Signal Processing: Image Communication 64 (2018) 118–129

Contents lists available at ScienceDirect

Signal Processing: Image Communication

journal homepage: www.elsevier.com/locate/image

Counting challenging crowds robustly using a multi-column multi-task convolutional neural network

Biao Yang [a,*], Jinmeng Cao [a], Nan Wang [b], Yuyu Zhang [a], Ling Zou [a]

[a] *College of Information Science and Engineering, Changzhou University, Changzhou, Jiangsu, China*
[b] *College of Information Science and Engineering, Ocean University of China, Qingdao, Shandong, China*

ARTICLE INFO

Keywords:
Crowd counting
Multi-column CNN
Multi-task
Per-scale loss
Density map

ABSTRACT

Counting challenging crowds from still images has a wide range of applications, such as surveillance event detection, public safety control, traffic monitoring, and urban planning. Early studies on crowd counting focused on extracting hand-crafted features and building effective regression models. However, previous approaches may encounter many challenges, such as partial occlusion, non-uniform density distribution, and variations in scale and perspective. A multi-column multi-task convolutional neural network (MMCNN) is proposed for robust crowd counting, which is achieved through summing up the density map estimated by the proposed network. A novel approach is used to generate the ground truth of density map that focuses on location and detailed information. A multi-column CNN is designed to address drastic scale variation exists in crowds. Per-scale loss is minimized to make the features of different scales highly discriminative. Meanwhile, a multi-task strategy is utilized to simultaneously estimate the density map, crowd density level, and background/foreground mask. Contrastive evaluations in benchmarking datasets are implemented with several state-of-the-art CNN-based crowd counting approaches. Results reveal the accuracy and robustness of our approach in counting challenging crowds. The proposed approach achieves the state-of-the-art performance in terms of mean absolute error and mean squared error. The counting approach can be also extended to other related tasks, such as anomaly detection.

1. Introduction

With the increase in population worldwide, threats in crowded environments, such as fights, riots, and stampedes, are intensifying. For example, on New Year's Eve in 2015, 35 people were killed in a massive stampede in Shanghai, China. Many other massive stampedes have occurred around the world, and these stampedes have taken away many lives. Therefore, something must be done to prevent these tragedies. Crowd counting in public places (e.g., in religious or sport events) can provide useful information for safety control to prevent crowd tragedies. However, it is impossible for people to artificially count the crowds within a short while, especially those congested ones. Therefore, vision based crowd counting, as a key part of intelligent video monitoring, have elicited increased attention from researchers in recent years.

Existing crowd counting approaches mainly fall into two categories, namely, counting by detection and regression. In counting by detection, crowd counting is achieved by detecting instances of people in a scene [1]. To resolve the partial occlusions exist in congested crowds, some researchers only detected noteworthy parts such as heads and shoulders [2]. However, counting by detection is time consuming due to the exhaustive scanning of an image space using a pre-trained detector with variable scales. Meanwhile, it is inaccurate because of the cluttered background. Inspired by the necessary of real-time crowd counting, counting by regression aims to achieve direct mapping between specific visual features and crowd counts without detecting or tracking individuals [3]. Then, these studies focus on extracting hand-crafted features, e.g. foreground areas, shapes, edges, and so on [4]. These features are verified to be effective for counting sparse crowds under normal conditions with a reasonable time consuming.

Despite the rapid development of crowd counting approaches mentioned above, some challenges (Fig. 1), e.g. severe occlusion, diverse density distributions, varying scales and non-uniform illumination, are still not well addressed. Recently, researchers have focused on deep learning. Most of them attempted to count crowds by estimating a density map (Fig. 2). They mapped an input crowd image to its corresponding density map to reveal the number of people per pixel present in the image. Researches argued that features automatically extracted with convolutional neural network (CNN) are more robust to hand-crafted features in handling challenges such as severe occlusion, non-uniform

[*] Corresponding author.
E-mail address: yb6864171@cczu.edu.cn (B. Yang).

https://doi.org/10.1016/j.image.2018.03.004
Received 25 October 2017; Received in revised form 7 March 2018; Accepted 7 March 2018
Available online 12 March 2018

Fig. 1. Challenging crowd images selected from (a) the UCF_CC_50 and (b) the Shanghai Tech datasets. All crowd images suffer from severe occlusion, diverse density distributions, varying scales and non-uniform illumination, and so on.

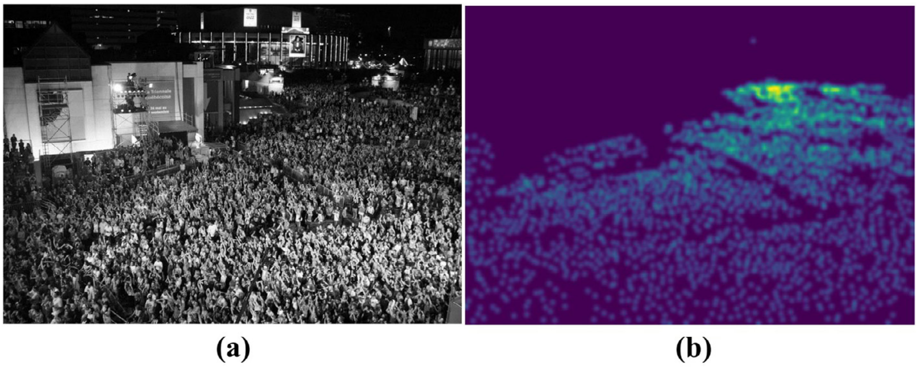

Fig. 2. Example of (a) a source crowd image and its (b) corresponding density map. The yellow regions of the density map indicate high density levels in those regions (best viewed in color).

illumination, varying appearances [5]. However, existed CNN based counting approaches cannot effectively address two problems, namely, non-uniform density distributions and drastic scale variation. Thus, we propose a multi-column multi-task convolutional neural network (MMCNN) to handle these two challenges.

The pipeline of the proposed counting approach can be divided into a training and an evaluation stage (Fig. 3). In both stages, the input images are converted into gray ones (if necessary) and then are normalized into 960 × 960 (except those images in UCSD and MALL datasets). Later, in the training stage, 16 non-overlapped patches are uniformly segmented from the image. For each patch, its ground truths (density map, density level, and background/foreground (BG/FG) mask) are calculated in advance. Unlike Sindagi et al. who classified the crowds into 10 density levels [6], we uniformly classify the crowds into 5 density levels: very high density, high density, medium density, low density, and very low density according to their actual counts. The aim of using less levels is to relieve the imbalance among different levels. Then, all patches and their corresponding ground truths are used together to learn parameters of feature extraction and multi-task learning networks through iterative optimization. In the evaluation stage, the input image is segmented into overlapped patches (240 × 240 pixels) with a fixed stride (20 pixels). These patches can be arbitrary sizes theoretically, however, the minimum and maximum sizes are experimentally set to 120 ×120 and 480 × 480, respectively. All patches are then fed to the well-learned feature extraction network in sequence. Outputs of three columns are fused together and then fed to the multi-task learning network. Notably, only density maps of different patches are employed to reconstruct the density map of the entire crowd image. Estimating density levels and BG/FG masks are used to refine the estimated density maps. Finally, the crowd count is estimated through integrating all values of the reconstructed density map. Extracting patches in different manners during training and evaluation stages are inspired by Marsden et al. [7]. They argued that extracting non-overlapped patches in the training stage is better than extracting overlapped patches because too much redundancy existed in the overlapped patches. The redundancy may lead to over-fitting and a poor generalization capability.

This work provides three novel contributions. First, we propose a new density map that focuses on both the location and detailed information. Second, a multi-column CNN is designed to extract features of different scales. Per-scale loss is minimized to make the learned features highly discriminative. Third, our model jointly estimate the density map, crowd density level, and BG/FG mask. Accuracy of the estimated density map is improved through jointly other correlated objectives. The rest of this paper is organized as follows. Section 2 provides a review of related work on crowd counting approaches. Section 3 presents the details of the proposed method. Section 4 shows the experimental results and analysis. The conclusions are presented in Section 5.

2. Related work

Various approaches have been proposed for vision based crowd counting. We will review some related works on traditional and CNN based counting approaches.

2.1. Review of traditional counting approaches

Most early studies focused on designing a detection framework to detect people in a scene and then estimated the crowd counts through detection results. Sabzmeydani et al. detected pedestrians using shapelet features [8]. Dalal et al. proposed a new feature named histograms of oriented gradients that is proofed to be very effective for pedestrian detection [9]. However, detecting the entire body maybe easily affected by partial occlusion. Thus, Gao et al. detected heads with a water filling algorithm and counted crowds through computing the number of detected heads [10]. Luo et al. built a head–shoulder model for depicting moving and stationary crowds [11]. Partial occlusion can be handled through detecting local parts. However, all detection approaches need to scan the input images with sliding windows of different scales. Thus, they are time consuming and are easily affected by cluttered background.

