Neurocomputing 472 (2022) 224–251

Contents lists available at ScienceDirect

Neurocomputing

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

A survey of crowd counting and density estimation based on convolutional neural network

Zizhu Fan [a], Hong Zhang [a], Zheng Zhang [b,d,*], Guangming Lu [b], Yudong Zhang [c], Yaowei Wang [d]

[a] School of Basic Science, East China Jiaotong University, Nanchang 330013, China
[b] Harbin Institute of Technology, Shenzhen 518055, China
[c] School of Informatics, University of Leicester, Leicester LE1 7RH, UK
[d] Peng Cheng Laboratory, Shenzhen 518055, China

ARTICLE INFO

Article history:
Received 5 November 2020
Revised 17 January 2021
Accepted 1 February 2021
Available online 8 November 2021
Communicated by Zidong Wang

Keywords:
Crowd counting
Crowd density estimation
Convolutional neural network
Deep learning

ABSTRACT

Crowd counting and crowd density estimation methods are of great significance in the field of public security. Estimating crowd density and counting from single image or video frame has become an essential part of a computer vision system in various scenarios. In this paper, we comprehensively review the recent research advancement on crowd counting and density estimation. First of all, we introduce the background of crowd counting and crowd density estimation. Second, the traditional crowd counting methods are summarized. Third, we focus on reviewing the crowd counting and crowd density methods based on convolutional neural network (CNN) models. Next, we report and discuss the experimental results of a number of typical methods on benchmark datasets. Finally, we present the promising future directions of crowd counting and crowd density.

1. Introduction

Crowd counting is to estimate the number of people in an image or video frame by using a counting method, and the crowd density estimation is to convert the input crowd image into its corresponding density map. The total number of people can be obtained by integrating the density map that indicates the number of people per pixel in the image [32]. Literature [79] used the following five levels of crowd density: very high density, high density, medium density, low density, and very low density to evaluate the crowding degree. There are many researchers who investigate the crowd counting and crowd density estimation methods, since these methods have important applications in many research fields.

1) Public safety management. Due to the development of society and city, the crowd density is increasing in shopping malls, subway stations, tourist attractions, gymnasiums and other large-scale gathering places (as shown in Fig. 1). In this case, it is easy to lead to safety accidents, such as the stampede on the Bund of Shanghai on December 31, 2014, and the serious stampede on November 22, 2010 in Phnom Penh, Cambodia, which killed more than 500 people and injured more than 700 people. These accidents are all crowded stampedes caused by insufficient preparation for mass activities, poor on-site management and improper handling. They lead to major casualties and bad social influence. If we can make an accurate analysis of the degree of the congestion [60,64,65,75,147,158] and crowd behavior [55,137], we can effectively prevent such incidents.
2) Public space design. It is necessary to analyze the density of customers in the shopping mall, and evaluate the degree of the customers' interest in a certain product. Then we can optimize the allocation of resources, allocate service personnel and goods, and improve the service quality of shopping malls, so as to develop efficient marketing strategies. In addition, by counting the inbound and outbound passenger flow of shopping malls, stations, airports, etc. [66], we can know whether the setting of entrances and exits is reasonable. Optimizing this setting can not only improve the flow of people and improve the safety, but also save people's waiting time [68,69,201].
3) Intelligent monitoring. In the past, the crowd monitoring was usually done manually, and the users needed to observe multiple monitoring TVs for a long time. The long-time observation would make the users tired and lead to unnecessary security risks. If the intelligent monitoring system can give safety warning when the crowd density is too high,

* Corresponding author.
E-mail address: darrenzz219@gmail.com (Z. Zhang).

https://doi.org/10.1016/j.neucom.2021.02.103
0925-2312/

Fig. 1. Dense crowd scenario from UCF_CC_50 dataset [23].

it can effectively prevent some problems [63]. In the school, automatically counting the students of each class through video monitoring can save the time of teachers, and know the student's study status in real time.

4) Person search and crowd simulation. In some crowded crime places, we need to detect [148,165,167,189] suspects in the crowd and track them. Using face detection algorithm to search suspects in a wide range of scenes is more efficient than traditional detection methods [76]. In addition, crowd scene can also be applied to crowd simulation technology.
5) Construction and management of smart buildings and smart parks. With the development of intelligence technology, intelligent technology can be used as components in the construction of smart buildings and smart parks to improve the performance of buildings. Crowd counting technology is an important part of intelligence technology, which can help buildings and parks operate efficiently and create a comfortable environment [72].

In addition to the above applications, many tasks are also related to the crowd counting, such as cell counting [10,14,43,48,50], vehicle counting [24,44,49,114,115,128], animal migration detection [56], environmental investigation [77,78], scene understanding [9,54,55,139], abnormality detection [61,62,123,138] and medical image analysis [155].

In recent years, some important progress has been made in the field of crowd counting and crowd density estimation, but there are still many challenging problems. With the increase of crowd density, there will be serious occlusion in the human body. In addition, the complexity of the background environment of the image, such as noise, uneven light and human body deformation, will affect the final detection results. To solve these problems, researchers have proposed many estimation methods for different problems, e.g., crowd counting based on detection method [6,7,36–39,129,130,133,134,149,163,178,179,185], crowd counting based on regression method [8,10,26,40,124,131,135], and density estimation method [13,14,107]. As shown in Fig. 2, the detection-based method uses bounding boxes to accurately locate each person in the image, and the regression-based method directly outputs the number of people in an image or the density map of the corresponding image. However, in the complex environment or larger crowded scenes, traditional methods are often unable to achieve accurate estimation. In recent years, CNN has been widely used in a great number of applications (image recognition, object detection, image segmentation, etc.) [126]. Due to CNN's strong learning ability, it can learn more feature information, which makes it perform well in the field of crowd counting. Different from the traditional methods, the convolution neural network method can extract high-level features conveniently and efficiently, and consequently obtain better performance than the traditional methods. For the dense crowd area in the image, the CNN method is used to get better prediction results. At present, the crowd counting method based on CNN [12,28,29,44,51,86] has become more popular than before (see Fig. 3).

