Communications in Information and Systems
Volume 17, Number 1, 1–24, 2017

Counting congested crowds under wild conditions with a multi-task Inception network

Biao Yang*, Jinmeng Cao, Nan Wang,
Yuyu Zhang, and Ling Zou

Counting of congested crowds has been widely applied in surveillance event detection, public safety control, and traffic monitoring. Early studies mainly focus on designing hand-crafted features. However, counting performance based on hand-crafted features maybe easily influenced by issues such as partial occlusion, scale variation, and illumination change. Convolutional neural network (CNN) has shown great success in visual crowd counting. In this work, a multi-task Inception network is proposed for crowd counting in congested crowds. Crowd count is predicted through summing up a density map estimated by the proposed network. Three Inception blocks are employed to automatically extract multi-scale features from different patches cropped from the crowd image. The network can jointly estimate the density map, crowd density level, and background / foreground separation. Counting performance obtained through multi-task learning is superior to that obtained through only estimating density map. Contrastive evaluations based on three benchmarking datasets are implemented with several state-of-the-art CNN-based crowd counting approaches. Results indicate the accuracy and robustness of our network in counting congested crowds. The multi-task Inception network almost outperforms the state-of-the-art counting approaches in terms of mean absolute error and mean squared error.

*This work has been supported by the National Natural Science Foundation of China under Grant No. 61501060 and No. 61703381, the Natural Science Foundation of Jiangsu Province under Grant No. BK20150271, Key Laboratory for New Technology Application of Road Conveyance of Jiangsu Province under Grant BM20082061708.

1. Introduction

Crowd counting plays a key role in many applications, *e.g.* crowd control, pedestrian behavior profiling, and crowd anomaly detection which is useful for safety control. Crowd disasters can be nipped in the bud by giving an alarm when the number of people flocking to certain areas exceeds a certain crowd level. Meanwhile, crowd counting can be extended to other areas, such as counting cells from microscopic images. However, it is difficult for people to quickly count the crowds, especially those congested ones. Therefore, vision based crowd counting approaches have elicited increased attention from researchers in recent years.

Vision based crowd counting approaches mainly fall into three categories, namely, counting by detection, clustering, and regression. In counting by detection, crowd counting is achieved through detecting instances of pedestrians in a scene [1]. However, detecting entire body is always time consuming due to the exhaustive scanning of an image space and the result is always inaccurate due to partial occlusions. To reduce computational consumption, some researchers tried to detect only noteworthy parts. For example,Gao *et al.* detected heads using a water filling algorithm and achieved crowd counting through computing the number of detected heads [2]. Luo *et al.* built a head-shoulder model for depicting moving and stationary crowds [3]. The number of people was estimated through clustering the head-shoulder instances. Compared with methods that detect entire bodies, crowd counting based on detecting local parts are more robust to partial occlusions. However, detecting local parts are still time consuming and will be easily affected by a cluttered background.

In counting by clustering [4], a crowd is assumed to be composed of individual entities, each of which possessing unique yet coherent motion patterns can be clustered to estimate the number of pedestrians in a scene. For example, Rao *et al.* proposed an approach to estimate crowd density using motion cues and hierarchical clustering [5]. Counting by clustering is easy to implement because it needs little priori information, such as object detectors or hand-crafted features. However, motion cues are generally extracted using dense optical flow, which is time consuming to calculate.

Unlike counting by detection and clustering, counting by regression aims to achieve direct mapping between specific features and crowd counting without detecting individuals or clustering motion patterns in the crowd [6]. Thus, it is suitable to count the crowds under cluttered background with reasonable time consumption. Many works focused on extracting hand-crafted features and feature editing. They first extracted foreground areas, shapes, edges, and other features from detected crowds [7]. Then, support vector machine [8], random forest [9], extreme learning machine [10], and Gaussian process regression [11] were used to predict crowd counts. For instance, Fradi *et al.* used foreground pixel counts and corner density for crowd analysis [11], while number of people was estimated through Gaussian process regression. Mahdi *et al.* utilized a combination of key point (corners) and segment-based features to to estimate crowd count [12]. Liang *et al.* also employed key points (speeded up robust feature points) as cues for crowd counting [13]. Aside from commonly used features, novel features which are highly suitable for crowd counting have been proposed. Shafiee *et al.* proposed a novel low-complexity, scale-normalized feature called histogram of moving gradients (HoMG) [14]. HoMG is a highly effective spatiotemporal representation of individuals and crowds within a video. Zhang *et al.* proposed a flow field texture representation approach to depict segmented crowds [15]. Chen *et al.* introduced a novel cumulative attribute concept for learning a regression model when only sparse and imbalanced data were available [16]. These features can achieve a satisfactory performance when the crowds are simple and sparse. However, they may fail if the crowds are heavily occluded or very dense. Mousse *et al.* extracted the convex hull from detected foreground pixels, and crowd counting was realized by fusing the obtained polygons with geometric properties [17]. This method is robust to partial occlusion, however, it can only be used in multi-camera networks with overlapping views.

Moreover, in spite of the rapid development of crowd counting approaches, many longstanding challenges are not well solved, *e.g.* partial occlusion, perspective distortion, background clutter, illumination change, scale and appearance variations. As illustrated in Fig.1, all crowd images (selected from benchmarking datasets, *e.g.* UCF_CC_50 [18], WorldExpo'10 [19] and Shanghai Tech [20] datasets) suffer from severe occlusion, scale variation, non-uniform density distribution and so on. Furthermore, scale variations in crowds with relatively low or medium density levels (the 2^{nd} and 3^{rd} examples) are more drastic than that in crowds with high density levels (the 1^{st} and 4^{th} examples), which are always captured in a bird's-eye view. That is to say, drastic scale variations are always observed in crowds that are not extremely dense. Recently, deep learning based crowd counting approaches have achieved great success while counting the crowds that encountered with cluttered background, non-uniform illumination, and varying appearances [19]–[31]. However, scale variation and non-uniform density

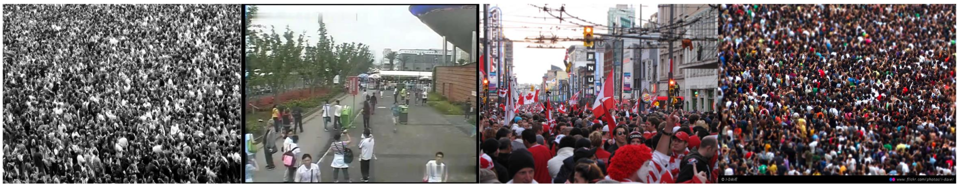

Figure 1: Examples of crowd images extracted from three benchmarking datasets. These crowd images suffer from severe occlusion, scale variations, non-uniform density distribution, perspective distortion, and so on.

distribution still prevent the state-of-the-art counting approaches from predicting precise crowd counts. Thus, in this work, we mainly focus on the abovementioned two issues.