In counting by detection, researchers tried to learn a direct mapping between specific features and crowd counts. For example, Fradi et al. used foreground pixels and corner density for crowd counting [12]. Hashemzadeh et al. utilized a combination of key point (corners) and segment-based features to estimate crowd counts [13]. Liang et al.

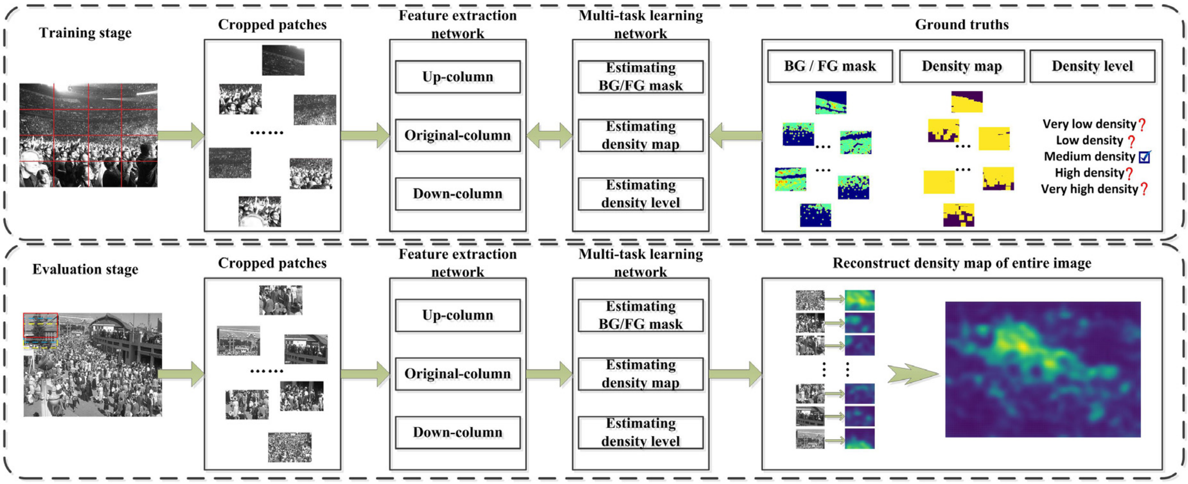

Fig. 3. Pipeline of the proposed MMCNN.

employed another kind of key points (speeded up robust feature) as cues for crowd counting [14]. Inspired by their work, Li et al. tried to combine multiple features, e.g. object size, shape, edges, and local binary patterns (LBPs) for crowd counting [15]. Aside from these commonly used features, researchers also proposed new features which are highly suitable for crowd counting. Shafiee et al. proposed a novel low-complexity, scale-normalized feature called histogram of moving gradients (HoMG) [16]. Zhang et al. modeled crowds as a flow and proposed a flow field texture representation approach to depict segmented crowds [17]. Chen et al. introduced a novel cumulative attribute for learning a regression model when only sparse and imbalanced crowd images were available [18]. These approaches are satisfactory when counting sparse crowds. 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 [19]. It is robust to partial occlusion, however, it can only be used in multi-camera networks with overlapping views.

2.2. Review of CNN based counting approaches

As discussed above, traditional counting approaches suffer from many challenges, e.g. severe occlusion, non-uniform density distribution, and varying scale. Recently, CNN has achieved great success in pattern recognition, such as object detection [20] and semantic segmentation [21]. According to a recent review on crowd counting [5], CNN based counting approaches have outperformed the traditional ones. In early studies, Fu et al. [22] and Pu et al. [23] estimated the density level of the crowds. However, accurate crowd counts are needed sometimes. Wang et al. proposed an end-to-end CNN regression model based on the AlexNet network [24] for crowd counting. Inspired by their success, Walach et al. performed layered boosting through iteratively adding CNN layers to the model to estimate the residual error of the initial prediction [25]. 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 [7]. For better counting performance, Sindagi et al. used contextual pyramid for generating high-quality crowd density and count estimation through incorporating global and local contextual information [26]. Xiong et al. used convolutional LSTM for crowd counting. Both spatial and temporal dependencies can be obtained through exploiting useful temporal information in video sequences [27]. Despite their successes in handling severe occlusion, illumination change, and cluttered background, they cannot resolve the problems of scale variation and non-uniform density distribution.

To resolve scale variation, Zhang et al. designed a multi-column CNN (MCNN) to extract features of different scales [28]. Inspired by their work, Onoro et al. developed a counting approach called Hydra-CNN that can also extract features of different scales [29]. Meanwhile, they down-sampled the input data to enhance the ability of scale-aware. Except for MCNN, Sam et al. argued that improved counting performance can be obtained through replacing multiple columns with the most suitable one [30], which is selected using a VGG based classifier. Zeng et al. designed a single column CNN with multi-scale blobs, which can be used to generate scale-relevant features for higher crowd counting performance [31]. Except for multi-scale features, Lokesh et al. combined both the high-level semantic information (face/body detectors) and the low-level features (blob detectors) for crowd counting under large scale variations [32]. All these approaches are effective in handling scale variation, however, they did not fully utilize the relations among different scales.

Multi-task learning is commonly used to handle non-uniform density distribution. For instance, Zhang et al. alternatively trained two related objectives, namely, crowd count and density map [33]. However, they did not perform well when counting extremely dense. Sindagi et al. proposed a cascaded multi-task CNN (Cascade-MTL) that jointly estimates the density level and density map [6]. Crowd counts could be further estimated from the density map. Through jointly estimating the density level with density map, non-uniform density distribution of the crowds could be partly handled. Actually, density distribution is highly related to the locations of crowds. Thus, crowd foregrounds can be added into the multi-task strategy to further improve the counting performance.

3. Proposed method

3.1. Generating combined density maps

The main objective of the proposed counting approach 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 position of each person is labeled, as well as the perspective images of different scenes. Notably, perspective images are pre-calculated using the approach proposed by Chan et al. [4]. Many studies followed [34] and defined the density map as a sum of Gaussian kernels (Fig. 4(a)) centered on object locations. This type of density map only focuses on the location information of each people in the crowd image. Later, Zhang et al. proposed to generate density map with a human-shaped kernel (Fig. 4(b)) that is comprised of a bivariate normal distribution [33].

This type of density map focuses on both the location and detailed information, thus is more suitable for characterizing the density distribution of crowds. However, it may fail when characterizing congested crowds where only heads can be reliably observed. Inspired by their work, we propose a novel way to generate density map that is composed of a location and a detailed density map. Similar to Lempitsky et al., the location density map is generated using a sum of Gaussian kernels centered on heads to encode location information (Fig. 4(d)). Notably, only head parts are used to generate density map in this work. Thus, edges of heads can be used to provide detailed information. However, it is tedious to label the edges of different heads in a congested crowds. But, the boundary (the white circle in Fig. 4(c)) between Gaussian kernel and the background approximately reveal the edge information. Thus, the detailed density map is generated through inversing the location density map (Fig. 4(e)). Finally, the used density map is generated based on abovementioned two density maps in a weighted manner (Fig. 4(f)).

Details to generate the used density map is given as follows. Initially, after obtaining the center positions of heads P_h in the patch, its location density map is generated as

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

where p is an arbitrary position in the ith patch P_i and N_g is a normalized 2D Gaussian kernel with variance σ_h (setting of σ_h can refer to [33]). 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. Then, detailed density map is defined through reversing the location density map. For each non-zero value x in location density map $D_i(p)$, its corresponding value in detailed density map $\widetilde{D}_i(p)$ is defined as $(1/\exp(x))$. Finally, the proposed density map $\widehat{D}_i(p)$ can be calculated as follows.

$$\widehat{D}_i(p) = \omega_1 D_i(p) + \omega_2 \widetilde{D}_i(p) \qquad s.t. \qquad \omega_1 + \omega_2 = 1 \tag{2}$$

where ω_1 and ω_2 are weights for different density maps. Both weights are set to 0.5 in this work. The effectiveness of the new density map will be discussed in Section 4.

3.2. MMCNN

MMCNN is designed to estimate a density map, which can be further used for crowd counting. MMCNN can be divided into two parts, including a feature extraction network and a multi-task learning network. Notably, our purpose is to study the problems of non-uniform density distribution and drastic scale variation, which remain challenging in state-of-the-art CNN based crowd counting approaches.

The feature extraction network is similar to the MCNN proposed by Zhang et al. [28]. We mainly make three changes. First, both down-sampling and up-sampling are utilized to the input patch to guarantee that effective features can be extracted from different scales. Second, deconvolutions are used to compensate for the loss in detail caused by early pooling layers. Last but not the least, per-scale loss is used and is minimized to make the learned features highly discriminative. The effectiveness of these measures to handle drastic scale variation will be discussed in Section 4.3.

The multi-task learning network is similar to the Cascade-MTL proposed by Sindagi et al. [6]. Aside from estimating density map and density level simultaneously, a BG/FG mask is added into the multi-task learning framework. Notably, only the density map is utilized to count the crowds, while other two correlated objectives are used to improve the estimated density map. Employment of density level can refine the estimated density map in intensity, while employment of BG/FG mask can refine it in distribution. Both refinements are beneficial for the estimated density map to achieve more accurate crowd counting.

