

# Crowd counting with crowd attention convolutional neural network

Jiwei Chen<sup>a,b</sup>, Su Wen<sup>c</sup> and Zengfu Wang<sup>a,b</sup>

<sup>a</sup>*Institute of Intelligent Machines, Chinese Academy of Sciences, Hefei, China*

<sup>b</sup>*Department of Automation, University of Science and Technology of China, Hefei, China*

<sup>c</sup>*Faculty of Mechanical Engineering and Automation, Zhejiang Sci-Tech University, Hangzhou, China*

## ARTICLE INFO

### Keywords:

Convolutional neural network  
Crowd counting  
Confidence map  
Density map

## ABSTRACT

Crowd counting is a challenging problem due to the scene complexity and scale variation. Although deep learning has achieved great improvement in crowd counting, scene complexity affects the judgement of these methods and they usually regard some objects as people mistakenly; causing potentially enormous errors in the crowd counting result. To address the problem, we propose a novel end-to-end model called Crowd Attention Convolutional Neural Network (CAT-CNN). Our CAT-CNN can adaptively assess the importance of a human head at each pixel location by automatically encoding a confidence map. With the guidance of the confidence map, the position of human head in estimated density map gets more attention to encode the final density map, which can avoid enormous misjudgements effectively. The crowd count can be obtained by integrating the final density map. To encode a highly refined density map, the total crowd count of each image is classified in a designed classification task and we first explicitly map the prior of the population-level category to feature maps. To verify the efficiency of our proposed method, extensive experiments are conducted on three highly challenging datasets. Results establish the superiority of our method over many state-of-the-art methods.

## 1. Introduction

Crowd counting by computer vision technology plays an important role in safety management [43], video surveillance [42], and urban planning [24]. The method of crowd counting can be also extended to other applications [21], such as cell counting, animal counting, and vehicle counting. However, due to the severe occlusion, scale variation, and high density in the crowd scene, crowd counting is still a challenging task.

To address these problems, a lot of efforts [31, 14] have been done in previous works including detection-based methods [41, 13, 17] and regression-based methods [3, 4, 12]. Detection-based methods [41, 13, 17] usually detect the instances of each person with pre-trained detectors [6, 37]. In the sparse crowd scene, they count the crowd accurately, while their accuracies are downgraded in the congested scene. Regression-based methods [3, 4, 12] regress the number of the crowd without detecting people. They implement an implicit mapping between low-level features and crowd counts. However, the location information of the crowd is omitted. So that many CNN-based methods with state-of-the-art results [45, 26, 30] are proposed recently. Most of them map the image to a density map that is more robust than the hand-crafted features. The quantity and location of the crowd at each pixel location are recorded in the density map. The crowd count can be obtained by integrating the density map.

Although CNN-based methods have achieved significant success in crowd counting, we find an important problem that needs to be solved urgently. Due to the complexity of crowd scenes, CNN-based methods usually mistake some

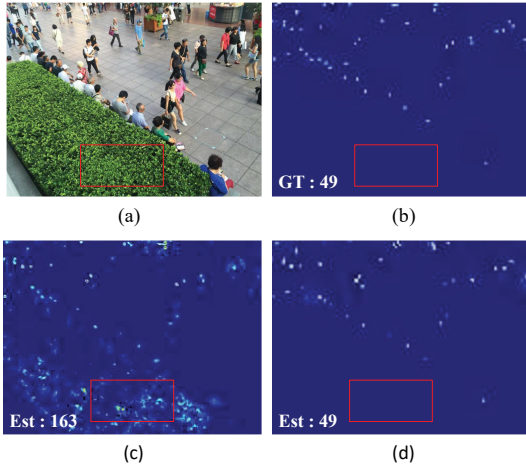
objects as the head of people. As shown in Fig. 1, there are no people inside the red box, however, MCNN [45] regards the dense shrubberies as human heads by mistake, which results in enormous errors of crowd counting.

To address the above problem, we propose a novel end-to-end model called CAT-CNN. An overview of the proposed CAT-CNN is shown in Fig. 2. It contains four modules: Multi-information Handling Module, Confidence Module, Density Map Estimation Module, and Fusion Module. The Multi-information Handling Module is utilized to extract robust features for crowd counting. Motivated by [2, 45], we leverage different convolution kernels to encode the input image at the beginning, then we fuse rich hierarchies from different convolutional layers, which is significant for extracting multi-scale features. In addition, the total crowd count of each image is classified [15] in a designed crowd count group classifier. To the best of our knowledge, we first explicitly map the weights of predicted class to feature maps to automatically contribute in encoding a highly refined density map. In the Confidence Module, we classify each pixel to obtain the probability of a human head at each pixel location to encode the confidence map. Unfortunately, the ground-truth confidence map is not provided in present crowd counting datasets. We propose a simple but effective way to obtain the ground-truth confidence map by pasting the ones template on a binary map. The intensive cost of manual labeling is saved. Meanwhile, to address the problem of unbalanced population distribution, we propose the weighted Binary Cross-Entropy Loss (BCELoss) to encode a robust confidence map for population distribution. In the Density Map Estimation Module, the estimated density map is encoded. In the Fusion Module, the estimated density map is multiplied by the pixel-level confidence map. With the guidance of confidence map, the position of human

✉ Corresponding author: zfwang@ustc.edu.cn (Z.

Wang)

ORCID(s): 0000-0003-1859-900X (Z. Wang)



**Figure 1:** Density estimation results. (a) Input image. (b) Ground-truth density map. (c) The density map estimated by MCNN. (d) The density map estimated by CAT-CNN. GT represents the ground-truth count. Est represents the estimated count.

head in the estimated density map gets more attention to encode the final density map and enormous misjudgements are avoided effectively. The final density map is integrated to get the crowd count. These modules work collaboratively to complete the crowd counting task. The training method of them is not as complex as these methods in [45, 26, 30]. They are trained jointly by minimizing their loss functions. They don't need pre-training and need to be trained only once.

Our contributions are summarized as follows:

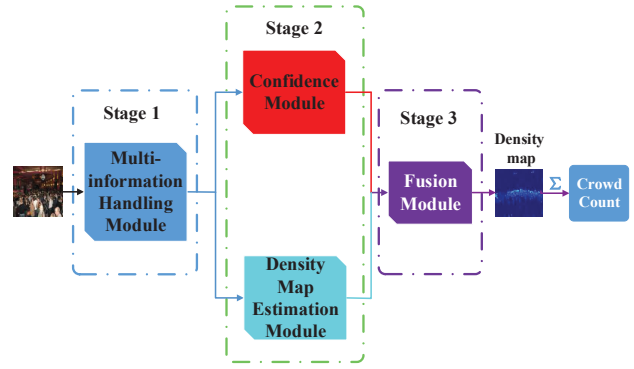
1. We propose the CAT-CNN that can adaptively assess the importance of a human head at each pixel location to avoid enormous misjudgements in crowd counting.
2. We design a novel classification model that can take input of arbitrary size for training in crowd counting. And we first explicitly map the prior information of the population-level category of images to feature maps to automatically contribute in encoding a highly refined density map.
3. Our CAT-CNN is a multi-stage and multi-supervision model. Meanwhile, it is robust to scale variations by the novel design in the Multi-information Handling Module. Extensive experiments demonstrate that our method outperforms many state-of-the-art methods on three highly challenging datasets ([12, 45, 43]).

## 2. Related work

In recent years, crowd counting has drawn much attention and various methods have been proposed, especially in deep learning. Next, we will give these methods some introductions.

### 2.1. Detection-based methods

Traditional detection-based algorithms such as Haar wavelets [33], HOG [6], and LBP [37] occupy an important



**Figure 2:** Overview of the proposed CAT-CNN. The Multi-information Handling Module is a feature extractor. The confidence map and estimated density map are generated respectively in the two middle modules. Then they are multiplied and further encoded to generate the high-precision density map in the Fusion Module.