In previous works, literatures [34,125] have studied and analyzed the traditional crowd counting methods. Loy et al. [34] classified the traditional crowd counting methods into three categories, i.e., the detection-based, regression-based, and density estimation-based approaches. Moreover, they evaluated the methods based on image and video, and analyzed the advantages and disadvantages of each method. Ryan et al. [125] focused on evaluating regression models with different features such as the local, global, and histogram features, and compared the effects of different features on regression results. Zitouni et al. [87] investigated and analyzed general aspects of techniques of crowd counting instead of specific algorithm. Li et al. [75] and Saleh et al. [73]

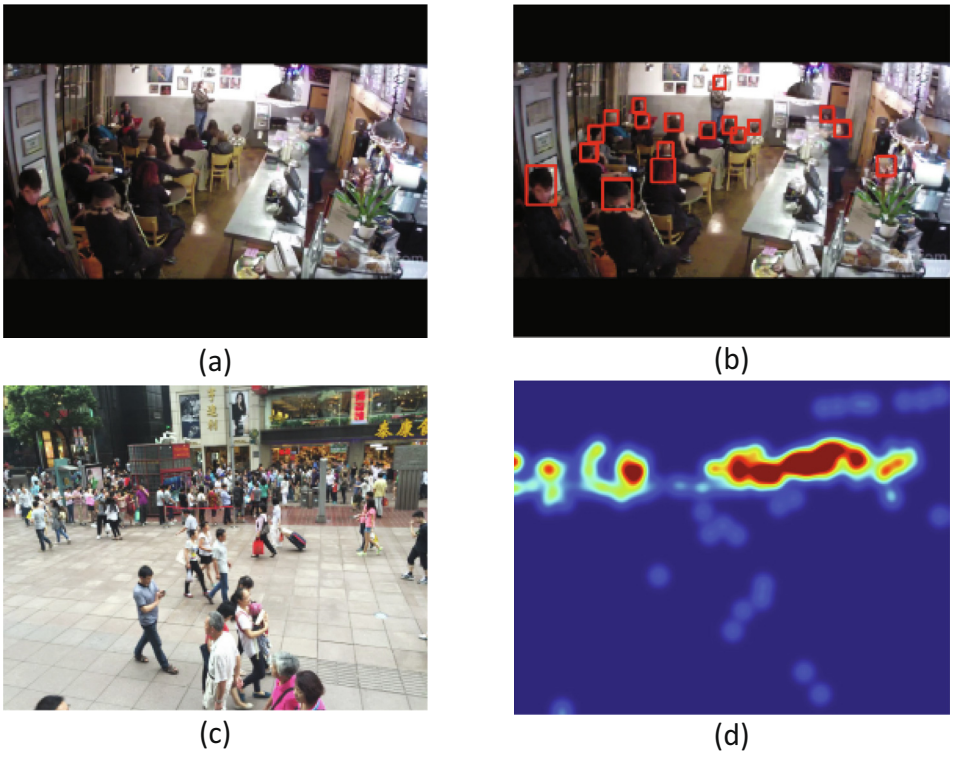


Fig. 2. Examples of detection-based and density-based methods. (a) the input image for the detection method, (b) the detection result of image (a), (c) the input image for the density map estimation method, (d) the density map corresponding to the input image in (c).