On the basis of the abovementioned discussions, structure of MMCNN is illustrated in Fig. 5. The legend on the top-right corner illustrates different colors of convolutional filters with different kernel sizes. Kernel sizes and filter numbers are labeled in the figure in detail. The left part of MMCNN is the feature extraction network, which consists of three columns, namely, up-, original-, and down-column. These columns process up-sampled, original, and down-sampled patches, respectively. Four convolutional layers with local receptive fields of different sizes exist in each column. Filters with large receptive fields are generally highly effective in modeling density maps that correspond to large heads. To reduce computational complexity, few filters are used for CNN columns with large receptive fields. The first two convolutional layers in each column are followed by max pooling layers with a stride of 2, due to which the outputs of conv1_2, 2_2, and 3_2 are down-sampled by a factor of 4. Each convolutional layer (except conv6, 7, 11, 12, and 13) 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. For the fourth convolutional layer in each column, deconvolutions are adopted for up-sampling. The times of deconvolutions depend on the size of output features produced by conv1_4, 2_4, and 3_4. Up-sampling is carried out to guarantee the same size ($W \times H$) for feature fusion and to compensate for the loss of details due to early pooling. The outputs of different columns are concatenated in the fusion layer. Moreover, these outputs are fed to three 1 × 1 convolutional layers (Conv11, 12, 13) for calculating the per-scale loss. Notably, these three convolutional layers will be removed after MMCNN is learned. The right part of MMCNN is the multi-task learning network. To estimate density level that is the more abstract among three objectives, three convolutional layers (Conv8, 9, 10) are utilized to further process the fused features. Then, a spatial pyramid pooling (SPP) [35] of three layers is used to remove the constraint of fixed size. 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, fused features are directly fed to a 1 × 1 convolutional layer (Conv7) for estimating BG/FG mask. To estimate the density map that contains more information than BG/FG mask, a 3 × 3 convolutional layer (Conv5) is utilized to process the features before they are fed to Conv6. Strides of all convolutional filters are set to 1.

3.3. Details of loss functions used in multi-task learning

The proposed MMCNN is trained in the manner of multi-task learning. The loss between the estimated density map and its ground truth is defined as $L_{density}$, which is calculated using Euclidean loss. $L_{density}$ is defined as follows:

$$L_{density} = \frac{1}{2N} \sum_{i=1}^{N} \|F_d(P_i, O) - \widehat{D}(P_i)\|_2 \tag{3}$$

where N is the number of training samples, O is a set of network parameters, P_i is the ith patch, $F_d(P_i, O)$ is the estimated density map of P_i, and $\widehat{D}(P_i)$ is the ground truth of $F_d(P_i, O)$.

Jointly estimating density level is beneficial to refine the intensity of estimated density map. The loss between the estimated density level and its ground truth is defined as L_{level}, which is calculated using the cross-entropy loss. L_{level} is defined as follows:

$$L_{level} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{M} [(\widehat{Y}(P_i) = j) F_c(P_i, O)] \tag{4}$$

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 $\widehat{Y}(P_i)$ is its ground truth.

Jointly estimating BG/FG mask is beneficial to refine the distribution of estimated density map. The loss between the estimated BG/FG mask and its ground truth is defined as L_{mask}. L_{mask} can be also calculated using Euclidean loss as follows:

$$L_{mask} = \frac{1}{2N} \sum_{i=1}^{N} \|F_m(P_i, O) - \widehat{M}(P_i)\|_2 \tag{5}$$

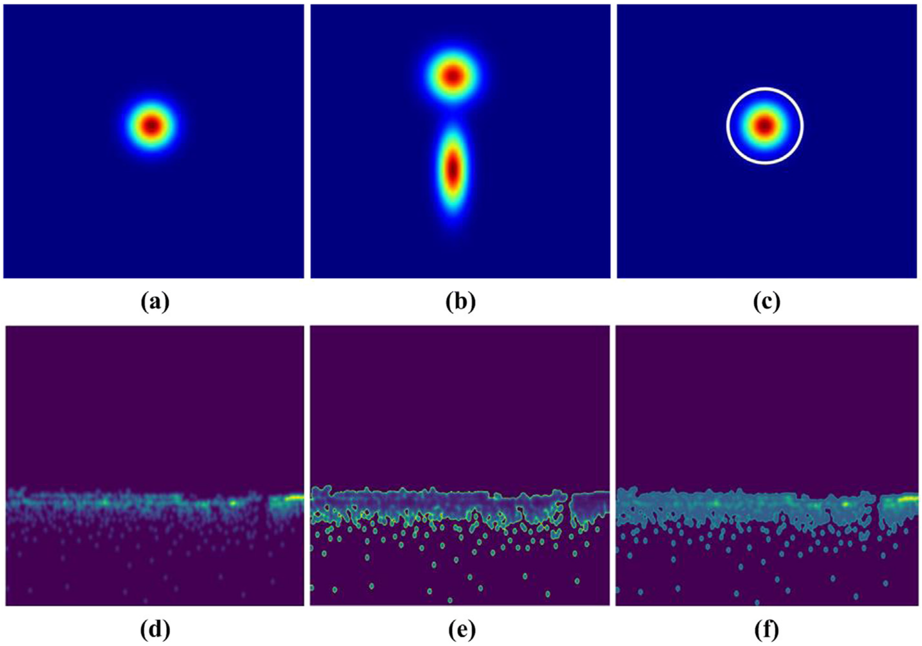

Fig. 4. Illustration of (a) Gaussian kernel, (b) human shaped kernel, (c) boundary (the white circle) between Gaussian kernel and the background, (d) location density map, (e) detailed density map, (f) the used density map. The yellow regions indicate a high crowd density level (best viewed in color).

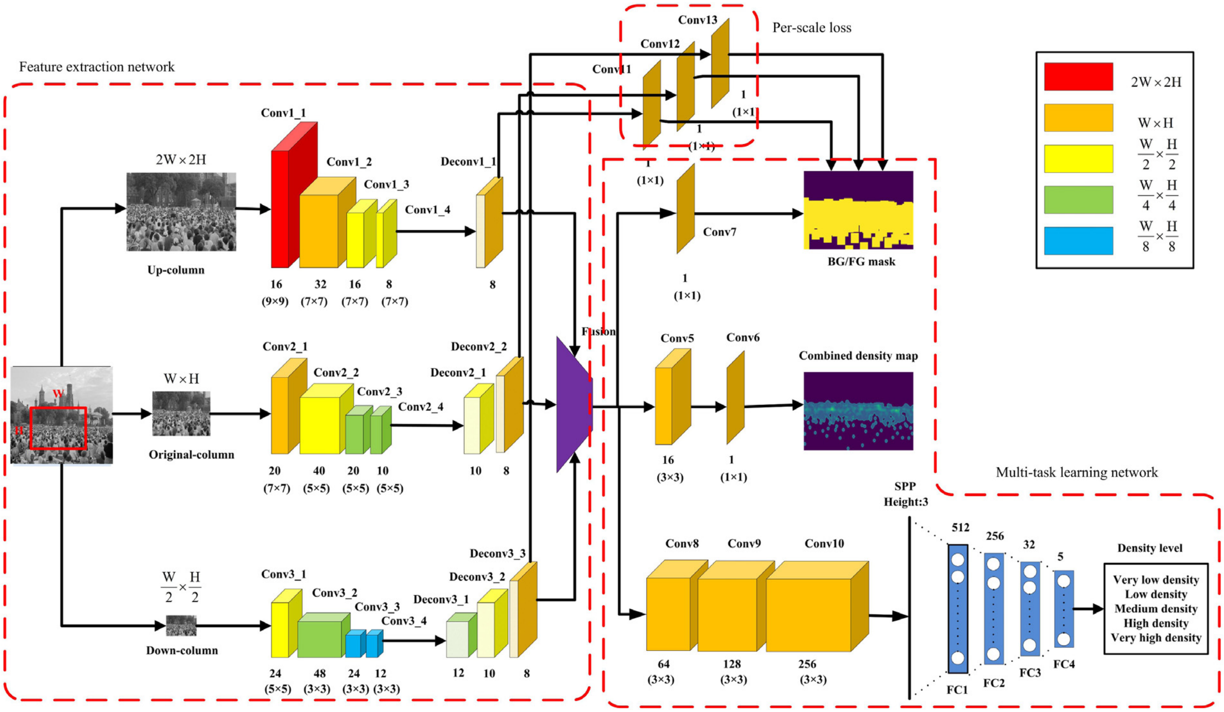

Fig. 5. Structure of MMCNN. Dropout, PReLU, and LRN are not listed for simplification.