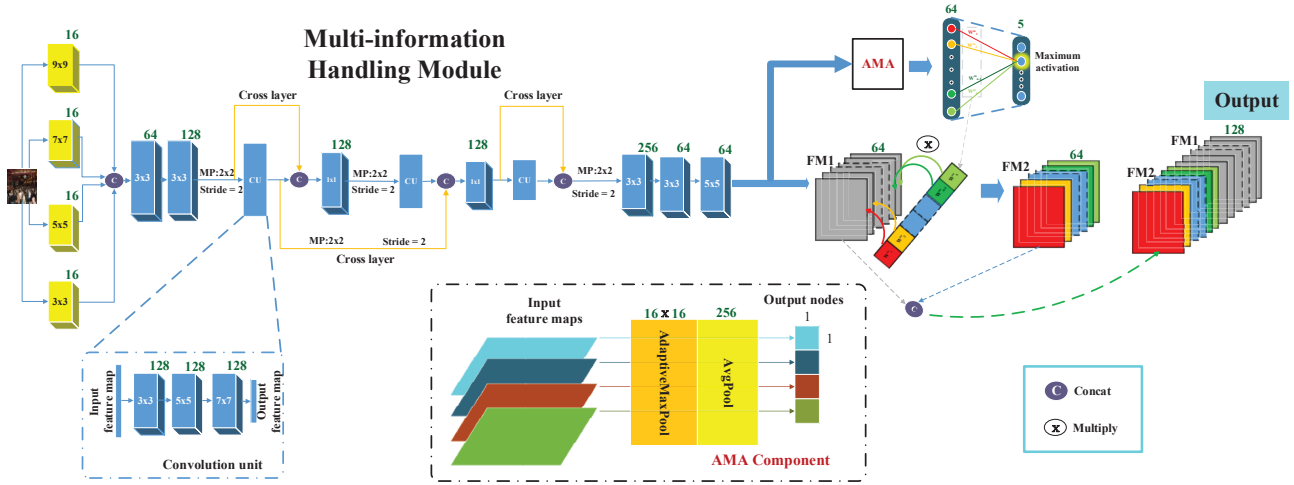
position in early works. Lin et al. [13] employed the Haar wavelet transform to detect the head-like contour. Zeng et al. [41] detected the head-shoulder of people by HOG and LBP. Lin et al. [17] proposed a part-template tree model for human detection. The crowd count can be obtained by summing the total number of positive samples in their methods. Detection-based methods perform very well in the sparse crowd. However, when the crowd becomes dense, some people are too small to be detected.

### 2.2. Regression-based methods

Since the accuracy of detection-based method is not very high in the highly congested scene, researchers attempt to use regression-based methods to handle this problem. The regression-based methods learn a mapping between high-level features and crowd counts. The high-level features are extracted from low-level information such as edge information [3], texture information [20], and segmentation information [24], then the crowd count is regressed according to high-level features. Chan et al. [3] proposed the Gaussian regression algorithm to learn a mapping between feature maps and crowd counts. Chan et al. [4] employed the Poisson regression algorithm to model the crowd count as the Poisson random variable. To decrease crowd counting errors, Idrees et al. [12] utilized some outstanding features such as people's heads to regress the crowd count. Although the regression-based methods can regress the crowd counts directly, the location information of each person is omitted.

### 2.3. CNN-based methods

Recently, due to the success of CNNs in many fields [32, 9, 19, 22, 11], CNN-based methods are widely used in crowd counting. The density map generated by CNNs records the count and location information of the crowd. Zhang et al. [45] proposed the MCNN to overcome scale variations. The MCNN leveraged three-branch CNNs with different convolution kernels to extract multi-scale features. Based on [45], Sam et al. [26] designed the Switch-CNN



**Figure 3:** The proposed architecture of the Multi-information Handling Module.

with a classifier to select the optimal branch to encode a density map according to the variation of crowd counts. To employ the temporal correlation across frames in video sequences to assist crowd counting, Xiong et al. [38] introduced the ConvLSTM that could extract bidirectional timing information. Sindagi et al. [30] proposed the CP-CNN to incorporate global and local contextual features to encode a high-quality density map. Li et al. [18] proposed the DecideNet model to adaptively leverage the estimations of detection and regression. Zhang et al. [44] proposed the MRA-CNN that could automatically focus on head regions by score maps. Hossain et al. [10] introduced a scale-aware attention mechanism to adapt the scale variation of crowds. Wang et al. [36] designed the data collector and labeler to automatically generate and annotate the crowd data to reduce over-fitting caused by limited training data. In this paper, we propose a novel method to avoid misjudgements that can result in enormous errors of crowd counting.

### 3. Proposed methods

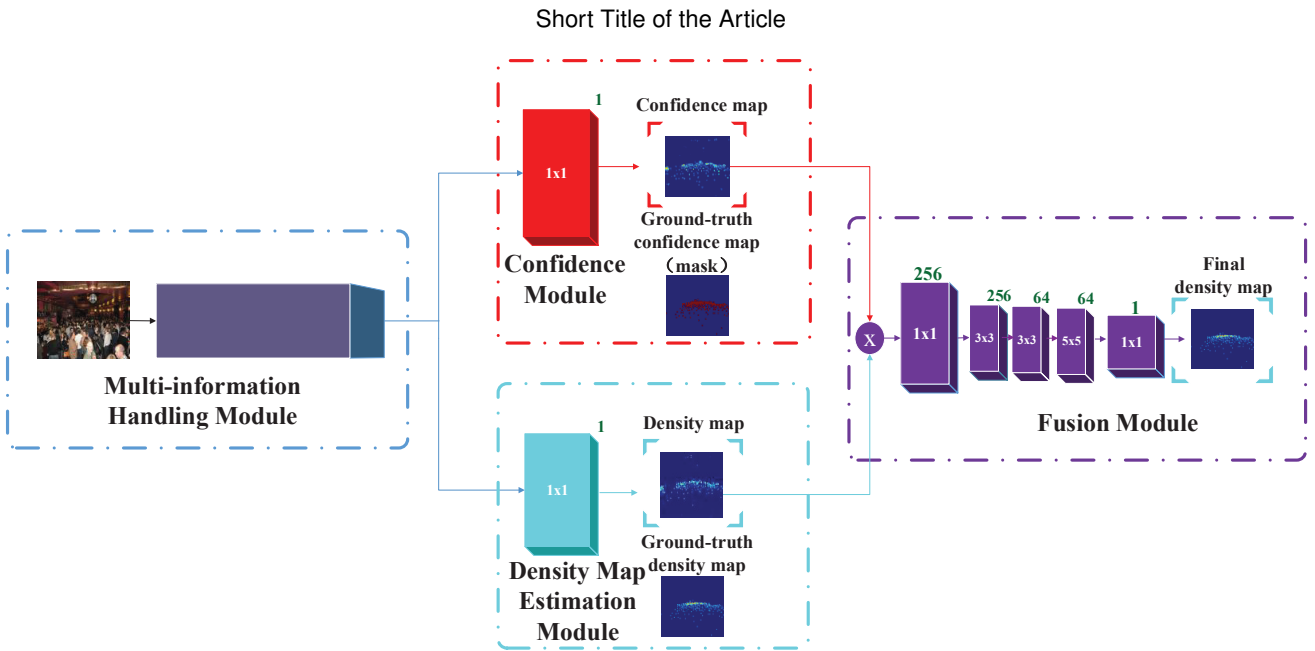
#### 3.1. Network architecture

An overview of the proposed CAT-CNN is shown in Fig. 2. Our CAT-CNN is composed of three stages. The first stage contains the first module where the features which can automatically adapt different scales and different crowd count groups are extracted. The second stage consists of two modules in the middle to encode confidence map and estimated density map respectively. The third stage contains the final module. With the guidance of the confidence map, final density map is encoded from the estimated density map in this stage. Next, we will elaborate these modules in each stage.

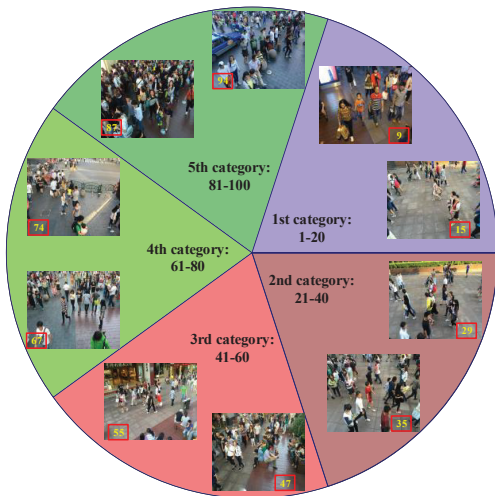
**Multi-information Handling Module:** This module is proposed to overcome the scale variation and explicitly map the prior information of the population-level category back to feature maps to automatically contribute in encoding a

highly refined density map. This module is illustrated in Fig. 3. To extract multi-scale features to overcome the scale variation, inspired by MCNN [45], we exploit four different kernels to convolve the input image at the beginning of this module. Besides, several  $2 \times 2$  max-pooling layers are designed inside this module. And we fuse convolutional layers of different depths to fully excavate the multi-scale features. To alleviate overfitting caused by redundant parameters, we widely employ the dilated convolution [40] in our CAT-CNN which can expand the receptive fields of convolution with fewer parameters. In this paper, all convolutions of different shapes are constituted by the  $3 \times 3$  convolution with corresponding dilation, except for the  $1 \times 1$  convolution. Every convolution is followed by a rectified linear unit (ReLU) [7].