The pipeline of the proposed counting approach is shown in Fig.2. The entire process can be divided into the training and the evaluation stages. In the training stage, the input image is uniformly segmented into 16 non-overlapped patches. For each patch, its ground truths of density map, crowd density level, and background/foreground (BG/FG) separation are calculated. Then, all the patches are delivered to the Inception network which is comprised of Inception blocks, pooling layers, and deconvolutions. This network is used to extract features of different scales which will be further transferred to the multi-task network. Based on the automatically extracted features and pre-computed ground truth, parameters of the network can be learned through minimizing a joint loss function, which is comprised of three sub-losses (losses of estimating density map, crowd density level, and BG/FG separation) in a weighted manner.

In the evaluation stage, the input image is segmented into overlapped patches using a given stride (stride is set to be 20 pixels in this work). The patch can be of arbitrary size. All patches are then imported to the well-learned Inception and multi-task networks in sequence. Among the multiple outputs of Multi-task network, only density maps of all patches are used to reconstruct the density map of the entire crowd image. Estimating density levels and BG/FG separations are used to refine the estimating results of density maps. The crowd count is then predicted through summing up all values of the reconstructed density map. Notably, several locations of the entire density map are repeatedly accumulated. Thus, the values of these locations should be normalized by the cumulative frequency in such locations. Extracting patches in different manners during training and evaluation stages are inspired by the work of Sindagi *et al.* [21]. They argued that extracting non-overlapped patches in the training stage is superior to extract overlapped patches because too much redundancy existing in the overlapped patches. The redundancy may lead to over-fitting and a poor generalization capability.

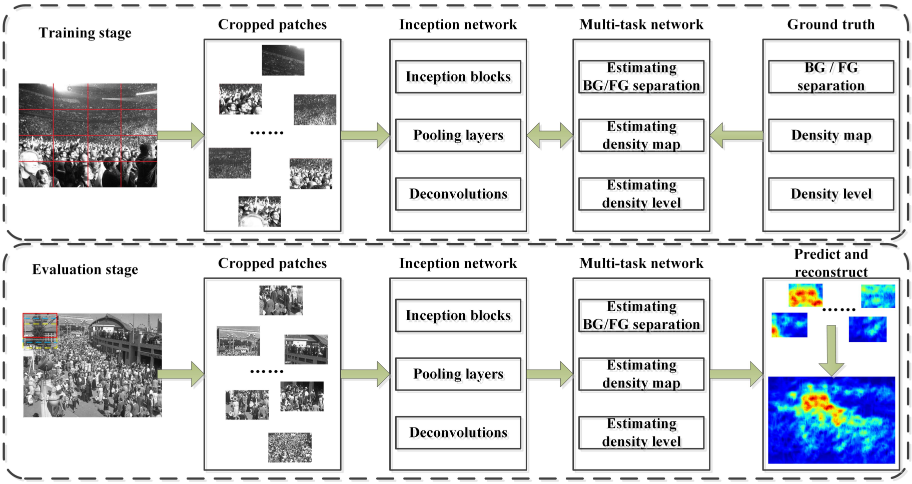

Figure 2: Pipeline of the proposed counting approach. Upper part is the training stage and lower part is the evaluation stage.

This work provides two novel contributions. First, three Inception blocks are used to extract multi-scale features from the crowd patches. Unlike existed multi-column CNN (MCNN) [20] and switching CNN [22], the Inception block is more suitable to resolve drastic scale variations. Deconvolutions are applied to compensate for the loss in detail caused by early pooling layers. Second, our model can jointly estimate the density map, crowd density level, and BG/FG separation. The multi-task learning strategy can improve the accuracy of the reconstructed density map.

The rest of this work is organized as follows. Section 2 provides a review of related work on crowd counting. Section 3 gives the details of the proposed approach. Section 4 gives the experimental evaluations and analysis. The conclusions are presented in Section 5.

2. Related work

Serious limitations exist in crowd counting using hand-crafted features, especially in dense crowds with heavy occlusions. CNN has achieved great success in pattern recognition fields, such as object detection [23] and semantic segmentation [24]. According to a recent review on crowd counting [25], CNN-based counting approaches are superior to traditional ones which mainly based on hand-crafted features. Thus, we mainly reviewed related work on CNN-based counting approaches. For crowd counting, CNNs have been trained alternatively with two related learning objectives, namely, crowd density and crowd count [19]. Wang *et al.* proposed an end-to-end CNN regression model based on the AlexNet network for counting people in highly dense crowds [26]. Inspired by their success, Walach *et al.* performed layered boosting and selective sampling [27]. To remove fully connected layers and make the model highly compact, Marsden *et al.* proposed a fully convolutional crowd counting approach to predict the number of people in highly congested scenes by estimating the density map [28]. Fu *et al.* classified a crowd image into five density levels, which are very high, high, medium, low, and very low density, instead of directly estimating density maps [29].