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

Inspired by scale attention [36], different scales consistently have different receptive regions, which are mainly location information. To handle scale variation, a per-scale loss is minimized to make the features of different scales highly discriminative, thereby improving counting performance. In consideration of the correlation between scale attention and the BG/FG mask (both focus on location information), PL_{mask} is defined as follows:

$$PL_{mask} = \frac{1}{2N} \sum_{i=1}^{N} \| F_m(P_i, O) - \widehat{M}(P_i) \|_2 + \sum_{j=1}^{M} \alpha \cdot \frac{1}{2N} \sum_{i=1}^{N} \| F_m^j(P_i, O_j) - \widehat{M}(P_i) \|_2 \quad (6)$$

Table 1
Parameters used to train MMCNN.

Parameters	Value
Base learning rate	0.0001
Learning policy	“inv”
Power	0.75
Gamma	0.001
Max iterations	2000000
Momentum	0.9
Weight decay	0.005
Optimization type	Adam

where $M (M = 3)$ represents the number of scales in MMCNN, α represents the weight for per-scale loss (each scale owns the same weight), O_j comprises the parameters of the CNN column with scale j, and $F_m^j(P_i, O_j)$ is the estimated BG/FG mask of P_i with parameter O_j.

Finally, the total loss of MMCNN can be defined as

$$L_{total} = \lambda_1 L_{density} + \lambda_2 L_{level} + \lambda_3 PL_{mask} \tag{7}$$

where λ_1, λ_2, and λ_3 are weights of different loss functions. We experimentally set $\lambda_1 = 1$, $\lambda_2 = 0.001$, and $\lambda_3 = 0.1$ according to the observed losses. Setting of λ_1, λ_2, and λ_3 are based on following reasons: (1) $L_{density}$ plays a main role in L_{total}, thus the value of $\lambda_1 \times L_{density}$ is about one order of magnitude larger than those of $\lambda_2 \times L_{level}$ and $\lambda_1 \times PL_{mask}$. (2) L_{level} is about two orders of magnitude larger than PL_{mask}, while $L_{density}$ and PL_{mask} own the same order of magnitude. We initialize α to be large in order to learn scale information according to Chen et al. [36]. Then, α is linearly decreased to make L_{total} focus on other works (like estimating the density map). Initial and final values of α are experimentally set to 60 and 0.01, respectively.

4. Experimental analysis

4.1. Implementing details of MMCNN and evaluation metrics

Training and evaluation of MMCNN are performed on NVIDIA GTX 1080 GPU (8G) using the Caffe framework and a cuDNN library. Table 1 lists the parameters used to train MMCNN. Batch size is set to 16 during training due to memory limitations. MSRA [37] is used for initialization of each convolutional layer.

We evaluate different crowd counting approaches with two indicators, namely, mean absolute error (MAE) and mean squared error (MSE), which are defined as follows:

$$\mathrm{MAE} = \frac{1}{W} \sum_{i=1}^{W} |E(i) - G(i)| \tag{8}$$

$$\mathrm{MSE} = \sqrt{\frac{1}{W} \sum_{i=1}^{W} (E(i) - G(i))^2} \tag{9}$$

where W is the total number of test frames, $G(i)$ is the actual number of people in the crowds, and $E(i)$ is the predicted number of people in the ith frame. Generally, MAE and MSE indicate the accuracy and robustness of the estimates, respectively.

For dense crowds, we define a metric named variance of head–head distances (VHHD) to represents the level of scale variation. VHHD is calculated using the variance of a vector, which contains the distances between each head and its nearest neighbor. A large VHHD indicates drastic scale variation. Notably, head locations should be labeled in advance and this metric is only used to evaluate the effectiveness of our approach in handling scale variation.

4.2. Datasets and evaluation settings

(1) UCSD [3]

This dataset was among the first ones to be created for crowd counting. It is an hour of video collected from a stationary digital camera overlooking a pedestrian walkway at UCSD. The dataset contains 2000 frames of size 238 × 158 from the video sequence, along with ground truth annotations of each pedestrian in every fifth frame. Linear interpolation is used to create the annotations for the rest of the frames. The dataset contains 49,885 pedestrian instances, and the number of individuals varies from 11 to 46, with an average of 25 individuals per image. It is also split into a training set (frames 600 to 1399) and a test set (the rest of the 1200 frames). For fairness, the procedure of training-evaluation was repeated five times and the average mean absolute error (MAE) and mean squared error (MSE) were used for evaluation. Generally, the UCSD dataset consists of low-density crowd images, and no variation in scene perspective exists across images. Notably, images in UCSD are initially resized to 240 × 160. In the training stage, each image is segmented into 16 non-overlapping patches. In the evaluation stage, each image is segmented into 60 × 40 patches with stride 10.

(2) MALL [38]

This dataset was collected using a surveillance camera installed in a shopping mall. It contains 2000 frames of size 320240 with 6000 labeled pedestrians. The first 800 frames were used for training, and the other 1200 frames were used for evaluation. The procedure of training-evaluation was also repeated five times and the average MAE and MSE were used for evaluation. The MALL dataset covers crowd density from sparse to dense and consists of 13 to 53 people. Aside from the varying density level, this dataset also poses other challenges, such as severe perspective distortion, partial occlusion, and different activity patterns (static and moving crowds). Notably, image size (640 × 480) in MALL is unchanged. In the training stage, each image is segmented into 16 non-overlapping patches. In the evaluation stage, each image is segmented into 160 × 120 patches with stride 20.

(3) UCF_CC_50 [39]

UCF_CC_50 is the first truly challenging dataset because it contains a wide range of densities and diverse scenes with varying illumination conditions and perspective distortion. This dataset was created from publicly available web images and includes such scenes as concerts, protests, stadiums, and marathons. A total of 63,075 individuals are labeled in the entire dataset, with an average of 1280 individuals per image. The number of individuals varies from 94 to 4543, indicating a large variation across the crowd images. However, this dataset contains only 50 images with varying resolution. This limited number of images may affect training and evaluation using this dataset. For this dataset, five-cross validation was used to evaluate different crowd counting approaches.

(4) WorldExpo’10 [33]

This dataset was proposed by et al., who focused on cross-scene crowd counting. This large-scale dataset contains 1132 annotated video sequences captured by 108 surveillance cameras from the Shanghai 2010 WorldExpo event. It consists of 3980 frames of size 576 × 720, with 199,923 labeled pedestrians. The number of individuals varies from 1 to 253, with an average of 50 individuals per image. The dataset is split into training and testing sets. The former contains 1126 one-minute video sequences from 103 scenes, and the latter contains five one-hour video sequences from five scenes. Each test scene consists of 120 labeled frames, with the number of pedestrians varying from 1 to 220. We trained our model on the training set and evaluated it on the testing set according to the official guidance. For a fair comparison, perspective maps are used to generate ground truth maps and ROI maps are used to postprocess the estimated density map.

(5) Shanghai Tech [28]

This dataset was recently introduced by Zhang et al. for training and evaluating crowd counting approaches. It contains 1198 images with 330,165 annotated heads. This dataset is split into Parts A and B. Part A contains 482 images randomly selected from the Internet, whereas Part B contains images obtained from the streets of metropolitan areas in Shanghai, China. Training and testing of Part A involved 300 and 182 images, respectively, and the number of individuals varied from 33 to 3139, with an average of 501 individuals per image. For Part B, 400 and 316 images were used in the training and test sets, respectively, and the number of individuals varied from 9 to 578, with an average of 123 individuals per image. Generally, this dataset is among the largest ones in terms of the number of annotated people. For each part, the procedure of training-evaluation was also repeated five times for fairness.

4.3. Evaluations of MMCNN

4.3.1. Evaluations of the combined density map

The combined density map focuses on both the location and detailed information of congested crowds. To verify its effectiveness, different density maps are evaluated using the proposed approach in benchmarking datasets. MAE and MSE are used for evaluation and the results are given in Table 2.

As shown in above table, the Gaussian density map performs the worst in terms of MAE and MSE because it only focuses on location information of pedestrians. The human shaped density map performs well in crowds with a low density level because it depicts the global outlines of pedestrians. Its MAE and MSE in UCSD and MALL datasets are a little lower than those of combined density map. However, it do not perform well when counting dense crowds because only heads can be reliably observed in these crowds. Combined density map obviously outperforms other density maps in dense crowds, especially the crowds in UCF_CC_50 and Shanghai Tech Part A datasets, due to its focus on location and detailed information of head parts. In consideration of fairness, Cascade-MTL and MCNN are trained with the combined density map in two challenging datasets (UCF_CC_50 and Shanghai Tech Part A). It can be observed from Table 3 that combined density map can achieve lower MAE/MSE compared with Gaussian density map or human shape density map.

4.3.2. Evaluations of the multi-scale strategy

The multi-scale strategy is our main novelty. To proof its effectiveness, two approaches are employed for comparison in benchmarking datasets. The first approach uses no multi-scale strategy (just use the original column) while the second one uses the same multi-scale strategy as ours but does not employ the per-scale loss. MAE and MSE are used to evaluate different approaches, and the results are given in Table 4.