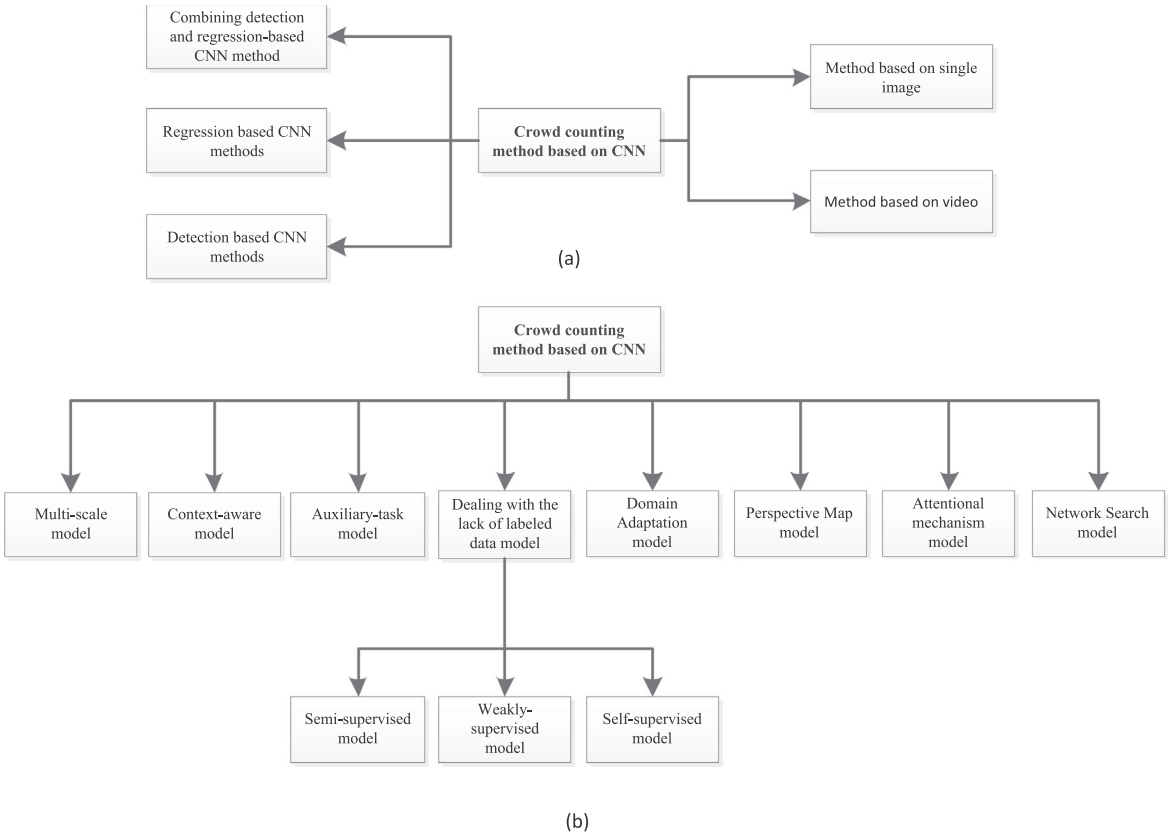


Fig. 3. Categorization of crowd counting method based on CNN. (a) The left side represents the categorization based on the mode of counting implementation, and the right side represents the categorization of the types of training data. (b) Categorization according to network attributes.

reviewed crowd counting and density estimation methods under visual monitoring. Kang et al. [49] analyzed the CNN-based density estimation methods for crowd counting, detection and tracking. Sindagi et al. [32] summarized the existing methods for crowd counting and density estimation in a single image, but they did not analyze detection-based methods [17,67,95], the methods based on video [95] and methods based on detection and regression [29,85,192]. In our work, we systematically and comprehensively analyzed the crowd counting and density estimation methods based on CNN, and also discussed the traditional methods that were not analyzed before.

The structure of this survey is organized as follows: the Section 2 describes the traditional methods of crowd counting, including direct detection method and indirect method. The Section 3 introduces the latest methods of crowd counting based on CNN, and systematically analyzes and summarizes them. The Section 4 introduces the relevant datasets and evaluation standards of crowd counting task, and reports the results of various methods on each dataset. The Section 5 summarizes this work and describes some future works in the related fields.

2. Traditional crowd counting methods

The traditional crowd counting method can be divided into two categories, i.e., direct detection methods and indirect methods. The direct detection method is to detect the objects and count the dense crowd. The indirect method is to take the whole crowd as an object, analyze the features of the crowd and then establish a mapping to the number of the dense crowd to indirectly calculate the number of the crowd. Since this paper focused on crowd counting based on CNN, traditional methods are only briefly introduced.

2.1. Direct detection method

Direct detection methods are divided into two classes. The first class is based on the detection of human's overall features [3–5,36]. This type of method mainly trains a classifier, e.g., Support Vector Machine (SVM), boosting or random forest [2], which is combined with the extracted features, such as Haar wavelet features [35,180], histogram oriented gradients (HOG) [36], edgelet [37], and shapelet [38], to detect the crowd. Wojek et al. [1] used a sliding detector to detect people in images and the total number of people in the scene can be counted. Methods based on the holistic detection have achieved good results under the condition of the relatively sparse population. However, in the case of high crowd density the performance is often poorer due to the serious occlusion between people which usually limits the use of the algorithm. In order to solve this problem, researchers proposed detection methods using the parts of the body. Literatures [6,7,27,39,136] mainly counted the number of people by detecting heads and shoulders. This method is slightly better than holistic detection. The idea of detection based on heads can be used as a reference for many other methods, typically not only the methods based on detection but also the ones based on density maps.

2.2. Indirect method

For the direct detection based on the whole body or the parts of the body, it is difficult to deal with the severe occlusion among crowds. In order to solve this problem, indirect method is used in the crowd counting. The main advantage of indirect method is to avoid the complex pedestrian detection process and avoid relying on the learning detector. The indirect method mainly learns a mapping from image to the total number of people or the corresponding density map of input image [8,10,100,145]. The procedure of this method contains two steps. The first step is to extract the features of the low level, which includes global features (such as the area of the region [40,102], perimeter and area perimeter ratio), local features (edge features [33,105,124,125,181], texture features [23,106], etc.). The second step is to learn a regression model, (linear regression [40,105], piecewise linear regression, ridge regression [10] and Gaussian regression [26], etc.). After extracting the features of the low level, different regression models are used to learn the mapping from low-level features to the number of people.

Davies et al. [33] found that there was an approximate linear relationship between pixel-features and the total number of people in images. They obtained the foreground image through the three-frame difference method. Then they counted the number of pedestrians in each frame using the manual method and established a linear equation to obtain the corresponding linear relationship. This approach is inapplicable when the number of the crowd is very large. However, as the crowd density becomes higher, the crowd occlusion becomes more serious, and the scene becomes more complex. It is impossible to accurately estimate the number of people with any single feature. In order to solve this problem, Idrees et al. [23] estimated the number of people using a variety of information sources. They first used three different and complementary methods, i.e., the method based on Fourier analysis, the method based on the detection of heads, the method based on the count of the interest points, to count image patches, respectively. Then the estimation result is globally constrained in a multi-scale Markov Random Field (MRF) framework. Consequently, they obtained the number of people in the whole image.