Inspired by [29], the crowd counts are quantized into five groups in each dataset and a crowd count group classifier is learned. A simple example about the process of statistics and classification of datasets is shown in Fig. 5. In the crowd count group classifier, the total crowd count of each image is classified. To feed arbitrarily sized images into the fully-connected (FC) layer for training without resizing images to maintain the original distribution of the crowd, inspired by [16, 29], we employ AdaptiveMaxPool and AvgPool (AMA) to design a novel model named AMA Component. As shown in Fig. 3, arbitrarily sized input can be fed to the AMA Component. And the output is always  $1 \times 1$  node. This component is placed between convolutional layer and FC layer to form the novel classification model. Inspired by [46], the major improvement is that AMA has one more AvgPool layer than SPP used in [29, 39]. The purpose of this design is to directly map the weights of the predicted population-level category back to feature maps, which is also the main distinction from [29, 39]. We first explicitly map the prior information of the population-level category back to feature maps to automatically contribute in encoding a highly refined density map.



**Figure 4:** The proposed architecture of our CAT-CNN.



**Figure 5:** A simple example about the process of statistics and classification of datasets: The crowd counts range from 1 to 100 in a dataset and they are quantized into five groups. The crowd count group classifier can divide the image into corresponding groups according to the crowd count in the image. The number in red box represents the crowd count in images.

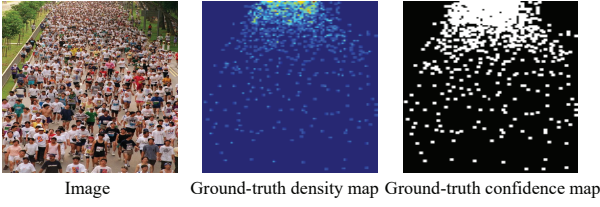
As shown in Fig. 3, we leverage the feature map (FM1) to feed the AMA Component and outputs are sent to the FC layer followed by Parametric ReLU [8]. The maximum activated neuron in FC layer represents the population-level category of input image. We explicitly use the prior of the population-level category by multiplying the weights of predicted category to feature maps (FM1). The mapped feature maps (FM2) can be used as a feature vector representation to characterize the population-level category. FM2 is concatenated with FM1 to serve as the output of this module

to retain more sufficient feature information for following modules. We choose minimizing the cross-entropy loss to optimize the novel classification task.

**Confidence Module:** Due to the complexity of crowd scenes, CNN-based methods often mistake some objects as human heads. If the prior of whether there are people at each pixel location can be used, this problem will be solved easily. We propose the Confidence Module based on probability. The probability of each pixel belonging to a person's head can be estimated by dense classifications.

The proposed Confidence Module is shown in Fig. 4. The output of the Multi-information Handling Module is fed into this module. With the supervision of ground-truth confidence map, the  $1 \times 1$  convolution followed by ReLU is utilized to encode the feature maps and transform them into predicted confidence map. The ground-truth confidence map is a binary map. We set the probability of the pixel belonging to a human head to 1. The predicted confidence map records the estimated possibility of a human head at each pixel location. The greater probability in the predicted confidence map, the more attention our model will pay in the density map by multiplying the confidence map in the Fusion Module. Due to the unbalanced population distribution, we propose the weighted BCEloss to obtain a robust confidence map for population distribution. The weighted BCEloss will be elaborated in Sec. 3.2.

**Density Map Estimation Module:** The estimated density map of human heads is encoded in this module. As shown in Fig. 4, the output of the Multi-information Handling Module is fed into this module. With the supervision of ground-truth density map, the  $1 \times 1$  convolution followed by ReLU is utilized to encode the feature map to the estimated density map. The ground-truth density map will be elaborated in Sec. 3.2. Euclidean loss is used to optimize the estimated density map.



**Figure 6:** Visualization results of the ground-truth density map and ground-truth confidence map (mask).

**Fusion Module:** Firstly, we distinguish the difference between confidence map and estimated density map in detail. The classification-based confidence map reflects the possibility of a human head at each pixel location. But it can't be used to count the crowd directly. The estimated density map contains the pivotal information of crowd counts. However, it usually regards some objects as human heads mistakenly. We combine both to avoid the misjudgements in crowd counting. As shown in Fig. 4, the estimated density map is multiplied by confidence map and the results are sent to the following convolutional layer and ReLU. With the guidance of the confidence map, the position of human head in the estimated density map gets more attention to encode the final density map. Then enormous misjudgements can be avoided in the final density map. Euclidean loss is used to optimize the final density map. The crowd count can be obtained by integrating the final density map. In [44], its attention mechanism is also important in MRA-CNN, which inspires us to explore the attention mechanism of MRA-CNN in the future.

### 3.2. Implementation

**Ground truth generation:** Following the method in [29], the head positions  $P$  are provided in each image  $I_i$ . All of annotated points  $A_i$  are convolved by a Gaussian kernel centered on each annotated point to encode the ground-truth density map. The Gaussian kernel with  $\sigma = 4.0$  has been normalized to 1. The density of a specific pixel  $p$  in one image can be regarded as the Gaussian function effects of its surrounding effective annotated points:

$$p \in I_i, D(p) = \sum_{P \in A_i} N(p; \mu_i, \sigma^2 I_{2 \times 2}), \quad (1)$$

where  $N$  represents the response of a 2D Gaussian function with its mean  $\mu_i$ .  $I_{2 \times 2}$  represents its isotropic  $2 \times 2$  covariance matrix with variance  $\sigma^2$ . We can integrate the density value at each pixel over the entire map to get the crowd count  $\sum_{p \in I_i} D(p) = Num_i$ .

Due to the enormous number of people in datasets, the cost of manually labeling the ground-truth confidence map is expensive. For example, the ShanghaiTech Part\_A [45] dataset contains 241,677 people. We propose a simple but effective method to obtain the ground-truth confidence map. Specifically, the ones template where all pixels are set to 1 is first generated. Then the ones template centered on

each annotated point is pasted on a binary map to encode the ground-truth confidence map (mask). The ones template and the Gaussian kernel have the same size that is set to  $15 \times 15$  in [29]. The ground-truth confidence map (mask) generation uses the same size on different datasets. In the ground-truth confidence map (mask), '1' represents 100% possibilities of a human head at this pixel location and '0' represents 0 possibility of a human head at this pixel location. The BCEloss that can predict the probability of a positive sample is utilized to optimize the confidence map by supervised learning. The greater probability in the predicted confidence map, the more attention our model will pay in the density map by multiplying the confidence map. Visualization results are shown in Fig. 6.

**Module Optimization:** Our CAT-CNN contains four modules. The whole loss function  $L_{whole}$  is given by Eq.(2):

$$L_{whole} = L_{fus} + L_{den} + \lambda_1 L_{con} + \lambda_2 L_{mul}, \quad (2)$$

where  $L_{mul}$ ,  $L_{con}$ ,  $L_{den}$ , and  $L_{fus}$  are the losses for Multi-information Handling Module, Confidence Module, Density Map Estimation Module, and Fusion Module, respectively.  $\lambda_1$  and  $\lambda_2$  are the hyper-parameters to balance the magnitude of two tasks.  $\lambda_1$  is set to 2 and  $\lambda_2$  is set to  $1e-2$ .

In the Multi-information Handling Module, we classify the crowd into five groups according to the crowd counts in each dataset. The cross-entropy loss function is used in this module:

$$L_{mul} = -\frac{1}{N} \sum_{i=1}^N \sum_{j=1}^K [(q^i = j) G_{mul}(X_i, \theta)], \quad (3)$$

where  $N$  represents the total number of training images.  $\theta$  represents a set of network parameters.  $X_i$  represents the  $i^{th}$  training image.  $q^i$  represents the ground-truth class and  $K$  represents the whole number classes.  $G_{mul}(X_i, \theta)$  represents the output of classification.