Despite significant improvements in CNN-based crowd counting, other challenges, such as scale variations and non-uniform density distribution, remain. Zhang *et al.* aimed to resolve scale variations and built a simple but effective MCNN to estimate crowd count from a crowd image [20]. Inspired by their work, Onoro *et al.* developed a scale-aware counting approach called Hydra-CNN that can estimate the density map in different scenarios without any explicit geometric information on the scene [30]. Sam *et al.* argued that improved performance can be obtained by choosing a given CNN column with the aid of a pre-trained switch classifier [22].But these approaches are more suitable for counting crowds with mild scale variation. Aside from scale variations, non-uniform density distribution is another problem that affects counting performance. A multi-task strategy can be used to handle this issue. Zhang *et al.* simultaneously estimated the density map and crowd count [19]. Sindagi *et al.* proposed a cascaded multi-task CNN (Cascaded-MTL) that jointly estimates the density level and density map [21]. The counting performance was improved by replacing the crowd count with the density level. Other tasks can also be performed during crowd counting by employing a multi-task strategy. For instance, Marsden *et al.* proposed a Resnet Crowd model for crowd counting [31]. The model can simultaneously predict the crowd count, density map, count class, and even abnormal events. Initializing network parameters with a well-trained Resnet model leads to enhanced counting performance when the training instances are insufficient.However, all these methods cannot effectively resolve non-uniform density distribution.

Despite the state-of-the-art CNN-based counting approaches can resolve problems such as partial occlusion, perspective distortion, illumination change, they still encounter with two challenges, namely, drastic scale variation and non-uniform density distribution. Therefore, some efforts will be made to handle these two challenges through improving existed CNN-based counting approaches.

3. Proposed method

3.1. Generating ground truths for input patches

The main objective of the proposed neural network is to learn mapping F: $X \rightarrow D$, where X is a set of features automatically extracted from training patches and D is a set of density maps of these patches. For each patch, the density map is generated based on the labeled locations of people in the crowds, as well as the perspective images of different scenes. Notably, perspective images are pre-calculated using the approach proposed by Chan *et al.* [7].

Many studies followed [32] and defined the density map as a sum of Gaussian kernels (Fig.3(a)) centered on object locations. This type of density map is proposed to characterize the density distribution of circle-like objects, such as cells and bacteria. Later, Zhang *et al.* found that a human-shaped kernel (Fig.3(b)) maybe more suitable for characterizing pedestrians in the crowds [19]. They argued that the shapes of pedestrians are more similar to ellipses than to circles. However, in real applications, only head parts can be reliably observed in congested crowds. Thus, the density map used in this work was generated using the Gaussian kernel. After obtaining the head position of pedestrian P_a in the patch, its density map is generated as

$$D_i(p) = \sum_{p \in P_i} \frac{1}{\|Z_i\|} N_g(p|P_a, \sigma_a) \tag{1}$$

where p is an aribitary position in the i^{th} patch P_i and N_g is a normalized 2D Gaussian kernel with variance σ_a (setting of σ_a can refer to [19]). To ensure that the integration of all values in the density map equals the total number of pedestrians in that patch, the entire distribution is normalized by Z_i, which is the actual count of the crowd in that patch.

Notably, the generated density map is used as one of the ground truths for training the multi-task Inception network. The generated density map is used as one of the ground truths for training the multi-task Inception network. The BG/FG separation was calculated by thresholding the density map with a given threshold (in this paper, a small threshold of 0.0001 is used

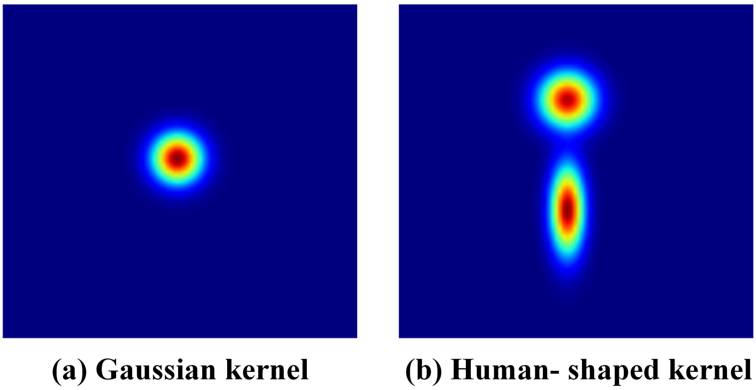

Figure 3: Two commonly used kernels for generating density maps: (a) Gaussian kernel and (b) human-shaped kernel. The red regions indicate strong activations, and the blue regions indicate weak activations (best viewed in color).

to guarantee that all foreground information is covered). The crowd density level is determined by the actual number of people in a particular patch according to the rule proposed by Fu *et al.* [29]. Both BG/FG separation and crowd density level are used as ground truths and applied in the training stage.

3.2. Multi-task Inception network

Partial occlusion, perspective distortion, illumination change can be well handled by state-of-the-art CNN-based crowd counting approaches. The purpose of this work is to study another two challenging problems, namely, non-uniform density distribution and drastic scale variations.

Inspired by the success of MCNN proposed by Zhang *et al.* [20], many researchers tried to extract features of different scales using multiple CNN columns. However, only high level features of different scales can be fused, whereas some useful low level features of different scales cannot be effectively fused in MCNN. In order to make full use of both low and high level features of different scales, an Inception network (containing three Inception blocks [33]) are used to automatically extract features. Compared with MCNN, the Inception network can fuse features of different scales at different depths. Thus, it is more likely to extract multi-scale features which are useful for crowd counting.

Another contribution of this work is the employment of multi-task learning, especially the introduction of BG / FG separation. Existed CNN-based crowd counting approaches are always shallow network and cannot estimate the density map that is totally consistent with the ground truth. Local optima is always obtained when training the neural network. Meanwhile, it is hard to converge when estimating density map with a very deep neural network. Multi-task learning strategy is beneficial to estimate a more accurate density map with a shallow CNN through jointly estimating other correlated objectives. Precisely, we follow the definition proposed by Fu *et al.* [29] that classifies crowds into five density levels: very high density, high density, medium density, low density, and very low density.

Jointly estimating the density map and the density level can make the intensity of density map more close to the ground truth. However, a portion of the background in density map is always recognized as the foreground or vice versa. To resolve this problem, a BG/FG separation is added into the multi-task learning strategy particularly. The BG/FG separation is similar to the density map. However, it only focuses on the distribution of crowds, whereas the density map concentrates on both distribution and intensity information. Thus, the BG/FG separation owns stronger distribution information than the density map.