Although the regression-based method can solve the problems of occlusion and complex background to a certain extent, it often ignores the spatial information of the image. Arteta et al. [50] applied density map to crowd counting for the first time and introduced how to convert manually annotated images into density maps. Compared with regression-based methods, the density map method can also show the congestion of the crowd, which has better applications in some cases. Many crowd counting methods are based on density map.

Since unbalanced datasets significantly reduce the performance of the video-based crowd counting (VCC) method, inspired by the success of using unbalanced datasets in image classification [47], Huang et al. [99] proposed a new cost sensitive sparse linear regression VCC method (CS-SLR-VCC). They first learned a sparse linear regression (SLR) model, and calculated the modeling error related to each training sample. In order to eliminate the adverse effects of data imbalance, all modeling errors were taken as the prior knowledge of relevant weighting factors of samples. Then, a cost sensitive SLR model was rebuilt, to obtain the optimal solution. Since both Gaussian regression and ridge regression models are expensive to calculate, they can only deal with a few features. Xu et al. [101] suggested that the use of more image features could improve the performance of crowd detection approaches. They proposed to estimate crowd density using a comprehensive feature set containing many features. In this paper, random forest regression model is combined with random projection method to embed random projection in tree nodes and introduce randomness into tree structure. First, it applied multi-layer background subtraction technique [195] to learn the foreground object from the region of interest (ROI), and then used the method proposed in [26] to correct the perspective deformation.

Pham et al. [13] established the mapping between the image patch features and the relative positions of all the objects in the image patch. In this way, spatial information was added in the learning process, which played a key role in generating the density map of the local image patch through Gaussian kernel density esti-

mation. They used a random forest to learn the relationship between images and their corresponding density maps. Wang and Zou [14] thought the method of integrating the density map to calculate the total number of people is effective, but the efficiency is very low. In order to solve this problem, they proposed a density estimation based visual object counting method (DEVOC), which uses subspace learning to accelerate the algorithm speed. Instead of learning the mapping between the global image features and the corresponding count, they learned the mapping between the features of the local image patches and the subspace formed by the corresponding local density maps. This method is as accurate as the mainstream methods. Moreover, it is faster than many mainstream methods.

3. Crowd counting method based on CNN

The main reason for the success of CNN in crowd counting and crowd density estimation task is its ability to learn nonlinear relations. It is more accurate to use CNN to learn the nonlinear relationship between images and the number of people in images or their corresponding density map. Wang et al. [41] used CNN in the task of crowd counting and crowd density estimation for the first time. Inspired by the success of CNN in image classification [31], they proposed a CNN model to count people in the region of interest (ROI) in the image. They replaced the last full connection layer with a single neuron to predict the counting by using the AlexNet. In addition, in order to reduce the error rate, they used expanded negative samples to enrich the training data, and achieved better results than the traditional methods.

However, with the increasing density of the crowd, the occlusion within the crowd is becoming more serious. It cannot get good results via simply using CNN to extract features. In order to solve this problem, Shang et al. [46] used a CNN, which took the whole image as an input network and directly outputs the counting. Aiming to address more noise and less target area, they used the context information to predict the counting of whole image and the local area. Literatures [12,15,90] tried to combine the multi-scale information of the crowd to enable the network detecting objects of various scales. On the other hand, researchers [20,21,25,112,122,161] found that different but related tasks can improve the performance of each other. They used multi-task learning to promote the performance of crowd counting.

In addition, many methods try to solve the problem of more practical crowd counting via different schemes. The existing CNN-based methods still have many parameters and require a lot of computing resources, which limits their practical applications, especially in embedded devices with limited memory. Inspired by the recursive convolution network [144], Ding et al. [143] proposed a deep recursive network structure. When the effect is equivalent, the network parameters are less, which can be better applied in the real time scenarios. Literatures [30,95,197] found that on the video dataset, capturing the time information between adjacent frames in the training can get better counting results. In many scenarios, labeled data are difficult to obtain. In order to solve this problem, literatures [20,25] using an auxiliary task which sorts the unmarked data to improve the performance of the network when the amount of labeled data is small.

According to the different types of training data, the methods used CNN are divided into two categories: 1) the method based on single image, 2) the method based on video. According to the training dataset used by the network and the output of the network, we categorize CNN based crowd counting methods into three categories:

(1) Detection based CNN method. This type of methods trains the image dataset annotated by the bounding box. During the test, the network can accurately detect every object in the image.
(2) Regression based CNN method. This method is trained using an image dataset annotated by point or using unsupervised methods. During testing, the network directly outputs the total number of people in the image or generates the density map corresponding to the image.
(3) Combining detection and regression-based CNN method. The detection-based CNN method and the regression-based CNN method are combined in a network for counting and positioning.

3.1. Detection based CNN methods

The detection-based method can accurately locate and count the objects in the image, which is very important in some scenarios.

Stewart et al. [17] proposed an end-to-end training network structure. They transformed each image into 1024 dimensional features through GoogLeNet, which contains rich location information. Then LSTM [11] as a controller, maps these features into a series of detection boxes. These bounding boxes are generated in descending confidence. When LSTM cannot find any detection box with confidence that is larger than the predetermined threshold, LSTM will be stop. They did not use non maximum suppression (NMS) as the successive step, because NMS only processes the bounding box without using the image information. This method is only applicable to the situation that the detection object has almost no overlap, and the effect is very poor when the overlap is high. They achieved good results by using recurrent neural networks to avoid multiple predictions of the same target. Li et al. [67] proposed an adaptive relational network which called HeadNet to extract context information to suppress missed detection. The network structure is shown in Fig. 4. The feature extraction network Resnet-101 was adopted to extract the features of the input image, and local structured feature modules are used to learn individual stability. Locally structured feature modules include S_{as}, S_{ar} and S_{lp}. Then the global adaptive module is used for encodes the pre-quantified intergroup conflict Z_i. Finally, the network outputs the confidence and bounding box.

