering and Information Technology (FEIT), University of Technology Sydney (UTS). Her research falls in the fields of image processing and analysis, computer vision and pattern recognition.

**Saeed Amirgholipour** is a PhD student at University of Technology Sydney, Australia. He received his Master degree in Computing Sciences from Isfahan university in 2009. Between 2011 to 2017, he was a Lecturer at Azad University, Iran. His research interest includes computer vision, deep learning video analytics, and image sentiment analysis.

**Yi Wang** received the BE and PhD degrees in 1010 computer science and technology from Jilin University, Jilin, China, in 2002 and 2009, respec tively. Since 2009, she has been with the Dalian University of Technology, China. She is currently an associate professor. Her research interests include machine learning, image processing, and computer vision.

**Michelle Zeibots** is a transport planner, specialising in the analysis of sustainable urban passenger transport systems. Her research, consultancy work and teaching draws together operational, behavioural and governance features relating to multi-modal urban transport networks.

**Xiangjian He** received his PhD degree in University of Technology, Sydney, Australia, in 1999. Since 1999, he has been with the University of Technology, Sydney, Australia. His research interests include image processing, network security, pattern recognition, computer vision and machine learning.