In the Confidence Module, we first propose employing classification at each pixel location to judge whether there are people. Due to the unbalanced population distribution in images, we propose the weighted BCELoss to encode a robust confidence map for population distribution. The weighted binary-cross entropy loss adopted in this module is shown as follows:

$$L_{con} = -w_i [y_i \log X_i + (1 - y_i) \log (1 - X_i)], \quad (4)$$

where  $X_i$  represents the confidence map that records the predicted probability of a positive sample.  $y_i$  indicates the ground-truth confidence map (mask).  $w_i$  is the weight given to the loss of each element. The derivation process of  $w_i$  is given as follows:

**Table 1**

Statistics of training datasets: Num represents the number of images; Range represents the range of crowd counts; Average represents the average crowd counts; Total represents the total number of labeled people.

| Dataset             | Resolution | Color    | Num  | Range      | Average | Total   |
|---------------------|------------|----------|------|------------|---------|---------|
| UCF_CC_50           | different  | Grey     | 50   | [94, 4543] | 1279.5  | 63,974  |
| 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  |
| WorldExpo'10        | 576 x 720  | RGB      | 3980 | [1, 253]   | 50.2    | 199,923 |

$$\begin{aligned}
 w_i &= f_i \cdot \bar{y}_i + b_i \cdot y_i, \\
 f_i &= \text{mean}(y_i), \\
 b_i &= 1 - f_i, \\
 \bar{y}_i &= 1 - y_i.
 \end{aligned} \tag{5}$$

$$w_i = \begin{cases} 10^{-6}, & \text{if } f_i = 0; \\ (1 - 2f_i)y_i + f_i, & \text{else.} \end{cases} \tag{6}$$

Where  $f_i$  represents the weight of crowd positions, and  $b_i$  represents the weight of background positions.  $\bar{y}_i$  represents the ground truth of background. For better convergence of our CAT-CNN, when  $f_i = 0$  which means that there is no person in the training image, we set  $w_i$  to  $10^{-6}$  instead of 0.

In the Density Map Estimation Module and Fusion Module, the Euclidean distance loss is used to optimize these modules:

$$\begin{aligned}
 L_{den} &= \frac{1}{2M} \sum_{i=1}^M \|D(X_i, \theta) - D_i\|_2^2, \\
 L_{fus} &= \frac{1}{2M} \sum_{i=1}^M \|F(X_i, \theta) - D_i\|_2^2,
 \end{aligned} \tag{7}$$

where  $D(X_i, \theta)$  represents the estimated density map in the Density Map Estimation Module.  $F(X_i, \theta)$  represents the final density map in the Fusion Module. These two modules have the same ground-truth density map  $D_i$ .  $M$  represents the total number of training samples.

Each module has its subtask. These four subtasks are completed synchronously in a synergistic manner. The optimized process is a multi-task learning [5, 34] which is helpful to reduce overfitting caused by limited data for training. Multi-task learning can extract the generalizable representation of similar data and assist our model in alleviating crowd counting errors. [39] also adopts the strategy of multi-task learning. There are two main differences between two works. Firstly, our proposed CAT-CNN is a multi-stage model. The confidence map is encoded in the second stage. In the third stage, the predicted confidence map is multiplied by the estimated density map to directly participate



**Figure 7:** Some representative examples from these datasets.

in the generation of final density map to avoid enormous misjudgements. In [39], the BG/FG mask is not directly involved in density map generation, but fine-tunes the density map as a multi-task learning. The rationality of two designs is guaranteed by different loss functions. Secondly, for the classification task in Fig. 3, inspired by [46], we design the AMA Component and the weights of the predicted category are directly mapped back to previous feature maps to contribute in generating a highly refined density map in our work, which is the second difference from [39].

## 4. Experiments

Our model is trained on three benchmark datasets separately: ShanghaiTech dataset [45], UCF\_CC\_50 dataset [12], and WorldExpo'10 dataset [43]. Statistics of them are summarized in Table 1. Some example images of these datasets are shown in Fig. 7. The data augmentation schemes in [29] are utilized to reduce overfitting. In the process of creating training datasets, 9 patches are cropped from each original image at random locations, and the size of each patch is  $1/4^{th}$  of the original image. Random flipping and uniform noise are used in the patch with a probability of 0.5. The Adam optimization algorithm with a batch size of 1 is used to optimize these loss functions in our model. Its parameter of learning rate is set to  $1e-6$ . The number of epochs is set to 5000. In every iteration, gradients for each loss function are calculated and corresponding parameters are updated. The curve of each loss function is expected to converge to smaller values. It may have a brief shock at the beginning. The experiments are conducted

**Table 2**

Estimation errors (MAE) of different configurations about the number of population-level categories.

|                     | population-level categories |              |       |       |            |
|---------------------|-----------------------------|--------------|-------|-------|------------|
|                     | 3                           | 5            | 7     | 10    | 15         |
| ShanghaiTech Part_A | 78.2                        | <b>66.7</b>  | 69.8  | 71.8  | 71.3       |
| ShanghaiTech Part_B | 15.3                        | <b>11.2</b>  | 13.2  | 12.9  | 12.2       |
| UCF_CC_50           | 308.4                       | <b>235.5</b> | 280.3 | 301.6 | 340.4      |
| WorldExpo'10        | 8.6                         | 7.2          | 8.0   | 8.5   | <b>7.1</b> |

**Table 3**

Estimation errors of different kernel sizes at the beginning of CAT-CNN.

|                       | Part_A      |              | Part_B      |             | UCF_CC_50    |              | WorldExpo'10 |            |
|-----------------------|-------------|--------------|-------------|-------------|--------------|--------------|--------------|------------|
|                       | MAE         | MSE          | MAE         | MSE         | MAE          | MSE          | MAE          | MSE        |
| 3×3                   | 76.8        | 125.9        | 14.3        | 24.1        | 289.6        | 441.5        | 8.4          | 11.3       |
| 3×3,5×5,7×7           | 78.2        | 128.1        | 13.4        | 22.7        | 306.4        | 465.6        | 9.2          | 13.8       |
| 3×3,5×5,7×7,9×9       | <b>66.7</b> | <b>101.7</b> | <b>11.2</b> | <b>20.0</b> | <b>235.5</b> | <b>324.8</b> | <b>7.2</b>   | 9.5        |
| 3×3,5×5,7×7,9×9,11×11 | 73.8        | 120.3        | 12.8        | 22.8        | 271.8        | 370.2        | 7.4          | <b>9.3</b> |

on NVIDIA GTX 1080Ti and Intel Core i7 with the Torch framework.

Following existing methods [45, 26, 30, 25, 18, 27, 28, 1], the Mean Absolute Error (MAE) and Mean Squared Error (MSE) are used to measure the count errors. They are defined as:

$$MAE = \frac{1}{N} \sum_{i=1}^N |z_i - \tilde{z}_i|, MSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (z_i - \tilde{z}_i)^2}. \quad (8)$$

Where  $N$  is the number of test images.  $z_i$  is the actual crowd count by integrating the ground-truth density map, and  $\tilde{z}_i$  is the estimated crowd count by integrating the final density map.

#### 4.1. Ablation study

To further demonstrate the effectiveness of different components in our CAT-CNN, we perform an ablation study by a discussion.

**Benefits of Multi-Scale Features:** The crowd at different distances from the camera have different scale characteristics in the image. So it is very important to extract multi-scale features. In [45], Zhang et al. proposed the three-column CNN structures to extract multi-scale features. However, every column needs to be pre-trained separately and the multi-column CNNs are hard to be trained. By contrast, our model is more convenient and effective to extract multi-scale features.