As shown in Table 4, the second approach outperforms the first one in most datasets that contain crowds of medium or high density levels. However, the first approach outperforms the second one in UCSD dataset, which contains crowds of low density level. It reveals that for such crowds, directly fusing features of multiple columns without any special processing (such as minimizing per-scale loss) may be worse than directly using features of original column. Therefore, minimizing the per-scale loss is necessary to guarantee the effectiveness of our multi-column CNN. MAE and MSE of our multi-scale strategy are lower than those of other two approaches (Table 3). For datasets (UCF_CC_50 and Shanghai Tech Part A) owning congested crowds, significant improvements can be observed in both MAE and MSE. Notably, our multi-scale inputs are more effective in capturing local details than single-scale input used in MCNN. Obvious decreases can be observed through replacing single-scale input in MCNN with multi-scale inputs in terms of MAE and MSE while counting dense crowds, such as UCF_CC_50 (MAE: 377.6 → 362.6, MSE: 509.1 → 476.2) and Shanghai Tech Part A (MAE: 110.2 → 105.3, MSE: 173.2 → 168.8).

4.3.3. Evaluations of the multi-task strategy

The multi-task strategy is another novelty of this work. To evaluate its effectiveness, different combinations of objectives, including estimating density map, estimating density map and density level, estimating density map and BG/FG mask, as well as our approach are tested in benchmarking datasets. MAE and MSE are used for evaluation and the results are presented in Table 5.

The first approach only estimates the density map. Thus, it performs worst in all used datasets among different approaches. The second approach jointly estimates the density map and density level. It refines the estimated density map in intensity, thus MAE and MSE are both decreased. The third approach simultaneously estimates the density map and BG/FG mask. It refines the estimated density map in distribution, thus MAE and MSE are also decreased compared with those of the first approach. In addition, counting performance of the second and the third approaches are similar to each other. Our multi-task strategy incorporates the advantages of both BG/FG mask and density level. Thus, its counting performance in benchmarking datasets is superior to that of other approaches, in terms of MAE and MSE (Table 5).

Furthermore, the effectiveness of the BG/FG mask is evaluated by comparing the density maps obtained using the different approaches. Fig. 6(a) shows a crowd image selected from the UCF_CC_50 dataset. Fig. 6(b) is the ground truth of the corresponding density map, and Fig. 6(c), (d) shows the density maps estimated using the different approaches. Density map A was estimated without BG/FG mask while density map B was estimated with BG/FG mask but without the minimization of per-scale loss. Obvious activations (yellow regions) can be seen in the background of density map A (Fig. 6(c)). However, these wrongly activated regions can be effectively restrained by jointly estimating the BG/FG mask even without minimizing per-scale loss (Fig. 6(d)). This founding reveals the effectiveness of BG/FG mask in handling non-uniform density distribution.

4.3.4. Evaluations in different benchmarking datasets

Results of MMCNN in UCSD dataset are illustrated in Fig. 7. G represents the ground truth count, C represents the estimated count and E represents the count error. To evaluate better the obtained results, we use "↑" and "↓" to represent the deviations between our approach and the state-of-the-arts. "↑" indicates that result of other approach is larger than ours while"↓"indicates the opposite. It can be observed from Fig. 7 that our approach performs well when counting sparse crowds. The largest count error is 2 (Fig. 7(d)) and one counting result (Fig. 7(b)) is completely identical to the ground truth.

Results of MMCNN in MALL dataset are illustrated in Fig. 8. Our approach also performs well in this dataset, which has crowds of low or middle density levels. The largest count error among the five examples is 3 (Fig. 8(a)). Counting result of Fig. 8(c) is totally identical to its ground truth.

To evaluate dense crowds, ground truth and estimated density maps on sample images are also illustrated. VHHD represents the degree of scale variation. Results of MMCNN in UCF_CC_50 dataset is shown in Fig. 9. It can be observed that counting performance in UCF_CC_50 dataset is not as good as that in abovementioned two datasets. The largest count error (404↑) occurs when counting the crowds in the Muslim square (Fig. 9(c)). However, the estimated density maps are similar to the ground truth ones even when there are shelters in the crowds (Fig. 9(c), (d) and (e)).

Results of MMCNN in WorldExpo'10 dataset is shown in Fig. 10. Each sample image is selected from a testing scene. Generally, our approach performs well in this dataset, especially for crowds with consecutive scale variations (Fig. 10(a), (b) and (e), as indicated by VHHD). The largest count error is below 11 among the five samples. From the perspective of density maps, the estimated ones are similar to the ground truth ones. Regions with crowds are activated accurately, while the background regions have little or none activations.

Results of MMCNN in Shanghai Tech dataset is shown in Fig. 11. The first three samples are selected from Part A, and the left are selected

Table 2
Evaluations of different density maps in benchmarking datasets.

	Gaussian density map		Human shape density map		Combined density map	
	MAE	MSE	MAE	MSE	MAE	MSE
UCSD	1.07	1.31	**1.00**	**1.13**	1.02	1.18
MALL	2.11	6.59	**1.94**	**5.57**	1.98	5.68
UCF_CC_50	341.1	356.7	337.2	343.6	**320.6**	**323.8**
WorldExpo'10	10.5	22.1	9.8	20.1	**9.1**	**18.7**
Shanghai Tech Part A	105.9	150.8	103.6	145.7	**91.2**	**128.6**
Shanghai Tech Part B	22.7	38.6	21.3	35.6	**18.5**	**29.3**

Table 3
Evaluations of combined density map in Cascade-MTL and MCNN.

MAE/MSE	UCF_CC_50			Shanghai Tech Part A		
	Gaussian density map	Human shape density map	Combined density map	Gaussian density map	Human shape density map	Combined density map
Cascade-MTL	322.8/341.4	N/A	315.9/336.3	101.3/152.4	N/A	96.7/142.8
MCNN	N/A	377.6/509.1	355.2/468.2	N/A	110.2/173.2	107.5/171.6

Table 4
Evaluations of the multi-scale strategy in benchmarking datasets.

	No multi-scale		Multi-scale without per-scale loss		Our multi-scale strategy	
	MAE	MSE	MAE	MSE	MAE	MSE
UCSD	**1.02**	1.21	1.04	1.26	**1.02**	**1.18**
MALL	2.18	6.17	2.31	6.87	**1.98**	**5.68**
UCF_CC_50	366.2	687.3	363.8	509.4	**320.6**	**323.8**
WorldExpo'10	9.5	19.4	9.9	20.6	**9.1**	**18.7**
Shanghai Tech Part A	171.4	163.2	166.3	152.4	**91.2**	**128.6**
Shanghai Tech Part B	20.8	31.1	21.7	32.3	**18.5**	**29.3**

Table 5
Evaluations of the multi-task strategy in benchmarking datasets.

	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
UCSD	1.09	1.41	1.05	1.31	1.06	1.21	**1.02**	**1.18**
MALL	2.98	9.98	2.46	7.83	2.32	7.53	**1.98**	**5.68**
UCF_CC_50	372.1	496.7	367.2	513.5	370.5	462.6	**320.6**	**323.8**
WorldExpo'10	11.4	23.6	10.2	21.3	10.1	20.7	**9.1**	**18.7**
Shanghai Tech Part A	108.9	168.7	101.5	174.3	105.7	171.8	**91.2**	**128.6**
Shanghai Tech Part B	27.1	42.3	23.6	36.7	22.4	32.9	**18.5**	**29.3**

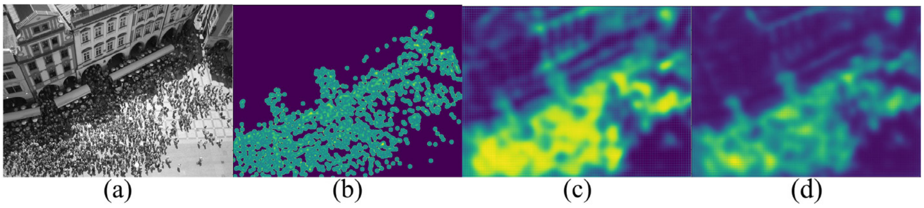

Fig. 6. Effectiveness of the BG/FG mask. (a) Original crowd image selected from the UCF_CC_50 dataset, (b) combined density map, (c) estimated density map A, and (d) estimated density map B. The yellow regions represent strong activations, which indicate high density levels in that region (best viewed in color).