On the basis of the abovementioned discussions, a multi-task Inception network was proposed for crowd counting (Fig.4). The legend on the bottom-left corner of Fig.4 illustrates the colors of different convolutional filters and their corresponding kernel sizes. The whole network is divided into an Inception network and a multi-task network. Inception network, which is used to extract features, is comprised of three Inception blocks, two max pooling layers and two deconvolution layers. Max pooling is used to reduce the feature dimension, thus reduce the numbers of convolutions. Deconvolutions are used after Inception block3 to compensate for the loss of detail due to early pooling layers. Inner structure of Inception block is illustrated in the top-right corner, represented by a blue rectangle. Take Inception block2 as an example, four convolutional layers (Conv2_1, 2_5, 2_6 and 2_7) of different kernel sizes (1×1, 3×3, 5×5, 7×7) are used to extract features of different scales. Three 1×1 convolutional layers (Conv2_2, 2_3, and 2_4) are used to reduce features because filters with large kernel sizes need more time for convolutions. Finally, outputs of Conv2_1, 2_5, 2_6, and 2_7 are concatenated as output_block2. Features of different scales can be fused at different depths through introducing Inception blocks, leading to capture more effective features for crowd counting than MCNN. Different from original Inception structure, we replaced the 3×3 max pooling layer with a 7×7

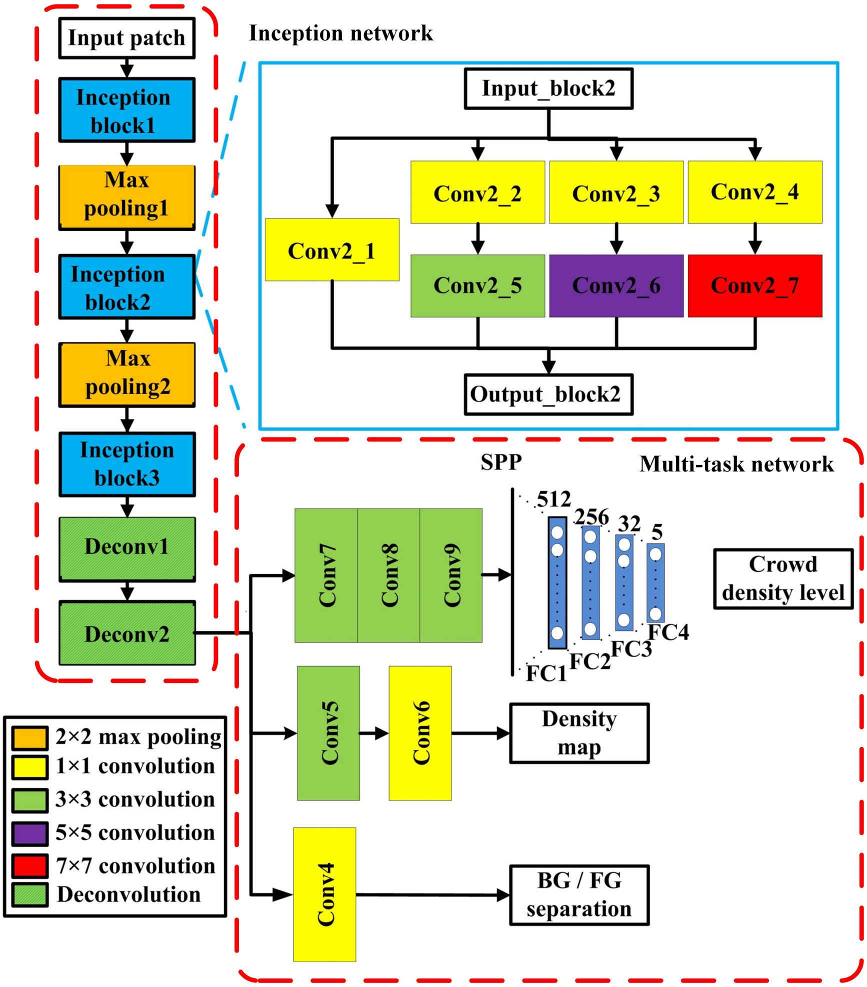

Figure 4: Structure of the multi-task Inception network that contains an Inception network and a multi-task network. Dropout, PReLU, and LRN are not listed for simplification (best viewed in color).

convolutional layer. The reason is that each Inception block is used to extract features of different scales, while keeping the feature size (max pooling may shrink the feature size).

The multi-task network is inspired by Sindagi *et al.* except for the employment of BG / FG separation. To estimate crowd density level that is more abstract than other two objectives, three convolutional layers (Conv7, 8, 9) are utilized to further process the output features of Inception network. Then, a spatial pyramid pooling (SPP) [34] of height three is used to eliminate the fixed size constraint of deep networks which contain fully connected layers. Fixed size outputs of SPP are fed to four fully connected layers, namely, FC1 (512 neurons), FC2 (256 neurons), FC3 (32 neurons), and FC4 (5 neurons) to estimate the density level. At the same time, features obtained by Inception network are directly fed to a 1×1 convolutional layer (Conv4) for estimating BG/FG separation. To estimate the density map that contains more information than BG/FG separation, a deeper neural network is wanted. Thus, a 3×3 convolutional layer (Conv5) is used to process the features before they are imported into Conv6. Each convolutional layer (except for conv4 and 6) is followed by a dropout layer (with parameter 0.3), a parametric rectified linear unit (PReLU) activation function, and a local response normalization (LRN) layer. Details of network parameters are given in Table 1. Strides of all convolutional filters are set to 1.

	Number of filters	Padding		Number of filters	Padding
Conv1_1	36	0	Conv3_1	64	0
Conv1_2	36	0	Conv3_2	64	0
Conv1_3	36	0	Conv3_3	64	0
Conv1_4	36	0	Conv3_4	64	0
Conv1_5	28	1	Conv3_5	48	1
Conv1_6	22	2	Conv3_6	44	2
Conv1_7	16	3	Conv3_7	40	3
Conv2_1	48	0	Conv4	1	0
Conv2_2	48	0	Conv5	16	1
Conv2_3	48	0	Conv6	1	0
Conv2_4	48	0	Conv7	64	0
Conv2_5	36	1	Conv8	128	0
Conv2_6	32	2	Conv9	256	0
Conv2_7	28	3	N/A	N/A	N/A

Table 1: Network parameters used in the multi-task Inception network. Kernel sizes of different convolutional filters are illustrated in Fig.4 and we will not list them in the table for simplification. Strides of all convolutional filters are set to 1.