3.2. Regression based CNN methods

Detection methods based on regression are divided into two categories: 1) regression counting method and 2) regression density map method. The regression counting method outputs the number of people in an image directly. The regression density map can get the counting and estimate the crowd density. Compared with detection-based CNN methods, generating density map based methods can get better prediction results in the higher density scenarios. Herein, we will discuss from the following aspects,

1) Multi-scale model. These methods mainly extract the information of different scales in the image through multi-column structure or feature pyramid network or scaling, which improves the robustness of scale aware.
2) Context-aware model. These methods add local and global context information to the network to improve the detection accuracy.
3) Auxiliary-task model. One or more tasks related to crowd counting are added into the network as auxiliary tasks to train together.

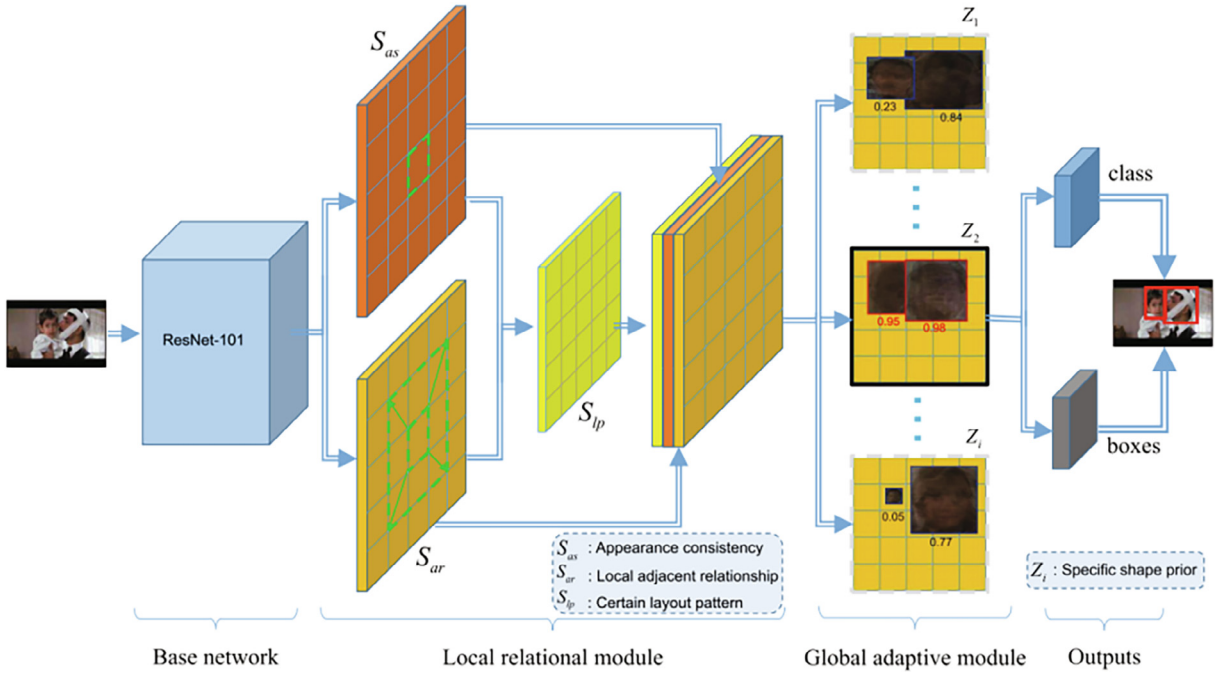


Fig. 4. network structure proposed by Li et al. [17].

4) Dealing with the lack of labeled data model. Furthermore, we divide the model into semi-supervised model, weakly-supervised model and self-supervised model.
5) Domain Adaptation model. These methods can count in any object domain.
6) Perspective Map model. Perspective map reflect the perspective distortion at each position in the image (i.e. how many pixels correspond to a person height of one meter). These methods add perspective map to the network to assist in generating density map.
7) Attentional mechanism model. These methods add attention mechanism to the network to improve the crowd counting performance.
8) Network Search model. These methods obtain network structure for crowd counting task by NAS (instead of hand-crafted the network).

3.2.1. Multi-scale model

Due to the increasing complexity of counting datasets, many models may lead to low performance when handle large-scale changes in crowd and high crowd density. A growing number of models are designed to extract multi-scale information from images. Boominathan et al. [42] proposed a network which was composed of deep network and shallow network to extract different scale information. The deep network is mainly used to capture high-dimensional information such as face and body. The shallow network is used to capture the head far away from the camera which is smaller. Deep network uses the VGG network. Boominathan removed the fifth pooling layer, which makes the density map of the network output is one eighth of the original map. The shallow network consists of three convolution layers. The convolution kernel size is 5×5. To ensure the maxpooling without loss count, this work used the average pooling in the shallow network. Then, the deep and shallow feature maps are fused, and the convolution layer of 1×1 is used. Finally, the density map is obtained by sampling the feature map to the original image size.