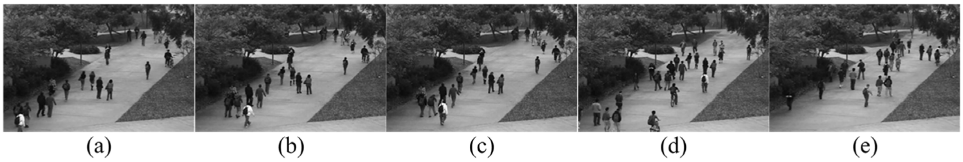

Fig. 7. Results of MMCNN in UCSD dataset. (a) G: 18, C: 17, E: 1 ↓; (b) G: 19, C: 19, E: 0; (c) G: 21, C: 20, E: 1 ↓; (d) G: 24, C: 26, E: 2 ↑; (e) G: 24, C: 23, E: 1 ↓.

from Part B. Our approach performs well in counting dense crowds, even with drastic scale variation (Fig. 11(e), as indicated by VHHD). The largest count error (89↑) occurs when counting the crowds in a park (Fig. 11(b)). From the perspective of density maps, the estimated ones are very similar to the ground truth ones. In addition, shelters in the crowds (like vehicles in Fig. 11(a) and flag in Fig. 11(b)) can be

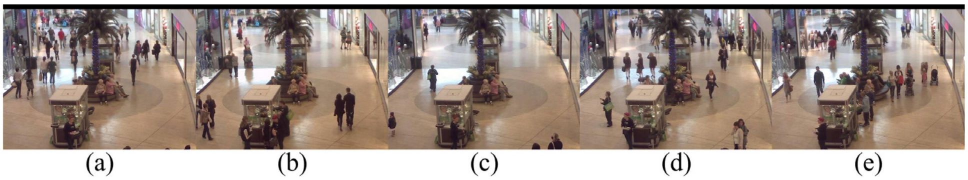

Fig. 8. Results of MMCNN in MALL dataset. (a) G: 33, C: 36, E: 3 ↑; (b) G: 26, C: 25, E: 1 ↓; (c) G: 14, C: 14, E: 0; (d) G: 28, C: 26, E: 2 ↓; (e) G: 27, C: 26, E: 1 ↓.

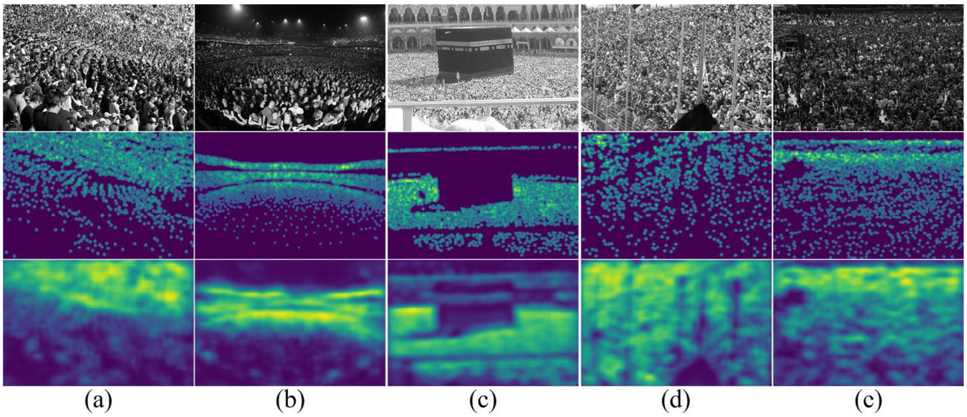

Fig. 9. Results of MMCNN in UCF_CC_50 dataset. First row: Input images. Middle row: Ground truth density maps. Last row: Estimated density maps. For different column: (a) VHHD: 35.6, G: 1940, C: 2190, E: 250↑; (b) VHHD: 28.3, G: 2961, C: 3197, E: 236↑; (c) VHHD: 13.0, G: 2105, C: 2509, E: 404↑; (d) VHHD: 56.3, G: 1037, C: 1087, E: 50↑; (e) VHHD: 28.7, G: 2372, C: 2112, E: 260↓.

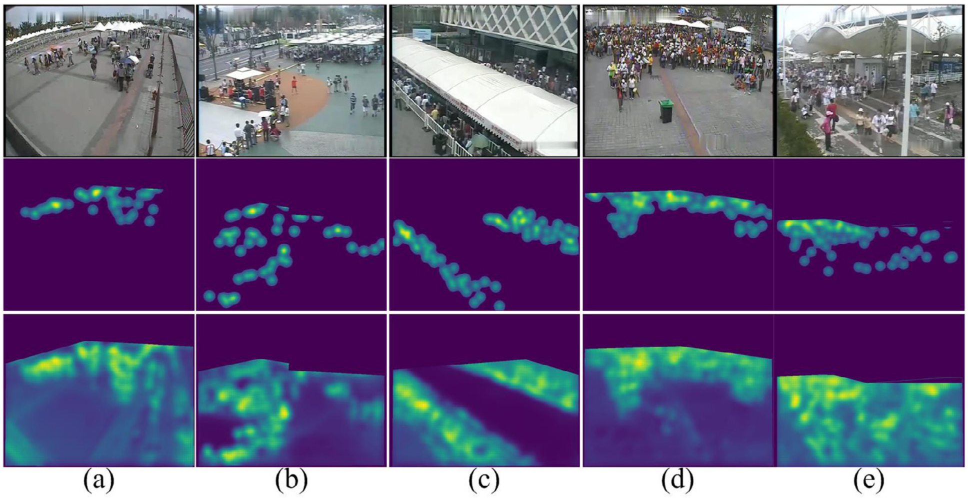

Fig. 10. Results of MMCNN in WorldExpo'10 dataset. First row: Input images. Middle row: Ground truth density maps. Last row: Estimated density maps. For different column: (a) VHHD: 159.2, G: 49, C: 54, E: 5↑; (b) VHHD: 173.7, G: 50, C: 47, E: 3↓; (c) VHHD: 125.4, G: 73, C: 76, E: 3↑; (d) VHHD: 30.2, G: 145, C: 151, E: 6↑; (e) VHHD: 146.0, G: 88, C: 99, E: 11↑.

effectively eliminated from the estimated density maps through jointly learning BG/FG masks.

4.4. Comparisons with state-of-the-art CNN based counting approaches

According to a recent review of crowd counting [5], CNN-based approaches greatly outperform traditional ones. Thus, only state-of-the-art CNN-based counting approaches are employed for comparison in benchmarking datasets. Both MAE and MSE are used for comparison. For each dataset, we select approaches that have reported results in their original works.

Table 5 presents the comparison results in UCSD dataset, which contains relatively sparse crowds. The cross-scene counting approach proposed by Zhang et al. [33], MCNN proposed by Zhang et al. [28], CNN-pixel counting proposed by Kang et al. [40], and switching CNN proposed by Sam et al. [30] are used for comparison. Among the approaches used for comparison, MCNN performs the best by jointly considering different scales of the crowd images. Cross-scene counting focuses on how to select similar scenes to fine-tune the existing counting model. It pays minimal attention on how to improve the counting performance of current scenes. Switching CNN uses a pre-calculated classifier to indicate which CNN column should be used for counting current patch. But it does not work well when counting the crowds of low density levels due to scale variation in current patch. CNN-pixel counting uses CNN-pixel and FCNN-skip networks to estimate a full-resolution density map. It is similar to our deconvolutions, which are utilized to compensate for the loss of detailed information in the early pooling. However, it pays no attention to different scales of the crowd

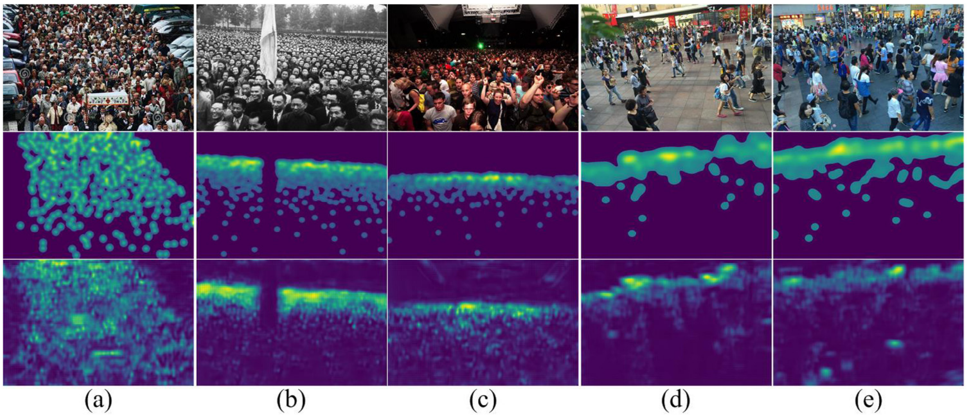

Fig. 11. Results of the proposed MMCNN method in Shanghai Tech dataset. First row: Input images. Middle row: Ground truth density maps. Last row: Estimated density maps. For different column: (a) VHHD: 75.3, G: 361, C: 326, E: 35↓; (b) VHHD: 230.8, G: 602, C: 691, E: 89↑; (c) VHHD: 260.1, G: 363, C: 348, E: 15↓; (d) VHHD: 255.1, G: 213, C: 208, E: 5↓; (e) VHHD: 450.7, G: 284, C: 269, E: 15↓.

Table 6
Comparisons with the state-of-the-arts in UCSD dataset.

	MAE	MSE
Cross-scene counting approach [33]	1.60(↑0.58)	3.31(↑2.13)
MCNN [28]	1.07(↑0.05)	1.35(↑0.17)
CNN-pixel counting [40]	1.12(↑0.10)	2.06(↑0.88)
Switching CNN [30]	1.62(↑0.60)	2.10(↑0.22)
The proposed approach	**1.02**	**1.18**

Table 7
Comparisons with the state-of-the-arts in MALL dataset.