3.3. Multi-task learning strategy

The main objective of the proposed network is to estimate the density maps of different patches, which are further used to reconstruct the density map of the crowd image. The loss between the estimated density map and its ground truth is defined as $L_{density}$, which can be calculated using Euclidean loss. $L_{density}$ is defined as follows:

$$L_{density} = \frac{1}{2N} \sum_{i=1}^{N} \|F_d(P_i, O) - D(P_i)\|_2 \tag{2}$$

where N is the number of patches, O is a set of network parameters, P_i is the i^{th} patch, $F_d(P_i, O)$ is the estimated density map of P_i, and $D(P_i)$ is the ground truth of $F_d(P_i, O)$.

The crowd density level and BG/FG separation were simultaneously estimated to resolve the non-uniform density distribution. The loss between the estimated density level and its ground truth is defined as L_{level}, which can be calculated using the cross-entropy loss. L_{level} is defined as follows:

$$L_{level} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{M} [(Y(P_i) = j) F_c(P_i, O)] \tag{3}$$

where M is the number of density levels (5 in our work), $F_c(P_i, O)$ is the estimated density level of P_i, and $Y(P_i)$ is its ground truth.

The BG/FG separation was used to prevent the network from mistaking the background as the foreground. The loss between the estimated BG/FG separation and its ground truth is defined as L_{mask}. L_{mask} can be also calculated using Euclidean loss as follows:

$$L_{mask} = \frac{1}{2N} \sum_{i=1}^{N} \|F_m(P_i, O) - M(P_i)\|_2 \tag{4}$$

where $F_m(P_i, O)$ is the estimated BG/FG separation of P_i and $M(P_i)$ is its ground truth.

The above mentioned three losses are jointly minimized in a weighted manner. The total loss function can then be defined as

$$L_{total} = \lambda_1 L_{density} + \lambda_2 L_{level} + \lambda_3 L_{mask} \tag{5}$$

where λ_1, λ_2, and λ_3 are the weights of different loss functions. Through estimating different objectives respectively, we found that L_{level} is almost two order greater in magnitude than $L_{density}$, while L_{mask} is the same order of magnitude as $L_{density}$.Thus,to highlight our main purpose (estimating density map), we set λ_1=1, λ_2=0.001, and λ_3=0.1. Effectiveness of three weights are verified through cross validation.

4. Experimental analysis

4.1. Implementing details and evaluation metrics

Training and evaluation of the multi-task Inception network are performed on NVIDIA GTX 1080 GPU (8G). Table 2 lists the parameters used to train the network. Batch size is set to 16 during training due to memory limitations. MSRA [35] is used to initialize each convolutional layer.

Parameters	Value
Base Learning rate	0.001
Learning policy	“inv”
Power	0.75
Gamma	0.001
Max iterations	60000
Momentum	0.9
Weight decay	0.005
Optimization Type	Adam

Table 2: Parameters used to train the multi-task Inception network.

Mean absolute error (MAE) and mean squared error (MSE) are used to evaluate different counting approaches. These two indicators are defined as follows:

$$MAE = \frac{1}{W}\sum_{i=1}^{W}|E(i) - G(i)| \tag{6}$$

$$MSE = \sqrt{\frac{1}{W}\sum_{i=1}^{W}(E(i) - G(i))^2} \tag{7}$$

where W is the total number of test frames, $G(i)$ is the actual crowd count in the i^{th} frame, and $E(i)$ is the predicted count in the i^{th} frame. In general, MAE and MSE indicate the accuracy and robustness of the model, respectively.

4.2. Datasets and evaluation settings

1) UCF_CC_50: UCF_CC_50 is a challenging dataset because it contains a wide range of densities and diverse scenes with non-uniform illumination conditions and perspective distortion. Totally 63,075 individuals are labeled in the entire dataset, with an average of 1,280 individuals per image. The number of individuals varies from 94 to 4,543, indicating a large variation across the crowd images. In this dataset, we employ five-cross validation to evaluate different crowd counting approaches.

2) WorldExpo'10: This large-scale dataset contains 1,132 annotated video sequences captured by 108 cameras from the Shanghai 2010 World Expo event. It consists of 3,980 frames, with 199,923 labeled pedestrians. The number of people varies from 1 to 253, with an average of 50 people per image. The dataset is split into training and testing sets. Training set contains 1,126 one-minute video sequences from 103 scenes, while testing set contains five one-hour video sequences from five scenes. Each test scene comprises 120 labeled frames, with the number of people varying from 1 to 220. The training-evaluation process is repeated five times and mean values of MAE / MSE are used for fairness.

3) Shanghai Tech: This dataset consists of 1,198 images with 330,165 annotated heads. This dataset is split into Part A and B. Part A contains 482 images randomly selected from the Internet, whereas Part B contains images captured from the streets of Shanghai. Training and testing sets of Part A involved 300 and 182 images, respectively, and the number of people varied from 33 to 3,139, with an average of 501 people per image. For Part B, 400 and 316 images were used in the training and testing sets, respectively, and the number of people varied from 9 to 578, with an average of 123 people per image. For each part, the procedure of training-evaluation is also repeated five times and mean values of MAE / MSE are used for fairness.

4.3. Evaluations of the multi-task Inception network

The multi-task strategy is a key contribution of this work. The density map, BG/FG separation, and crowd density level are jointly estimated, whereas most similar works only focus on the first one. To evaluate the proposed multi-task strategy, different combinations of objectives, including estimating density map, estimating density map and density level, estimating density map and BG/FG separation, as well as our approach are tested using benchmarking datasets. MAE and MSE are used to evaluate different approaches, and the results are presented in Table 3.

	Estimating density map		Estimating density map and density level		Estimating density map and BG/FG separation		Our multi-task strategy	
	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE
UCF_CC_50	383.1	502.9	377.5	511.6	379.5	492.6	**322.1**	**333.6**
World Expo '10	11.1	22.5	9.8	20.6	10.1	19.7	**8.9**	**19.5**
Shanghai Tech Part A	113.7	186.2	103.6	154.1	105.3	161.3	**91.6**	**133.2**
Shanghai Tech Part B	26.3	41.5	21.3	31.7	21.4	30.9	**18.2**	**28.7**

Table 3: Evaluations of the multi-task strategy in benchmarking datasets.