Similar to [42], Zhang et al. [12] proposed a multi-column CNN (MCNN) to capture multi-scale head features. The network structure is shown in Fig. 5 in which three networks are used to respectively extract different features of the crowd image. Finally, the features of the three scales are fused by 1×1 convolution layer. Because different branches have different receptive fields, different sizes of heads can be detected. However, this kind of model using multiple networks has many parameters and large amount of calculation, so it is impossible to predict the real-time crowd counting [27]. Moreover, some branches of multi array network are inefficient in many datasets, which cannot extract different scale head features. Sam et al. [15] also used the idea of three subnetworks. But unlike [12], they used pre-training. First, the input image is parted into nine patches, each of which is entered into the subnetwork in each column for pre-training. The image patches are divided into three categories by minimize the training error in that column. Three categories of image are used to train a Switch-CNN. The Switch-CNN is used to classify each image patch into the correct subnetwork, and the error of image patch in that subnetwork is the smallest. Finally, the accurate prediction of all image patches constitutes the accurate crowd estimation of the original image and achieves better results than MCNN [12]. However, the model faces the same problem as MCNN. That is, the model has more parameters and large amount of calculation, so it cannot predict the real-time crowd counting.

Inspired by [45], Onoro et al. [44] designed a scale-aware counting model, i.e., Hydra CNN, which can estimate the density of objects in different crowded scenarios without explicit scene information and perspective. Hydra CNN consists of several feature extractors and a decoder. Each extractor extracts different scale image patch features. All features are connected and the decoder outputs these features as density map. Zeng et al. [90] considered that it is complex and difficult to use multi-column network to extract multi-scale features. Therefore, they proposed a new multi-scale convolutional neural network (MSCNN) for crowd counting. By introducing multi-scale blobs (MSB) similar to naive perception module [108], the network can extract multi-scale features in a single-column structure, with high crowd counting performance and low calculation cost. MSCNN structure is shown in Fig. 6. The main structure of MSCNN includes three parts: feature

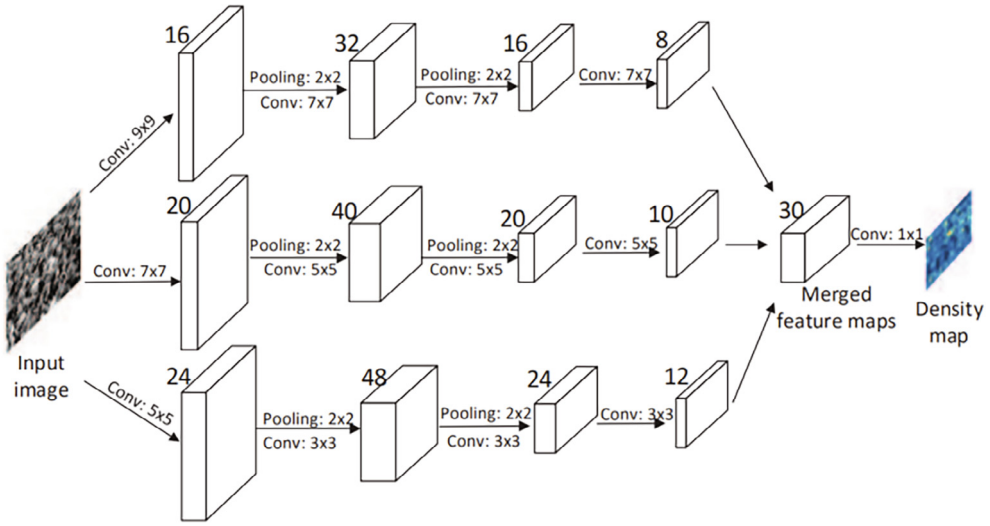


Fig. 5. Multi-column CNN structure proposed by Zhang et al. [12].

remapping, multi-scale feature extraction and density map regression. The first convolution layer is a common convolution layer with a single kernel, which is applied to extract image features. The MSB is used to extract multi-scale features. It contains multiple filters and multi-layer perceptron (MLP) [109] with different kernels sizes (9×9, 7×7, 5×5, 3×3). MLP convolution layer is a pixel level fully connected layer, which contains 11 filters for regression of the density map. Because the parameters of single column network structure (MSCNN) are less than those in the previous network, it is better than the previous method in terms of the training time. Moreover, compared with other methods, MSCNN has higher precision and better robustness.

The CSRNet proposed by Zhang et al. [18] does not adopt the multi-column mode. The structure advantage of Multi-Column is not significant compared with that of single-column. It is difficult to train. CSRNet model includes front-end and back-end networks. Similar to [15,22,42], the front-end network uses the pretrained VGG16 network that removes the full connection layer, which is used to extract image features. The size of output features is eighth of the original input image. The back-end network utilizes an extended convolutional neural network layer to expand the sensing domain while maintains the resolution, and generates high-quality crowd distribution density map. Kang et al. [51] consider that although the crowd counting based on density map has made great progress in the past few years, it still has no good effect in the face of serious occlusion and large-scale change. Because of the scale change, the size of the object usually changes dramatically. Different from MCNN which extracts features of different scales by increasing the number of CNN columns, so as to detect different sizes of human heads. The network structure is shown in Fig. 7, images of different scales are sent to the complete convolution network to get the density map of corresponding scale. On each scale, the subnetwork changes the input feature map into an attention map, and then uses a softmax function across all scales to multiply by the density of each scale. Finally, the convolution kernel of 1×1 is used to fuse all density map to get the final density map.