	MAE	MSE
CNN-boosting [25]	2.01(↑0.03)	N/A
MoCNN [41]	2.75(↑0.77)	13.40(↑7.72)
Weighted V-LAD [42]	2.41(↑0.43)	9.12(↑3.44)
The proposed approach	**1.98**	**5.68**

image. Compared with the best approach mentioned above, namely, MCNN, our counting approach further improved the multi-scale strategy and added multi-task strategy. Counting results in Table 6 indicate that our approach outperforms the state-of-the-art methods in terms of MAE and MSE.

Table 7 shows the comparison results in MALL dataset, which contains crowds of low or medium density levels. CNN boosting proposed by Walach et al. [25], MoCNN proposed by Kumagai et al. [41], and weighted V-LAD proposed by Sheng et al. [42] are used for comparison. The MAE and MSE of these approaches are reported in their original work, except for the MSE of the CNN boosting approach. Among the approaches used for comparison, CNN boosting employs a boosting strategy to reduce the residual errors during training the counting model, thus achieves the lowest MAE. MoCNN uses adaptive integration of multiple CNNs that are specialized to a specific appearance for crowd counting. Weighted V-LAD builds LAF to explore the spatial context and local information of crowds, and the VLAD encoding method is used for counting. All of these approaches have special designs to improve counting performance. However, our approach outperforms them in terms of MAE and MSE due to our special designs of multi-scale and multi-task strategies.

Table 8 shows the comparison results in UCF_CC_50 dataset, which has extremely dense crowds. This dataset is commonly used to evaluate different crowd counting approaches, and we merely select a few results from typical studies. The cross-scene counting approach, MCNN, Hydra-CNN proposed by Onoro et al. [29], CNN pixel counting, CrowdNet proposed by Boominathan et al. [32], MSCNN proposed by Zeng et al. [31], Cascade-MTL proposed by Sindagi et al. [6], and switching CNN are used for comparison. Among the approaches mentioned above, cross-scene counting approach and CNN-pixel counting perform unsatisfactory because both of them pay no attention in scale issues or non-uniform density distribution of dense crowds. CrowdNet also performs poorly in this dataset because no obvious semantic information can be obtained due to severe occlusions. Hydra-CNN is similar to MCNN, but the former also down-samples the input data into different scales. Thus, Hydra-CNN achieves better counting performance than that of MCNN. Unlike them, MSCNN uses a single column with multi-scale blobs instead of multi-columns. Its counting performance is between MCNN and Hydra-CNN. Cascade-MTL estimates the crowd counts in a multi-task manner, which jointly estimates the density map and the density level. As shown in Table 8, Cascade-MTL counting, switching CNN, and our approach achieve similar performance in crowd counting. Switching CNN performs a little better than our approach in terms of MAE, but ours outperforms it greatly in terms of MSE, which indicates our robustness in counting extremely dense crowds. The reason is that there may exist multiple scales in the patch. Switching CNN performs well in counting patches with single scale, whereas ours performs evenly in all patches (single scale or multiple scales).

Table 8
Comparisons with the state-of-the-arts in UCF_CC_50 dataset.

	MAE	MSE
Cross-scene counting approach [33]	467.0(↑146.4)	498.5(↑174.7)
MCNN [28]	377.6(↑57.0)	509.1(↑185.3)
Hydra-CNN [29]	333.7(↑13.1)	425.2(↑101.4)
CNN-pixel counting [40]	406.2(↑85.6)	404.0(↑80.2)
CrowdNet [32]	452.5(↑131.9)	N/A
MSCNN [31]	363.7(↑43.1)	468.4(↑144.6)
Cascade-MTL [6]	322.8(↑2.2)	341.4(↑17.6)
Switching CNN [30]	**318.1**(↑2.5)	439.2(↑115.4)
The proposed approach	320.6	**323.8**

Table 9 shows the comparison results in WorldExpo '10 dataset. We employ the cross-scene counting approach, MCNN, CNN pixel counting, and switching CNN for comparison. This dataset is officially split into training and testing sets. We only list the MAE because only MAEs are reported in their original works. Among different approaches, our approach achieves the lowest MAE due to our special designs in handling scale variation and non-uniform density distribution. In addition, switching CNN has a similar counting performance.

Table 10 shows the comparison results in Shanghai Tech dataset, which has crowds of both medium density level (Part B) and high density level (Part A). The cross-scene counting approach, MCNN, FCN proposed by Marsden et al. [7], Cascade-MTL, MSCNN and switching CNN are used for comparison. Marsden et al. used an FCN structure for crowd

Table 9
Comparisons with the state-of-the-arts in WorldExpo10 dataset.

	MAE
Cross-scene counting approach [33]	12.9(↑3.8)
MCNN [28]	11.6(↑2.5)
CNN-pixel counting [40]	13.4(↑4.3)
Switching CNN [30]	9.4(↑0.3)
The proposed approach	**9.1**

counting. It is a shallow network with a relatively few parameters. However, it only outperforms cross-scene counting approach. MCNN and Cascade-MTL outperform the abovementioned two approaches. But their counting performance is inferior to that of switching CNN and our approach. Similar to the result in Table 8, switching CNN outperforms our approach slightly in term of MAE in Part A, which has extremely dense crowds. Our approach outperforms switching CNN in term of MSE in the same dataset, which indicates that our approach is more robust in counting extremely dense crowds. For Part B that has crowds of medium density level, our approach outperforms switching CNN in terms of MAE and MSE. MSCNN has the best counting performance in Shanghai Tech dataset due to its multi-scale input. Our counting performance is similar to that of MSCNN except for MAE in Part A. However, our approach is much superior to MSCNN in counting extremely dense crowds (UCF_CC_50) as shown in Table 8.

The comparisons with state-of-the-art CNN-based counting approaches in different benchmarking datasets show that our approach almost outperforms the others in terms of MAE and MSE when counting crowds with low and medium density levels. However, in datasets with extremely dense crowds (e.g., UCF_CC_50 and Shanghai Tech Part A), switching CNN has a slight advantage over our approach in term of MAE. The reason may be that switching CNN can select the most suitable CNN column for a patch with extremely dense crowds. These crowds always have relatively unique scales, whereas less dense crowds have drastic scale variation, which is more suitable to use our multi-scale strategy. However, extremely dense crowds occupy a large part of the global crowd count. Accurately estimating the number of people in these crowds is more likely to result in a small MAE. However, our approach outperforms switching CNN in term of MSE because jointly considering multiple scales by minimizing per-scale loss makes our approach more robust in crowd counting. The use of a multi-task strategy also prevents our approach from predicting a large count error. In general, the proposed MMCNN is comparable with state-of-the-art approaches, such as switching CNN, but demonstrates a more robust counting performance.

It is obvious that crowds in UCF_CC_50 are most challenging to count. Take the crowds in a concert as example, the ground truth count is 3406, whereas the estimated count is only 2097. It can be observed from Fig. 12, that the estimated density map (Fig. 12(c)) is very similar to its ground truth (Fig. 12(b)), except for the regions in the top-left corner (approximately represented by white rectangles). As shown in the original crowd image (Fig. 12(a)), people in these regions can hardly be observed by the naked eyes. Although our counting approach has special designs in handling drastic scale variation and non-uniform density distribution, it can hardly capture the detailed information in these regions. Other technologies, such as super-resolution [43], could handle such a problem, which will be study later.

Table 11
Computational complexity of different counting approaches.

	MCNN	Cascade-MTL	MMCNN
PARAMs	1.9 M	5.6 M	6.8 M
Forward time	2.0 ms	4.8 ms	4.9 ms
Backward time	1.9 ms	6.4 ms	6.3 ms

4.5. Computational complexity

The number of neural networks parameters (PARAMs) are used to evaluate the spatial computational complexity, which is obedient to $O(\sum_{l=1}^{D} k_l^2 \cdot C_{l-1} \cdot C_l)$ (D indicates the depth, k_l represents the kernel size of the lth filter, C_{l-1} represents the number of input kernels and C_l represents that of output kernels). In addition, PARAMs of MCNN and Cascade-MTL are used for comparison (Table 11). MMCNN and Cascade-MTL have greater PARAMs than that of MCNN due to deeper structure and the existence of fully connected layers. PARAMs of MMCNN is slightly greater than Cascade-MTL because the former is wider than the latter.

The input patch largely influence the temporal computational complexity, which is obedient to $O(\sum_{l=1}^{D} M_l^2 \cdot k_l^2 \cdot C_{l-1} \cdot C_l)$ (M_l represents the size of feature map of the lth filter). It is obvious that MMCNN and Cascade-MTL own much higher temporal computational complexity than that of MCNN due to deeper structure. We also calculate the forward and backward times of different approaches with the same patch (240 × 240 pixels). It is obvious that MMCNN and Cascade-MTL have similar running time for one iteration, and the time used by MCNN is much shorter. In consideration of the image size (960 × 960 pixels), patch size (240 × 240 pixels) and stride (20 pixels), density maps should be estimated in 1296 ($[(960-240)/20]^2$) patches. Thus, it takes about 14 s (1296 × 11 ms) to count a crowd image. Of course, it will take less time with bigger stride. Notably, our work mainly purchases better accuracy and robustness. Later, we will try to simplify our counting model with no loss of accuracy/robustness.