As shown in Fig. 3, at the beginning of our model, the convolution kernels with different scales are used to map the head to feature maps from input image. From Table 3, it can be found that the performance of using four type convolution kernels is best in most cases. As shown in Fig. 3, inside the network, feature maps from convolutional layers of

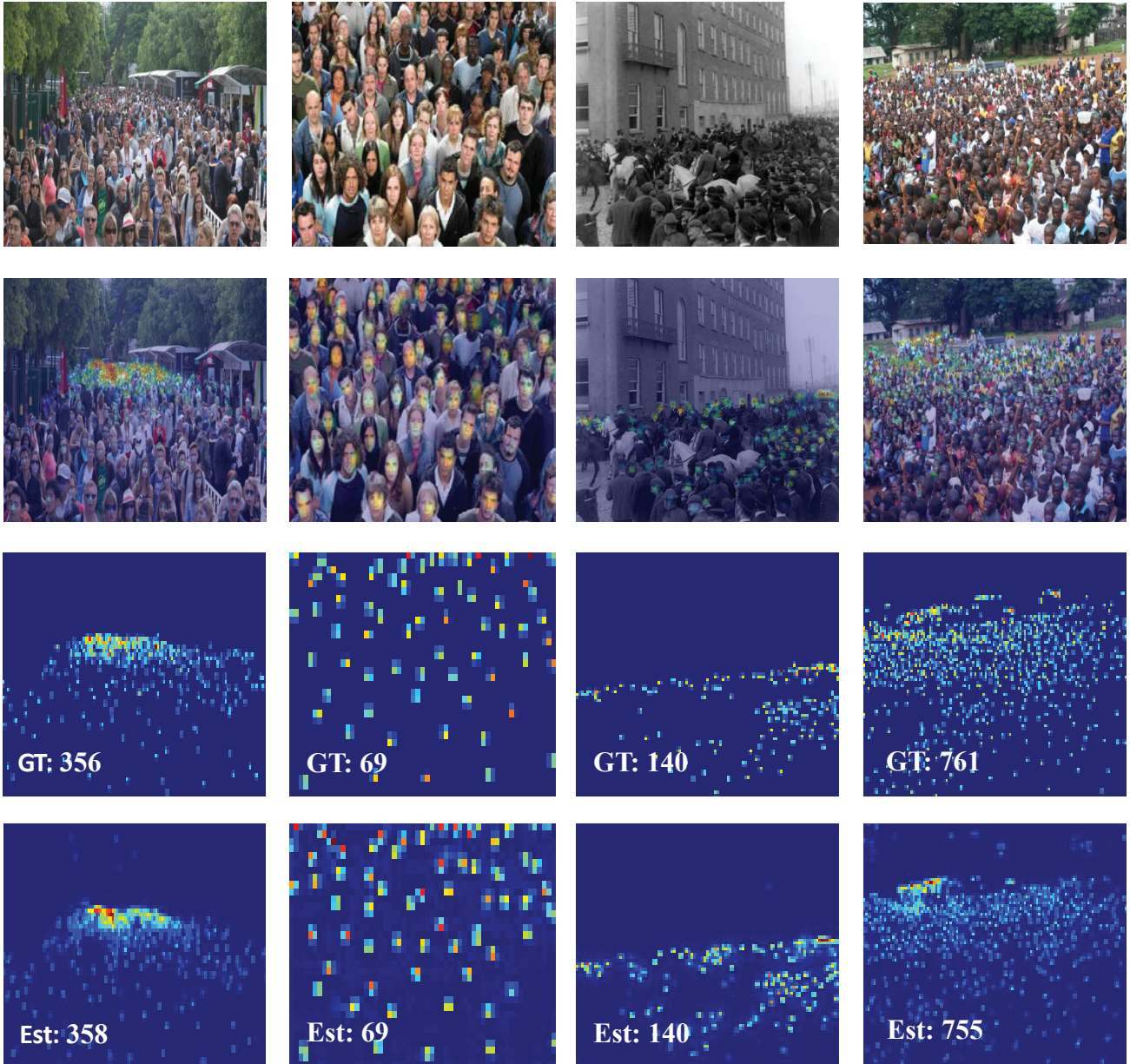
different depths are merged by cross-layer connection to automatically adapt the scale variations in the crowd. Higher layers encode the semantic concept of the person, whereas lower layers extract rich discriminative features from the person. Both of them can provide complementary information on the same person with different levels. Hence, from Table 4, we can observe that fusing features from convolutional layers of different depths is very effective to alleviate the crowd counting errors on these benchmarking datasets.

**Benefits of The Prior of Population-Level Categories:** In our CAT-CNN, we classify the crowd counts of images in each dataset into 5 categories according to experiments. The results are shown in Table 2. When the number of categories is set to 15 on the WorldExpo'10 dataset, the performance is improved slightly compared with the third column. We think that due to a large number of images in this dataset, 15 categories are suitable. However, it can be observed that the performance of 5 categories is best in most cases. We argue that due to the limitations of model capacity and training data, 5 categories are sufficient for most datasets. To prove the effectiveness of explicit use of the prior of population-level category, in the comparative experiment, FM1 and FM2 respectively serve as the output of the Multi-information Handling Module. The results are shown in Table 5. We can observe that by only using FM2 where we explicitly use the prior of population-level categories, the performance is better, with the MAE/MSE 5.4/6.2 lower than only using FM1 where we do not explicitly use the prior of population-level categories. When FM1 and FM2 are concatenated and used simultaneously, the performance is the best, with the MAE/MSE 10.0/13.6 lower than only using FM1. We argue that both FM1 and FM2 can encode the crowd count feature. The concatenation of them can provide more sufficient feature information for following modules.

**Table 4**

Estimation errors of different configurations about the cross-layer connection.

|                                | Part_A      |              | Part_B      |             | UCF_CC_50    |              | WorldExpo'10 |            |
|--------------------------------|-------------|--------------|-------------|-------------|--------------|--------------|--------------|------------|
|                                | MAE         | MSE          | MAE         | MSE         | MAE          | MSE          | MAE          | MSE        |
| Without cross-layer connection | 75.5        | 124.9        | 16.3        | 28.2        | 264.2        | 382.8        | 10.1         | 14.2       |
| With cross-layer connection    | <b>66.7</b> | <b>101.7</b> | <b>11.2</b> | <b>20.0</b> | <b>235.5</b> | <b>324.8</b> | <b>7.2</b>   | <b>9.5</b> |