Zhang et al. [132] developed a single-column scale-adaptive CNN (SaCNN), which combines multi-layer feature maps to adapt the changes of scale. In addition, they introduced density map loss and relative counting loss to jointly optimize the model. The relative counting loss is helpful to reduce the prediction error, which improves the generalization ability of the network in the face of unfamiliar crowd scenes. The loss of density map is helpful to generate high quality density map. Sang et al. [136] improved the SaCNN method proposed in [132]. By analyzing the head size estimation parameters of geometric adaptive Gaussian kernel in SaCNN, they found that this head size estimation method has a large error in sparse scenes. They optimized this head size estimation to obtain better density map and more accurate head size estimation. The absolute counting loss function and density map loss are combined to enhance the network generalization ability under the scenario with fewer pedestrians.

In order to adapt to the great changes of scale, Zhang et al. [132] used a small fixed filter to extract multi-scale features in a single column network structure. But Zou et al. [128] think that Zhang's approach is not feasible, since there are many connection methods, which need to carefully design the intermediate fusion process, and it ignores the potential of some later layers. To this end, Zou

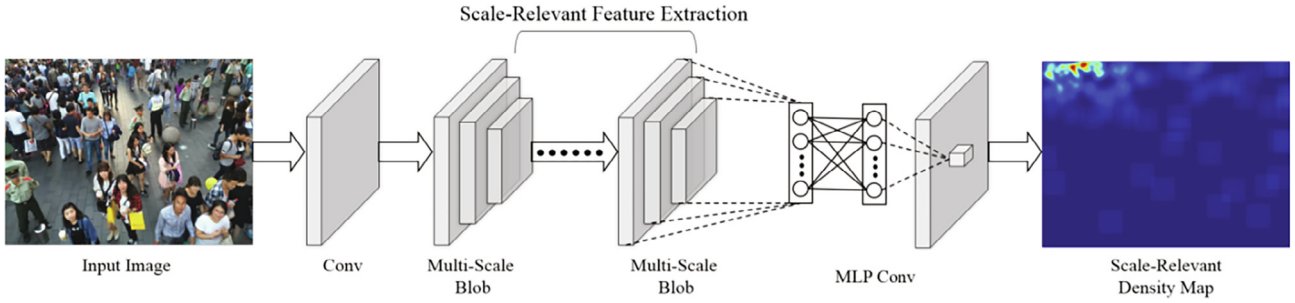


Fig. 6. Multi-scale convolutional neural network for crowd counting proposed by Zeng et al. [90].

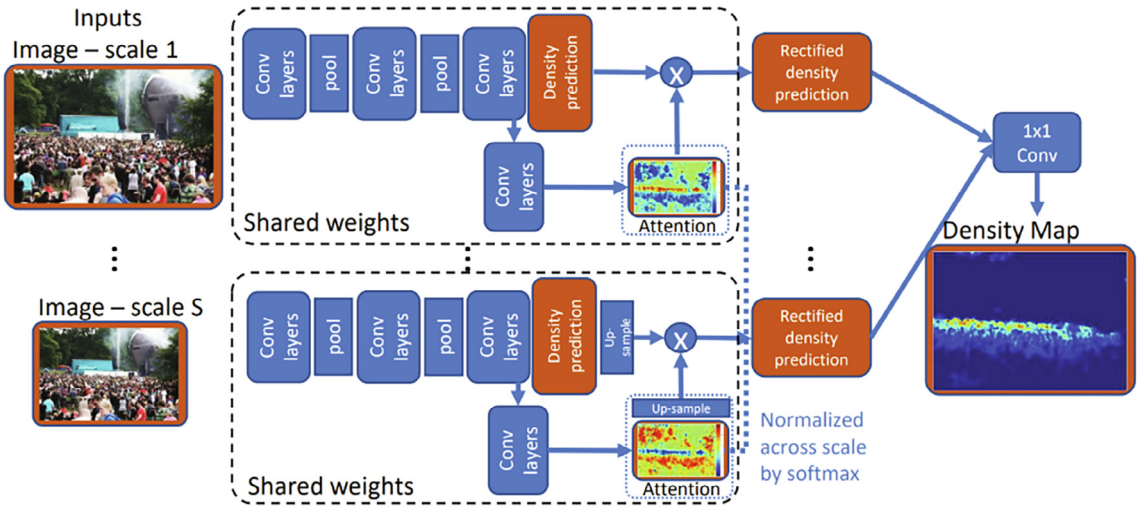


Fig. 7. Crowd counting network structure of image pyramid proposed by Kang et al. [51].

et al. [128] proposed a kind of Deformation Aggregation Network (DANet). In order to address the change of scales, the authors further used deformable convolution to achieve accurate location by increasing the offset of spatial sampling points, which can transform invariable features in different scenes. The structure of DANet is shown in Fig. 8. The whole architecture is composed of two parts: one is the backbone composed of eight blocks, and the other is multi-layer aggregation using deformation blocks to obtain offset information and adjustable weight, and then performs adaptive fusion to obtain the final density map.

In order to address perspective and scale change, Yang et al. [146] proposed a network called MS-GAN that consists of two subnetworks: a generator and a discriminator. The multi-scale generator predicts the density map, and the discriminator refines the generated density map. The generator uses a fully convoluted network, which consists of four convolution blocks, each of which contains two convolution layers, and three inception models used to extract multi-scale features. Finally, the fully connected layer is used to generate the final density map. Similar to [146], Cao et al. [170] also used modules similar to the inception architecture to extract multi-scale features, and each of these convolutional layers has a different size of convolution kernel. Finally, the final density map is obtained by deconvolution. Deb et al. [154] advocated adding the dilated filter [155] to the multicolumn convolution neural network, which would greatly improve the ability of the network to obtain multi-scale information without using perspective during training and testing. The dilated filter can effectively use multi-scale information by exponentially increasing the acceptance domain of the network without increasing the parameters.