5. Conclusions

A CNN-based crowd-counting approach is proposed in this work. The proposed approach focuses on two challenges in crowd counting: drastic scale variation and non-uniform density distribution. We propose the use of a combined density map, which is proven to be more effective than the traditional Gaussian density map and the recently used human-shaped density map. Then, a multi-column CNN is designed to resolve the problem of drastic scale variation. Up- and down-sampled data are fed to imported into the multi-column CNN, and features of different scales are reliably extracted by minimizing the per-scale loss. To resolve the problem of non-uniform density distribution, we jointly estimate three objectives, namely, crowd density map, BG/FG mask, and density level. Evaluations in five benchmarking datasets verify the effectiveness of our multi-scale and multi-task strategies. Comparisons with state-of-the-art CNN-based crowd counting approaches indicate the superiority of the proposed counting approach. Our future work will focus on embedding other tasks (e.g. crowd anomaly detection) into current framework. We also plan to further study the employment of super-resolution to handle the limitation proposed in Fig. 12.

Table 10
Comparisons with the state-of-the-arts in Shanghai Tech dataset.

	Part A		Part B	
	MAE	MSE	MAE	MSE
Cross-scene counting approach [33]	181.8(↑90.6)	277.7(↑149.1)	32.0(↑13.5)	49.8(↑20.5)
MCNN [28]	110.2(↑19.0)	173.2(↑44.6)	26.4(↑7.9)	41.3(↑12.0)
FCN [7]	126.5(↑35.3)	173.5(↑44.9)	23.7(↑5.2)	33.1(↑3.8)
Cascade-MTL [6]	101.3(↑10.1)	152.4(↑23.8)	20.0(↑1.5)	31.1(↑1.8)
Switching CNN [30]	**90.4**(↓0.8)	135.0(↑6.4)	21.6(↑3.1)	33.4(↑4.1)
MSCNN [31]	**83.8**(↓7.4)	**127.4**(↓1.2)	**17.7**(↓0.8)	30.2(↑0.9)
The proposed approach	91.2	**128.6**	**18.5**	**29.3**

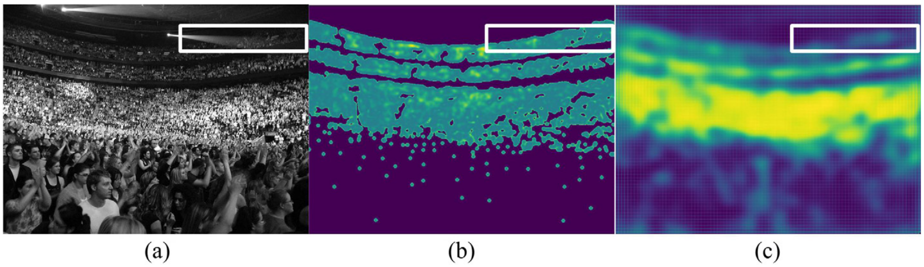

Fig. 12. Limitations of the proposed counting approach. (a) Input image, (b) ground truth density map, and (c) estimated density map.

Acknowledgments

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. Dollar, C. Wojek, B. Schiele, et al., Pedestrian detection: an evaluation of the state of the art, IEEE Trans. Pattern Anal. Mach. Intell. 34 (4) (2012) 743–761.

[2] P.F. Felzenszwalb, R.B. Girshick, D. McAllester, et al., Object detection with discriminatively trained part-based models, IEEE Trans. Pattern Anal. Mach. Intell. 32 (9) (2010) 1627–1645.

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

[4] A.B. Chan, M. Morrow, N. Vasconcelos, Analysis of crowded scenes using holistic properties, in: Performance Evaluation of Tracking and Surveillance Workshop at CVPR, 2009, pp. 101–108.

[5] V.A. Sindagi, V.M. Patel, A survey of recent advances in CNN-based single image crowd counting and density estimation, Pattern Recognit. Lett. (2017).

[6] V.A. Sindagi, V.M. Patel, CNN-based cascaded multi-task learning of high-level prior and density estimation for crowd counting, 2017. arXiv preprint arXiv:1707.09605.

[7] M. Marsden, K. McGuiness, S. Little, et al., Fully convolutional crowd counting on highly congested scenes, 2016. arXiv preprint arXiv:1612.00220.

[8] P. Sabzmeydani, G. Mori, Detecting pedestrians by learning shapelet features, in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2007, pp. 1–8.

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

[10] C. Gao, J. Liu, Q. Feng, et al., People-flow counting in complex environments by combining depth and color information, Multimedia Tools Appl. 75 (15) (2016) 9315–9331.

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

[12] H. Fradi, J. Dugelay, Low level crowd analysis using frame-wise normalized feature for people counting, in: International Workshop on Information Forensics and Security, WIFS, 2012, pp. 246–251.

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

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

[15] Y. Li, E. Zhu, X. Zhu, et al., Counting pedestrian with mixed features and extreme learning machine, Cogn. Comput. 6 (3) (2014) 462–476.

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

[17] X. Zhang, H. He, S. Cao, et al., Flow field texture representation-based motion segmentation for crowd counting, Mach. Vis. Appl. 26 (7–8) (2015) 871–883.

[18] 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, 2013, pp. 2467–2474.

[19] M.A. Mousse, C. Motamed, E.C. Ezin, People counting via multiple views using a fast information fusion approach, Multimedia Tools Appl. 76 (5) (2017) 6801–6819.

[20] S. Ren, K. He, R. Girshick, et al., Faster R-CNN: Towards real-time object detection with region proposal networks, in: Advances in Neural Information Processing Systems, 2015, pp. 91–99.

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

[22] M. Fu, P. Xu, X. Li, et al., Fast crowd density estimation with convolutional neural networks, Eng. Appl. Artif. Intell. 43 (2015) 81–88.

[23] S. Pu, T. Song, Y. Zhang, et al., Estimation of crowd density in surveillance scenes based on deep convolutional neural network, Procedia Comput. Sci. 111 (2017) 154–159.

[24] C. Wang, H. Zhang, L. Yang, et al., Deep people counting in extremely dense crowds, in: Proceedings of the 23rd ACM International Conference on Multimedia, 2015, pp. 1299–1302.

[25] E. Walach, L. Wolf, Learning to count with CNN boosting, in: European Conference on Computer Vision, Springer International Publishing, 2016, pp. 660–676.

[26] V.A. Sindagi, V.M. Patel, Generating high-quality crowd density maps using contextual pyramid CNNS., in: 2017 IEEE International Conference on Computer Vision, ICCV, IEEE, 2017, pp. 1879–1888.

[27] F. Xiong, X. Shi, ., D.Y. Yeung, Spatiotemporal modeling for crowd counting in videos, in: 2017 IEEE International Conference on Computer Vision, ICCV, IEEE, 2017, pp. 5161–5169.

[28] 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, 2016. pp. 589–597.

[29] D. Onoro-Rubio, R.J. Lopez-Sastre, Towards perspective-free object counting with deep learning, in: European Conference on Computer Vision, Springer International Publishing, 2016, pp. 615–629.

[30] D.B. Sam, S. Surya, R.V. Babu, Switching convolutional neural network for crowd counting, 2017. arXiv preprint arXiv:1708.00199.

[31] L. Zeng, X. Xu, B. Cai, et al., Multi-scale convolutional neural networks for crowd counting, 2017. arXiv preprint arXiv:1702.02359.

[32] L. Boominathan, S.S.S. Kruthiventi, R.V. Babu, Crowdnet: A deep convolutional network for dense crowd counting, in: Proceedings of the 2016 ACM on Multimedia Conference, ACM, 2016, pp. 640–644.

[33] 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, 2015, 833–841.

[34] V. Lempitsky, A. Zisserman, Learning to count objects in images, in: Advances in Neural Information Processing Systems, 2010, pp. 1324–1332.

[35] 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, 2014, pp. 346–361.

[36] L.C. Chen, Y. Yang, et al., Attention to scale: Scale-aware semantic image segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 3640–3649.

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

[38] Hal Daum III, Frustratingly Easy Domain Adaptation, ACL, 2009.

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

[40] D. Kang, D. Dhar, A.B. Chan, Crowd Counting by Adapting Convolutional Neural Networks with Side Information, 2016.

[41] S. Kumagai, K. Hotta, Mixture of counting CNNS: Adaptive integration of CNNS specialized to specific appearance for crowd counting, 2017. arXiv preprint arXiv:1703.09393.

[42] B. Sheng, C. Shen, G. Lin, et al., Crowd counting via weighted vlad on dense attribute feature maps, IEEE Trans. Circuits Syst. Video Technol. 99 (2016) 1–11.

[43] Y. Xian, Z.I. Petrou, Y. Tian, et al., Super-resolved fine-scale sea ice motion tracking, IEEE Trans. Geosci. Remote Sens. (2017).