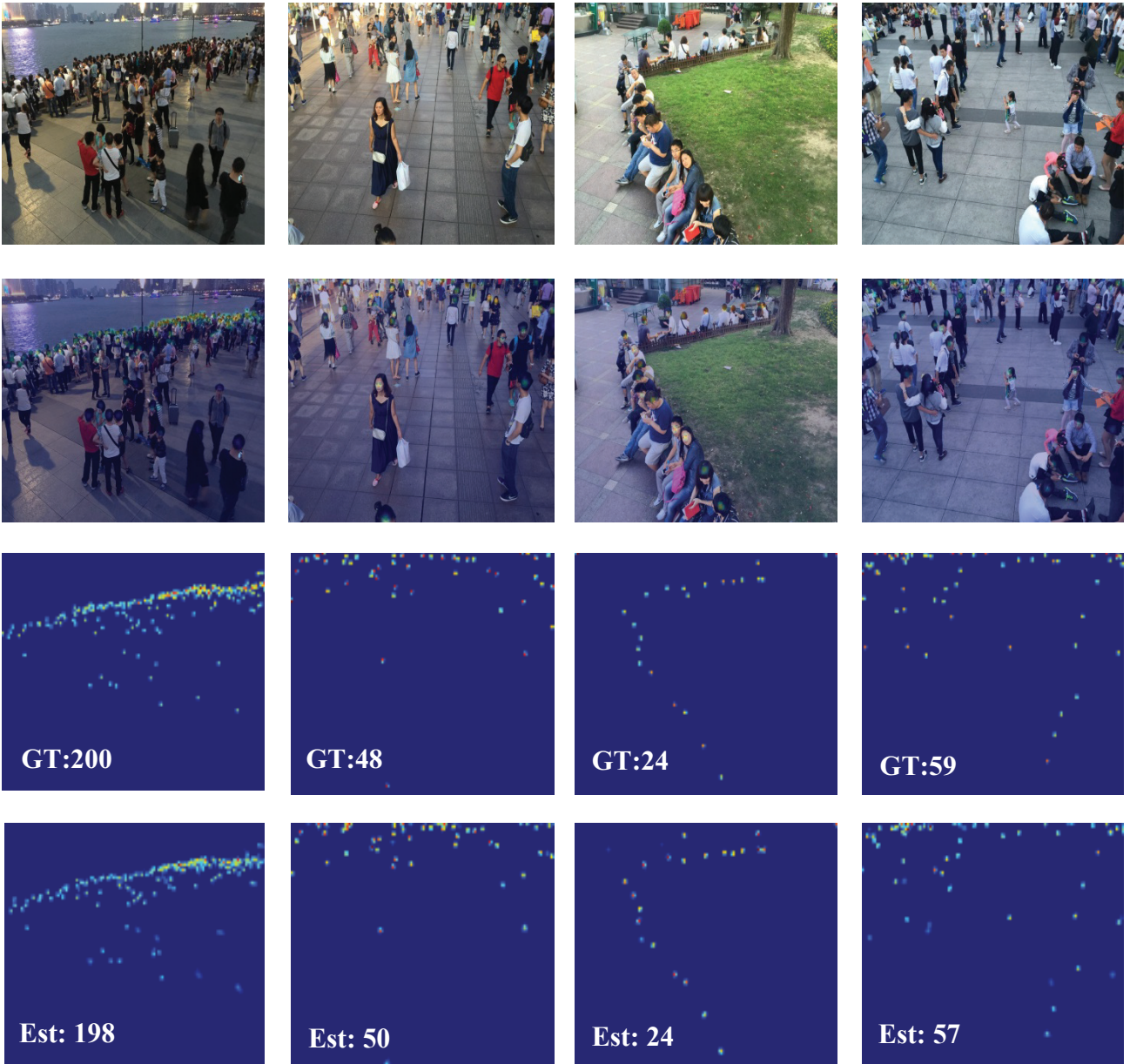


**Figure 8:** Results of the proposed CAT-CNN on ShanghaiTech Part\_A. First Row: test image. Second Row: test image overlaid by the estimated confidence map. Third Row: ground-truth density map. Fourth Row: the final estimated density map.

**Benefits of The Confidence Map:** To demonstrate the effectiveness of the confidence map, performances of our model with and without confidence map are compared. In the comparative experiment, the Confidence Module is removed from our CAT-CNN. Comparison results are given

in Table 6. We can see that the performances of our model are further enhanced by a confidence map generated in the Confidence Module, with the MAE/MSE 14.5/27.2 lower than that without a confidence map. The estimated count of each image and its actual count are shown in Fig. 10.





**Figure 9:** Results of the proposed CAT-CNN on ShanghaiTech Part\_B. First Row: test image. Second Row: test image overlaid by the estimated confidence map. Third Row: ground-truth density map. Fourth Row: the final estimated density map.

**Table 5**

Estimation errors of different configurations about the output of Multi-information Handling Module on ShanghaiTech Part\_A.

| Method      | MAE         | MSE          |
|-------------|-------------|--------------|
| Only FM1    | 76.7        | 115.3        |
| Only FM2    | 71.3        | 109.1        |
| FM1 and FM2 | <b>66.7</b> | <b>101.7</b> |

We can see that the green line (estimated count with confidence map) is closer to the red line (actual count) than the blue line (estimated count without confidence map), which

**Table 6**

Estimation errors of different configurations about the confidence map on ShanghaiTech Part\_A.

| Method                 | MAE         | MSE          |
|------------------------|-------------|--------------|
| Without confidence map | 81.2        | 128.9        |
| With confidence map    | <b>66.7</b> | <b>101.7</b> |

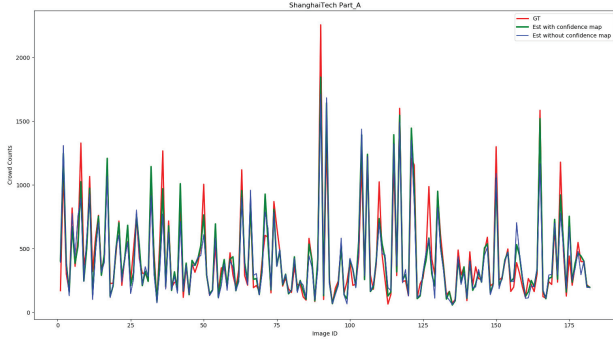
denotes that the confidence map plays an important role in reducing the error of crowd counting.

To visualize the effectiveness of confidence map, the confidence map is resized to the same size as the input image. And it is overlaid on the input image with 70% trans-

**Table 7**

Comparison of real time in Cascade-MTL, MRA-CNN, and CAT-CNN. Time represents the time to process a frame.

| Method           | Time (s) |      |      |      |      | Average |
|------------------|----------|------|------|------|------|---------|
|                  | S1       | S2   | S3   | S4   | S5   |         |
| Cascade-MTL [29] | 0.08     | 0.08 | 0.08 | 0.08 | 0.08 | 0.08    |
| MRA-CNN [44]     | 0.02     | 0.02 | 0.02 | 0.02 | 0.02 | 0.02    |
| CAT-CNN(OURS)    | 0.03     | 0.03 | 0.03 | 0.03 | 0.03 | 0.03    |



**Figure 10:** The crowd count comparison of different configurations about the confidence map on ShanghaiTech Part\_A. X-axis: the ID of images, Y-axis: the crowd counts of images.

parency to display both of them clearly. They are shown in the second row of Fig. 8. We can observe that only the position of the human head is highlighted, which denotes that our CAT-CNN encodes an accurate confidence map to avoid enormous misjudgements. The ground-truth density map and final high-precision density map are also shown in Fig. 8. We can observe that our CAT-CNN encodes a high-precision density map to count the crowd very well. Similar results can be found in Fig. 9.

**Evaluations of The Real Time:** To verify the real-time performance of our proposed method, some experiments are conducted on a video surveillance dataset (WorldExpo'10 dataset) using NVIDIA GTX 1080Ti and Intel Core i7. Its test set contains 600 frames from 5 different scenes (S1-S5). The resolution of each frame is 576×720. Comparisons with other state-of-the-art methods are given in Table 7. It can be observed that our proposed method obtains competitive results.

## 4.2. Shanghaitech dataset

Zhang et al. [45] introduced the large-scale ShanghaiTech dataset. It consists of two parts: Part\_A with 300 training images and 182 test images, Part\_B with 400 training images and 316 test images. In Part\_A, the crowd density is high while the crowd density is relatively low in Part\_B. We generate the ground-truth density map by using the method in Sec. 3.2.

In Table 9, we compare our method with other recent state-of-the-art methods. The LBP and RR are traditional algorithms for crowd counting. In [36], the data collector and labeler were proposed to generate and annotate the

crowd data. In [45], the MCNN with three-column CNNs was proposed to extract multi-scale features to adapt scale variations. In [26], the Switch-CNN with a classifier was proposed to select an optimal branch to generate the density map. In [30], global and local features were used to generate a high-quality density map. In [25], the top-down feedback TDF-CNN was proposed to get initial accurate prediction. In [18], detection and regression were used for crowd counting simultaneously. In [27], the GANs were proposed to mitigate blurring in the estimated density map. In [44], head regions were focused on automatically. In [28], the deep negative correlation learning was proposed to reduce over-fitting. In [10], a scale-aware attention model was proposed to adapt the scale variation of crowds. In [1], a growing CNN was proposed to adapt the variability seen from the crowd by increasing its capacity. In [35], the interference from background could be removed to automatically alleviate the mapping between input images and crowd counts. We can find that most of them pay attention to the extraction of multi-scale features in the crowd. Because the scale variation of crowds restricts the performance of proposed methods.

In Table 9, it can be observed that all of the CNN-based methods have an absolute advantage over the traditional algorithms. In the CNN-based methods, our method achieves the best results on both Part\_A and Part\_B, which indicates that the accurate judgement of human head is important for improving the accuracy of crowd counting. We can also observe that Part\_A is more challenging than Part\_B. The crowd density is higher and the training is harder. As shown in Fig. 11 that is from Part\_A, the final estimated density map and its ground truth are very similar, except for the region in red rectangles where people are more difficult to recognize and these samples are more difficult to train. In the future, we plan to introduce the hard example mining technology to break through the limitation to further improve counting accuracy.

## 4.3. UCF\_CC\_50 dataset

Idrees et al. [12] collected 50 images from internet to produce the challenging UCF\_CC\_50 dataset. The number of annotated heads in each image ranges from 94 to 4,543. The total number of people in the dataset is 63,974. It is a challenging dataset because of its dense crowd, limited image, and low resolution. We generate the ground-truth density map by using the method in Sec. 3.2. The 5-fold cross validation is performed to evaluate proposed methods.