As shown in Table 3, counting performance through estimating multiple objectives obviously outperform counting performance through estimating density map. The reason has been discussed in section 3.2. Among different combinations, our approach is superior to others in both MAE and MSE. The reason of the superiority is that our approach focus on both distribution and intensity information, whereas others only focus on one kind of information. Aside from our multi-task strategy, the employment of density level achieves weak advantage over the employment of BG/FG separation in refining the estimated density maps. It reveals that intensity information maybe more important to distribution information in improving counting performance.

Several examples selected from the UCF_CC_50 dataset are shown in Fig.5. The first row represents original crowd images, the second row represents the ground truths of density maps, and the third row represents the estimated density maps. Ground truth (G) of crowd counts and corresponding estimated counts (E) are listed as follows: (a) G: 1046 E: 1085 (b) G: 3406 E: 1977 (c) G: 581 E: 437 (d) G: 440 E: 521. In general, the estimated density maps are approximately close to their ground truth, in both distribution and intensity. This indicates the proposed method could predict the counts of congested crowds, even under drastic scale variation and non-uniform density distribution. However, there are obvious large biases in Fig.5 (b) and (c). Through comparing the estimated density maps and their ground truths, we can find that people in some particular regions are really hard to capture by the network. Yellow rectangles on the 2^{nd} and 3^{rd} original images highlight such regions where people appeared blurry and hard to observe. Other technologies, such as super-resolution [36], may handle such a problem. But we not include such algorithms in the present work.

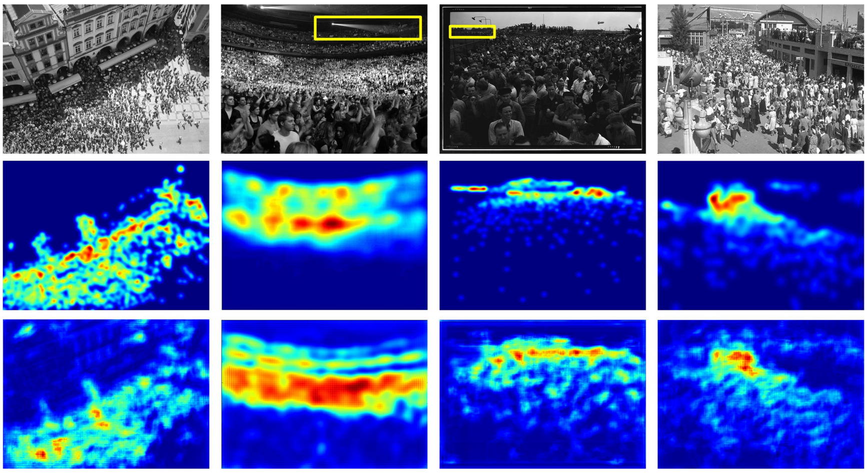

Figure 5: Some estimated density maps selected from the UCF_CC_50 dataset. The first row represents original crowd images, the second row represents the ground truths of density maps, and the third row represents the estimated density maps. In each density map, red regions indicate the existence of dense crowds while blue regions indicate that few people existed. Yellow rectangles on the original images of (b) and (c) indicate the regions where people appeared blurry and hard to observe (best viewed in color).

4.4. Comparisons with state-of-the-art approaches

Comparisons with state-of-the-art counting approaches are performed in three benchmarking datasets. Both MAE and MSE are used for evaluation. Table 4 illustrates the comparison results in UCF_CC_50 dataset, which owns an extremely high density level. This dataset is commonly used for evaluating different counting approaches, and we only selected some recent studies, *e.g.* cross-scene counting approach proposed by Zhang *et al.* [19], MCNN proposed by Zhang *et al.* [20], Hydra-CNN proposed by Onoro *et al.* [30], CNN-pixel counting proposed by Kang *et al.* [37], cascaded-MTL proposed by Sindagi *et al.* [21], and switching CNN proposed by Sam *et al.* [22], for comparison. Both MAE and MSE of these approaches are mentioned in their original works. Among the abovementioned approaches, cross-scene counting approach is the earliest approach that tried to count crowds using CNN. However, it focuses on how to select suitable scenes for fine-tuning an existing counting model and pays less attention on how to improve the counting performance on congested crowds. Thus, it has the highest MAE and the next-highest MSE. CNN-pixel counting utilizes CNN-pixel and FCNN-skip networks to directly estimate the entire density map of the crowd image. Thus, it could capture better detailed information than other approaches that employed max-pooling to reduce feature dimension. In this aspect, deconvolutions used in our approach could achieve similar performance as CNN-pixel counting in capturing detailed information. Furthermore, our approach tries to handle drastic scale variation and non-uniform density distribution through introducing Inception blocks and multi-task strategy. As a result, both MAE and MSE of our approach are much lower than those of CNN-pixel counting. Other approaches used for comparison focus on either scale or distribution variations. For instance, MCNN extracts features from the crowd image at different scales. Hydra-CNN is similar to MCNN, but the former also down-sample the input into different scales to better resolve scale variation. Different from them, switching CNN utilizes a pre-calculated classifier to indicate which CNN column should be used for estimating the density map. Thus, it could choose the best CNN column to extract features from crowds with certain density levels. However, using only one CNN column cannot cope with drastic scale variation. As a result, switching CNN can accurately estimate the counts of several extremely dense patches that take a large portion of the total counts, but it may perform unsatisfactory on patches with scale variation, thus resulting in an unreliable counting performance. As shown in the table, switching CNN has the lowest MAE but its MSE is much higher than our MSE. Aside from counting approaches designed for addressing scale variation, cascaded-MTL predicts the crowd counts in a multi-task manner, which jointly estimates the density map and the crowd density level. These two objectives are complementary and the final counting performance is improved. As shown in the table, MAE of cascaded-MTL is similar to our MAE. But MSE of cascaded-MTL is a little higher than our MSE. In general, our approach performs well in both MAE and MSE while comparing with several state-of-the-art CNN based approaches. This finding reveals the robustness of our approach when counting extremely dense crowds, such as the crowds in UCF_CC_50 dataset.