In Table 10, Other recent state-of-the-art methods are compared with our method. In [23, 12], traditional feature extractors were used to extract crowd features. In [38], the ConvLSTM employed the temporal correlation to assist crowd counting. Other methods on Shanghaitech dataset are still compared on this dataset. In Table 10, we can observe that the density map learned by CNN is more robust than the hand-crafted features extracted in [23, 12]. Our method outperforms all other methods on MAE metric, while we obtain a competitive MSE score which denotes the robustness

**Table 8**

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 [36] | 4.3        | 59.1       | 43.7       | 17.0       | 7.6        | 26.3       |
| MCNN [45]         | 3.4        | 20.6       | 12.9       | 13.0       | 8.1        | 11.6       |
| TDF-CNN [25]      | 2.7        | 23.4       | 10.7       | 17.6       | 3.3        | 11.5       |
| IG-CNN [1]        | 2.6        | 16.1       | 10.2       | 20.2       | 7.6        | 11.3       |
| Xiong. et al [38] | 6.8        | 14.5       | 14.9       | 13.5       | 3.1        | 10.6       |
| DDCN [35]         | 4.8        | 16.2       | 12.4       | 10.9       | 4.9        | 9.8        |
| Switch-CNN [26]   | 4.4        | 15.7       | 10.0       | 11.0       | 5.9        | 9.4        |
| DecideNet [18]    | 2.0        | 13.1       | <b>8.9</b> | 17.4       | 4.8        | 9.2        |
| D-ConvNet-V1 [28] | <b>1.9</b> | 12.1       | 20.7       | 8.3        | 2.6        | 9.1        |
| CP-CNN [30]       | 2.9        | 14.7       | 10.5       | 10.4       | 5.8        | 8.9        |
| MRA-CNN [44]      | 2.4        | 11.4       | 9.3        | 10.5       | 3.7        | 7.5        |
| ACSCP [27]        | 2.8        | 14.1       | 9.6        | <b>8.1</b> | 2.9        | 7.5        |
| CAT-CNN(OURS)     | 2.2        | <b>9.8</b> | 10.2       | 11.2       | <b>2.5</b> | <b>7.2</b> |

**Table 9**

Estimation errors on the ShanghaiTeach dataset.

| Method              | Part_A      |              | Part_B      |             |
|---------------------|-------------|--------------|-------------|-------------|
|                     | MAE         | MSE          | MAE         | MSE         |
| LBR + RR            | 303.2       | 371.0        | 59.1        | 81.7        |
| SE Cycle GAN [36]   | 123.4       | 193.4        | 19.9        | 28.3        |
| MCNN [45]           | 110.2       | 173.2        | 26.4        | 41.3        |
| Switch-CNN [26]     | 90.4        | 135.0        | 21.6        | 33.4        |
| TDF-CNN [25]        | 97.5        | 145.1        | 20.7        | 32.8        |
| DecideNet+R3 [18]   | -           | -            | 20.8        | 29.4        |
| ACSCP [27]          | 75.7        | 102.7        | 17.2        | 27.4        |
| MRA-CNN [44]        | 74.2        | 112.5        | 11.9        | 21.3        |
| CP-CNN [30]         | 73.6        | 106.4        | 20.1        | 30.1        |
| D-ConvNet-V1 [28]   | 73.5        | 112.3        | 18.7        | 26.0        |
| Hossain et al. [10] | -           | -            | 16.9        | 28.4        |
| IG-CNN [1]          | 72.5        | 118.2        | 13.6        | 21.2        |
| DDCN [35]           | 71.5        | 110.4        | 13.8        | 20.1        |
| CAT-CNN(OURS)       | <b>66.7</b> | <b>101.7</b> | <b>11.2</b> | <b>20.0</b> |

**Table 10**

Estimation errors on the UCF\_CC\_50 dataset.

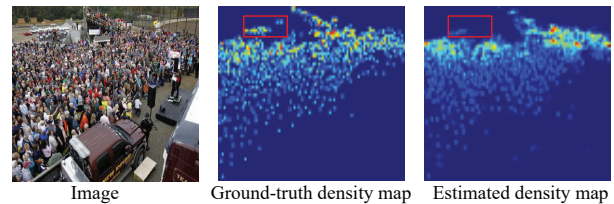
| Method                | MAE          | MSE          |
|-----------------------|--------------|--------------|
| Rodriguez et al. [23] | 655.7        | 697.8        |
| Lempitsky et al. [12] | 419.5        | 541.6        |
| MCNN [45]             | 377.6        | 509.1        |
| SE Cycle GAN [36]     | 373.4        | 528.8        |
| TDF-CNN [25]          | 354.7        | 491.4        |
| Switch-CNN [26]       | 318.1        | 439.2        |
| CP-CNN [30]           | 295.8        | 320.9        |
| IG-CNN [1]            | 291.4        | 349.4        |
| ACSCP [27]            | 291.0        | 404.6        |
| D-ConvNet-V1 [28]     | 288.4        | 404.7        |
| DDCN [35]             | 286.2        | 479.6        |
| Xiong. et al [38]     | 284.5        | <b>297.1</b> |
| Hossain et al. [10]   | 271.6        | 391.0        |
| MRA-CNN [44]          | 240.8        | 352.6        |
| CAT-CNN(OURS)         | <b>235.5</b> | 324.8        |

of proposed methods. We argue that the limited data for training cause this result as the UCF\_CC\_50 dataset only contains 50 images.

#### 4.4. WorldExpo'10 dataset

The WorldExpo'10 dataset [43] consists of 3,980 annotated frames collected from 1,132 video sequences. The video sequences contain 108 different scenes. This dataset is divided into a training set with 3,380 frames and a test set with 600 frames from 5 different scenes (S1-S5). This dataset provides the region of interest (ROI) and perspective map for each scene. For fair comparisons, we follow the experimental setting in [43] to generate density maps.

In Table 8, we compare our method with other recent state-of-the-art methods. MAE is used to evaluate the results of different methods, which is suggested in [43]. We can observe that our proposed method obtains the best results

**Figure 11:** Limitation of the proposed method

on the average MAE of five different scenes. It can be also observed that we get competitive results in the three of five scenes. By reviewing all test images, we find that due to the effect of ROI, the images in these three scenes are no longer complicated. There are fewer people and little background noise. Therefore, it is difficult to play our strengths.

## 5. Conclusion

In this paper, we proposed an end-to-end model named CAT-CNN in crowd counting. Our CAT-CNN can adaptively assess the importance of a human head at each pixel location to avoid enormous misjudgements. To obtain a robust confidence map for population distribution, the weighted BCELoss is proposed. The crowd counts are classified into five groups in each dataset and we first explicitly map the prior of the population-level category to feature maps to automatically contribute in encoding a highly refined density map. We evaluate our method on three benchmark datasets separately. Extensive experimental results indicate that our method outperforms many state-of-the-art methods. In the future, we will focus on hard example mining technology to handle the problem proposed in Fig. 11.

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

**Please cite @article{CHEN2020, title={Crowd counting with crowd attention convolutional neural network}, author={Chen, Jiwei and Wen, Su and Wang, Zengfu}, journal={Neurocomputing}, volume={382}, pages={210–220}, year={2020}, publisher={Elsevier} }**

## References

- [1] Babu Sam, D., Sajjan, N.N., Venkatesh Babu, R., Srinivasan, M., 2018. Divide and grow: Capturing huge diversity in crowd images with incrementally growing cnn, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3618–3626.
- [2] Boominathan, L., Kruthiventi, S.S., Babu, R.V., 2016. Crowdnet: A deep convolutional network for dense crowd counting, in: Proceedings of the 24th ACM international conference on Multimedia, ACM. pp. 640–644.
- [3] Chan, A.B., Liang, Z.S.J., Vasconcelos, N., 2008. Privacy preserving crowd monitoring: Counting people without people models or tracking, in: 2008 IEEE Conference on Computer Vision and Pattern Recognition, IEEE. pp. 1–7.
- [4] Chan, A.B., Vasconcelos, N., 2009. Bayesian poisson regression for crowd counting, in: Computer Vision, 2009 IEEE 12th International Conference on, IEEE. pp. 545–551.
- [5] Collobert, R., Weston, J., 2008. A unified architecture for natural language processing: Deep neural networks with multitask learning, in: Proceedings of the 25th international conference on Machine learning, ACM. pp. 160–167.
- [6] Dalal, N., Triggs, B., 2005. Histograms of oriented gradients for human detection, in: Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on, IEEE. pp. 886–893.
- [7] Glorot, X., Bordes, A., Bengio, Y., 2011. Deep sparse rectifier neural networks, in: Proceedings of the fourteenth international conference on artificial intelligence and statistics, pp. 315–323.
- [8] He, K., Zhang, X., Ren, S., Sun, J., 2015. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification, in: Proceedings of the IEEE international conference on computer vision, pp. 1026–1034.
- [9] He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778.
- [10] Hossain, M., Hosseinzadeh, M., Chanda, O., Wang, Y., 2019. Crowd counting using scale-aware attention networks, in: 2019 IEEE Winter Conference on Applications of Computer Vision (WACV), IEEE. pp. 1280–1288.
- [11] Hu, Y., Chang, H., Nian, F., Wang, Y., Li, T., 2016. Dense crowd counting from still images with convolutional neural networks. Journal of Visual Communication and Image Representation 38, 530–539.
- [12] Idrees, H., Saleemi, I., Seibert, C., Shah, M., 2013. Multi-source multi-scale counting in extremely dense crowd images, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2547–2554.
- [13] Li, M., Zhang, Z., Huang, K., Tan, T., 2008. Estimating the number of people in crowded scenes by mid based foreground segmentation and head-shoulder detection, in: 2008 19th International Conference on Pattern Recognition, IEEE. pp. 1–4.
- [14] Li, T., Chang, H., Wang, M., Ni, B., Hong, R., Yan, S., 2015. Crowded scene analysis: A survey. IEEE transactions on circuits and systems for video technology 25, 367–386.
- [15] Li, T., Wang, Y., Hong, R., Wang, M., Wu, X., 2018. pdisvpl: Probabilistic discriminative visual part learning for image classification. IEEE MultiMedia 25, 34–45.
- [16] Lin, M., Chen, Q., Yan, S., 2013. Network in network. arXiv preprint arXiv:1312.4400 .
- [17] Lin, Z., Davis, L.S., 2010. Shape-based human detection and segmentation via hierarchical part-template matching. IEEE Transactions on Pattern Analysis and Machine Intelligence 32, 604–618.
- [18] Liu, J., Gao, C., Meng, D., Hauptmann, A.G., 2018. Decidenet: Counting varying density crowds through attention guided detection and density estimation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5197–5206.
- [19] Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3431–3440.
- [20] Marana, A., Costa, L.d.F., Lotufo, R., Velastin, S., 1998. On the efficacy of texture analysis for crowd monitoring, in: Proceedings SIB-GRAP'98. International Symposium on Computer Graphics, Image Processing, and Vision (Cat. No. 98EX237), IEEE. pp. 354–361.
- [21] Marsde, M., McGuinness, K., Little, S., Keogh, C.E., O'Connor, N.E., 2018. People, penguins and petri dishes: adapting object counting models to new visual domains and object types without forgetting .
- [22] Ren, S., He, K., Girshick, R., Sun, J., 2015. Faster r-cnn: Towards real-time object detection with region proposal networks, in: Advances in neural information processing systems, pp. 91–99.
- [23] Rodriguez, M., Laptev, I., Sivic, J., Audibert, J.Y., 2011. Density-aware person detection and tracking in crowds, in: 2011 International Conference on Computer Vision, IEEE. pp. 2423–2430.
- [24] Ryan, D., Denman, S., Fookes, C., Sridharan, S., 2009. Crowd counting using multiple local features, in: 2009 Digital Image Computing: Techniques and Applications, IEEE. pp. 81–88.
- [25] Sam, D.B., Babu, R.V., 2018. Top-down feedback for crowd counting convolutional neural network, in: Thirty-Second AAAI Conference on Artificial Intelligence.
- [26] Sam, D.B., Surya, S., Babu, R.V., 2017. Switching convolutional neural network for crowd counting, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, p. 6.
- [27] Shen, Z., Xu, Y., Ni, B., Wang, M., Hu, J., Yang, X., 2018. Crowd counting via adversarial cross-scale consistency pursuit, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5245–5254.
- [28] Shi, Z., Zhang, L., Liu, Y., Cao, X., Ye, Y., Cheng, M.M., Zheng, G., 2018. Crowd counting with deep negative correlation learning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5382–5390.
- [29] Sindagi, V.A., Patel, V.M., 2017a. Cnn-based cascaded multi-task learning of high-level prior and density estimation for crowd counting, in: 2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), IEEE. pp. 1–6.
- [30] Sindagi, V.A., Patel, V.M., 2017b. Generating high-quality crowd density maps using contextual pyramid cnns, in: 2017 IEEE International Conference on Computer Vision (ICCV), IEEE. pp. 1879–1888.

- [31] Sindagi, V.A., Patel, V.M., 2018. A survey of recent advances in cnn-based single image crowd counting and density estimation. *Pattern Recognition Letters* 107, 3–16.
- [32] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A., 2015. Going deeper with convolutions, in: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1–9.
- [33] Viola, P., Jones, M.J., 2004. Robust real-time face detection. *International journal of computer vision* 57, 137–154.
- [34] Wang, J., Sun, Y., Zhang, W., Thomas, I., Duan, S., Shi, Y., 2016. Large-scale online multitask learning and decision making for flexible manufacturing. *IEEE Transactions on Industrial Informatics* 12, 2139–2147.
- [35] Wang, L., Yin, B., Tang, X., Li, Y., 2019a. Removing background interference for crowd counting via de-background detail convolutional network. *Neurocomputing* 332, 360–371.
- [36] Wang, Q., Gao, J., Lin, W., Yuan, Y., 2019b. Learning from synthetic data for crowd counting in the wild, in: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- [37] Wang, X., Han, T.X., Yan, S., 2009. An hog-lbp human detector with partial occlusion handling, in: *2009 IEEE 12th international conference on computer vision*, IEEE. pp. 32–39.
- [38] Xiong, F., Shi, X., Yeung, D.Y., 2017. Spatiotemporal modeling for crowd counting in videos, in: *2017 IEEE International Conference on Computer Vision (ICCV)*, IEEE. pp. 5161–5169.
- [39] Yang, B., Cao, J., Wang, N., Zhang, Y., Zou, L., 2018. Counting challenging crowds robustly using a multi-column multi-task convolutional neural network. *Signal Processing: Image Communication* 64, 118–129.
- [40] Yu, F., Koltun, V., 2016. Multi-scale context aggregation by dilated convolutions, in: *ICLR*.
- [41] Zeng, C., Ma, H., 2010. Robust head-shoulder detection by pca-based multilevel hog-lbp detector for people counting, in: *2010 20th International Conference on Pattern Recognition*, IEEE. pp. 2069–2072.
- [42] Zhan, B., Monekosso, D.N., Remagnino, P., Velastin, S.A., Xu, L.Q., 2008. Crowd analysis: a survey. *Machine Vision and Applications* 19, 345–357.
- [43] Zhang, C., Li, H., Wang, X., Yang, X., 2015. Cross-scene crowd counting via deep convolutional neural networks, in: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 833–841.
- [44] Zhang, Y., Zhou, C., Chang, F., Kot, A.C., 2019. Multi-resolution attention convolutional neural network for crowd counting. *Neurocomputing* 329, 144–152.
- [45] Zhang, Y., Zhou, D., Chen, S., Gao, S., Ma, Y., 2016. Single-image crowd counting via multi-column convolutional neural network, in: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 589–597.
- [46] Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A., 2016. Learning deep features for discriminative localization, in: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2921–2929.



**Jiwei Chen** is currently pursuing the Ph.D. degree in the Hefei Institutes of Physical Science, University of Science and Technology of China. His research interests include crowd counting, computer vision and machine learning.



**Wen Su** received Ph.D. degree in control science and engineering from University of Science and Technology of China in 2018 and B.E. degree in engineering from Automation Department, University of Science and Technology of China in 2013, respectively. Now, she works in virtual reality laboratory in Zhejiang Sci-Tech University. At present her research interests are image segmentation and depth scene understanding based on deep learning.



**Zengfu Wang** received the B.S. degree in electronic engineering from the University of Science and Technology of China in 1982 and the Ph.D. degree in control engineering from Osaka University, Japan, in 1992. He is currently a Professor with the Institute of Intelligent Machines, Chinese Academy of Sciences, and the Department of Automation, University of Science and Technology of China. He has published more than 300 journal articles and conference papers. His research interests include computer vision, human–computer interaction and intelligent robots. He received the Best Paper Award at the ACM International Conference on Multimedia 2009 and the IET Image Processing Premium Award 2017.