Table 5 illustrates the comparison results in World Expo '10 dataset, which contains crowds of a medium density level. Cross-scene counting approach, MCNN, CNN-pixel counting, and switching CNN are used for comparison based on their reported results. Notably, only MAEs are reported in

	MAE	MSE
Cross-scene counting [19]	467.0	498.5
MCNN [20]	377.6	509.1
Hydra-CNN [30]	333.7	425.2
CNN-pixel counting [37]	406.2	404.0
Cascade-MTL [21]	322.8	341.4
Switching CNN [22]	**318.1**	439.2
Our approach	322.1	**333.6**

Table 4: Comparisons with the state-of-the-arts in UCF_CC_50 dataset.

original works, therefore, only MAEs are reported in original works, therefore, only MAEs are listed in Table 5. As shown in the table, cross-scene counting and CNN-pixel counting approaches perform the worst due to the same reasons as we discussed in analyzing Table 4. MCNN performs better than the abovementioned two approaches due to its focus on handling scale variation. Both switching CNN and our approach achieved satisfactory performance. Our approach is a little better than switching CNN due to our advantage in handling drastic scale variation that is commonly existed in crowds of a medium density level.

	MAE
Cross-scene counting [19]	12.9
MCNN [20]	11.6
CNN-pixel counting [37]	13.4
Switching CNN [22]	9.4
Our approach	**8.9**

Table 5: Comparisons with the state-of-the-arts in WorldExpo'10 dataset.

Table 6 gives the comparison results in Shanghai Tech dataset, which has crowds of both medium (Part B) and high (Part A) density levels. The cross-scene counting approach, MCNN, FCN proposed by Marsden *et al.* [28], cascaded-MTL, and switching CNN are employed for comparison. Marsden *et al.* counted the crowds with a FCN structure. It outperforms cross-scene counting approach through replacing fully connections with fully convolutions. MCNN performs better than FCN in Part A due to its ability in extracting multi-scale features. However, FCN performs better than MCNN in Part B due to its deeper structure compared with MCNN. It reveals that for crowds of medium density level, a deep neural network maybe more effective than a wide neural network. Cascaded-MTL, switching CNN, and our approach perform much better than the abovementioned three approaches. Similar to the results of Table 4, switching CNN outperforms our approach slightly in terms of MAE on Part A, which has extremely dense crowds. Our approach outperforms switching CNN in terms of MSE in the same dataset. For Part B that has crowds of medium density level, our approach achieves the best performance in both terms of MAE and MSE.

	Part A		Part B	
	MAE	MSE	MAE	MSE
Cross-scene counting [19]	181.8	277.7	32.0	49.8
MCNN [20]	110.2	173.2	26.4	41.3
FCN [28]	126.5	173.5	23.76	33.12
Cascade-MTL [21]	101.3	152.4	20.0	31.1
Switching CNN [22]	**90.4**	135.0	21.6	33.4
Our approach	91.6	**133.2**	**18.2**	**28.7**

Table 6: Comparisons with the state-of-the-arts in Shanghai Tech dataset.

Comparisons with state-of-the-art CNN-based approaches in different benchmarking datasets show that our approach almost outperforms the others in terms of MAE and MSE when counting crowds with a low or medium level. However, switching CNN has a slight advantage over our approach in terms of MAE while counting extremely dense crowds (*e.g.*,UCF_CC_50 and Shanghai Tech Part A). The reason is that switching CNN can select a most suitable CNN column for counting extremely dense crowds that take a large portion of the global counts, whereas our approach focuses more on drastic scale variation of the crowds. Notably, both counting approaches are based on patches. Thus, switching CNN may achieve a low MAE (perform well on some extremely dense patches) while counting congested crowds, whereas our approach (perform slightly worse on extremely dense patches but balanced on all patches) may achieve a low MSE that reveals our robustness in counting. Meanwhile, multi-task learning used in estimating crowd counts also contributes to the robustness of the proposed counting approach.

5. Conclusions

A multi-task Inception network is proposed for counting congested crowds in this work. The proposed model tries to resolve two challenging issues in vision based crowd counting: drastic scale variation and non-uniform density distribution. Counting is achieved through estimating the density maps of different patches, which are further used to reconstruct the density map of the entire image. Several Inception blocks are used to extract multi-scale features, aiming at handling the consecutive scale variations. This multi-scale strategy is comparable to several state-of-the-art multi-scale strategies, *e.g.* MCNN and switching CNN. Meanwhile, non-uniform density distribution of congested crowds is handled through jointly estimating three objectives, namely, crowd density map, BG/FG separation, and crowd density level. Evaluations in three benchmarking datasets revealed the effectiveness of our multi-task strategy in improving counting performance. Comparisons with several state-of-the-art CNN-based crowd counting approaches indicate our superiority in both accuracy and robustness. Our future work will focus on counting crowds with blurry appearance using other technologies, such as super-resolution.

Acknowledgement

This work has been supported by the National Natural Science Foundation of China under Grant No. 61501060 and No. 61703381, the Natural Science Foundation of Jiangsu Province under Grant No. BK20150271, Key Laboratory for New Technology Application of Road Conveyance of Jiangsu Province under Grant BM20082061708.

References

[1] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, et al., *Object detection with discriminatively trained part-based models*, IEEE transactions on pattern analysis and machine intelligence **32** (2010), no. 9, 1627–1645.

[2] C. Gao, J. Liu, Q. Feng, et al., *People-flow counting in complex environments by combining depth and color information*, Multimedia Tools and Applications **75** (2016), no. 15, 9315–9331.

[3] J. Luo, J. Wang, H. Xu, et al., *Real-time people counting for indoor scenes*, Signal Processing **124** (2016), 27–35.

[4] B. Antic, D. Letic, D. Culibrk, et al., *K-means based segmentation for real-time zenithal people counting*, in "16th IEEE International Conference on Image Processing (ICIP)", pages 2565–2568, 2009.

[5] A. S. Rao, J. Gubbi, S. Marusic, et al., *Estimation of crowd density by clustering motion cues*, The Visual Computer, **31**(11):1533–1552,2015.

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

[7] A. B. Chan, M. Morrow, and N. Vasconcelos, *Analysis of crowded scenes using holistic properties*, in "Performance Evaluation of Tracking and Surveillance workshop at CVPR", pages 101–108, 2009.

[8] O. Arandjelovic, *Crowd detection from still images*, in "Proceedings of the British machine vision association conference. BMVA Press", pages 1–10, 2008.

[9] B. Xu and G. Qiu, *Crowd density estimation based on rich features and random projection forest*, in "IEEE Winter Conference on Applications of Computer Vision (WACV)", pages 1–8, 2016.

[10] Y. Li, E. Zhu, X. Zhu, et al., *Counting pedestrian with mixed features and extreme learning machine*, Cognitive Computation **6** (2014), no. 3, 462–476.

[11] H. Fradi and J. Dugelay, *Low level crowd analysis using frame-wise normalized feature for people counting*, in "International Workshop on Information Forensics and Security (WIFS)", pages 246–251, 2012.

[12] M. Hashemzadeh and N. Farajzadeh, *Combining keypoint-based and segment-based features for counting people in crowded scenes*, Information Sciences **345** (2016), 199–216.

[13] R. Liang, Y. Zhu, and H. Wang, *Counting crowd flow based on feature points*, Neurocomputing **133** (2014), 377–384.

[14] P. Siva, M. J. Shafiee, M. Jamieson, *et al.*, *Scene invariant crowd segmentation and counting using scale-normalized histogram of moving gradients (HoMG)*, `arXiv:1602.00386`, (2016).

[15] X. Zhang, H. He, S. Cao, et al., *Flow field texture representation-based motion segmentation for crowd counting*, Machine Vision and Applications **26** (2015), no. 7-8, 871–883.

[16] K. Chen, S. Gong, T. Xiang, et al., *Cumulative attribute space for age and crowd density estimation*, in "Proceedings of the IEEE conference on computer vision and pattern recognition", pages 2467–2474, 2013.

[17] M. A. Mousse, C. Motamed, and E. C. Ezin, *People counting via multiple views using a fast information fusion approach*, Multimedia Tools and Applications **76** (2017), no. 5, 6801–6819.

[18] C. C. Loy, K. Chen, S. Gong, et al., *Crowd counting and profiling: Methodology and evaluation*, Modeling, Simulation and Visual Analysis of Crowds. Springer New York, 347–382, 2013.

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

[20] Y. Zhang, D. Zhou, S. Chen, et al., *Single-image crowd counting via multi-column convolutional neural network*, in "Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition", pages 589–597, 2016.

[21] V. A. Sindagi and V. M. Patel, *Cnn-based cascaded multi-task learning of high-level prior and density estimation for crowd counting*, `arXiv:1707.09605`, (2017).

[22] D. B. Sam, S. Surya, and R. V. Babu, *Switching convolutional neural network for crowd counting*, `arXiv:1708.00199`, (2017).

[23] S. Ren, K. He, R. Girshick, et al., *Faster R-CNN: Towards real-time object detection with region proposal networks*, Advances in neural information processing systems, 91–99, 2015.

[24] L. C. Chen, G. Papandreou, I. Kokkinos, et al., *Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs*, `arXiv:1606.00915`, (2016).

[25] 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 (2017).

[26] C. Wang, H. Zhang, L. Yang, et al., *Deep people counting in extremely dense crowds*, Proceedings of the 23rd ACM international conference on Multimedia, 1299–1302, 2015.

[27] E. Walach and L. Wolf, *Learning to count with CNN boosting. European Conference on Computer Vision*, Springer International Publishing,

660–676, 2016.

[28] M. Marsden, K. McGuinness, S. Little, et al., *Fully convolutional crowd counting on highly congested scenes*, `arXiv:1612.00220`, (2016).

[29] M. Fu, P. Xu, X. Li, et al., *Fast crowd density estimation with convolutional neural networks*, Engineering Applications of Artificial Intelligence **43** (2015), 81–88.

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

[31] M. Marsden, K. McGuinness, S. Little, et al., *ResnetCrowd: A residual deep learning architecture for crowd counting, violent behaviour detection and crowd density level classification*, `arXiv:1705.10698`, (2017).

[32] V. Lempitsky and A. Zisserman, *Learning to count objects in images*, Advances in Neural Information Processing Systems, 1324–1332, 2010.

[33] C. Szegedy, W. Liu, Y. Jia, et al., *Going deeper with convolutions*, in "Proceedings of the IEEE conference on computer vision and pattern recognition", pages 1–9, 2015.

[34] K. He, X. Zhang, S. Ren, et al., *Spatial pyramid pooling in deep convolutional networks for visual recognition*, in "European Conference on Computer Vision. Springer, Cham", pages 346–361, 2014.

[35] K. He, X. Zhang, S. Ren, et al., *Delving deep into rectifiers: Surpassing human-level performance on imagenet classification*, in "Proceedings of the IEEE international conference on computer vision", pages 1026–1034, 2015.

[36] Y. Xian, Z. I. Petrou, Y. Tian, et al., *Super-resolved fine-scale sea ice motion tracking*, IEEE Transactions on Geoscience and Remote Sensing, 2017.

[37] D. Kang, Z. Ma, and A. B. Chan, *Beyond counting: Comparisons of density maps for crowd analysis tasks-counting, detection, and tracking*, `arXiv:1705.10118`, (2017).

Department of Information Science and Engineering
Changzhou University

Changzhou of Jiangsu Province, China 213164
E-mail address: yb6864171@cczu.edu.cn

Department of Information Science and Engineering
Changzhou University
Changzhou of Jiangsu Province, China 213164
E-mail address: 16106212@smail.cczu.edu.cn

Department of Information Science and Engineering
Ocean University of China
Qingdao of Shandong Province, China 266100
E-mail address: wangnanseu@163.com

Department of Information Science and Engineering
Changzhou University
Changzhou of Jiangsu Province, China 213164
E-mail address: 16106209@smail.cczu.edu.cn

Department of Information Science and Engineering
Changzhou University
Changzhou of Jiangsu Province, China 213164
E-mail address: zouling@cczu.edu.cn