Most previous crowd counting methods extract different features through different convolution kernels. Only L2 loss is used to optimize the model. Moreover, there is no interation between different convolution subnetworks, just trying to minimize their own errors. This leads to poor performance on other scales and no pursuit of scale consistency. Shen et al. [59] proposed an Adversarial Cross-Scale Consistency Pursuit (ACSCP). They used the U-Net [160] to estimate the image at the pixel level. In addition, inspired by GAN's success in image translation [159], the author put forward a kind of adversarial training loss, which is used to reduce the fuzzy effect of density map instead of the traditional L2 norm loss. They used two complementary density map generators: one is the input of the large image and the other is the small image which is cut from the large image. According to the fact that the total number of people in a whole image is equal to the sum of

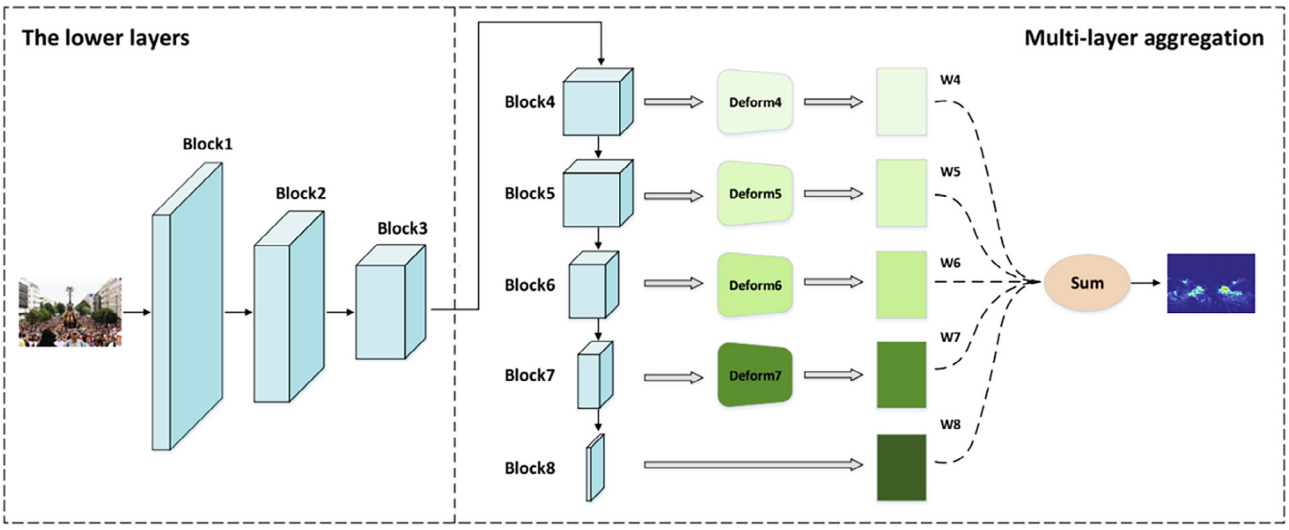


Fig. 8. DANet proposed by Zou et al [128].

people in small images which cut by the whole image, a cross-scale regular standard was proposed. The cross-scale regularization criteria provide strong regularization constraints for cross-scale crowd density estimation, and the method has achieved good performance on several datasets.

Jiang et al. [184] proposed a Trellis Encoder-Decoder network (TEDnet), which is composed of a multi-scale encoder and a multi-path decoder, can generate high-quality density map. In the encoder, a multi-scale convolution kernel is used to extract information of different scales. In the decoder, multi-scale information is decoded to the position of heads in crowded scene. In addition, a new combination loss was proposed to enhance the similarity between local consistency and spatial correlation in density map. With this combination loss, TEDnet can alleviate the problem of gradient disappearance. Liu et al. [198] proposed a Deep Structured Scale Integration Network (DSSINet) for crowd counting. This network uses a representation model based on the structured feature and hierarchically structured loss function to solve the problem of scale change. Different from the traditional method of weighted average or concatenation directly fusing multiple features, they introduced a structural feature enhancement model based on conditional random fields (CRFs), and refined multi-scale features through message passing mechanism. Then, they used an extended loss of multi-scale structure similarity to enhance DSSINet's ability to learn different population scales, so as to obtain high-quality density map.

Xu et al. [194] introduced a Learning to Scale Model (L2SM) to handle large density changes in crowd counting. L2SM can automatically scale different regions, make them have similar density, and significantly improve the quality of density map. When L2SM is added to CNN model for counting, the network can also be end-to-end training. L2SM first extracts the density map of image patch through the density estimation model, and then sorts them according to their density. Then the density map of each image patch level is normalized automatically by using the online center learning based on multipolar center loss (MPCL). Experiments on three datasets show the effectiveness of this method.

At present, most crowd counting methods exploiting deep neural network (DNN) are for single image, while most of existing multi-view counting methods use foreground features [69]. The performance of these methods is limited by the effectiveness of foreground feature. In order to solve the task of counting in a wide range of areas, a deep multi-view crowd counting method was proposed in [187], in which the network can fusion images from multiple cameras for counting and density estimation from a 3D perspective. Firstly, DNN was used to extract feature maps from different camera images, and then these feature maps were projected into a three-dimensional plane according to the camera geometry, so that the features of the same person could be roughly aligned among different feature maps. Finally, these aligned feature maps were fused together to predict the final density map. According to the different information types of fusion, three methods of multi-view fusion are proposed in the literature. The first fusion is the post fusion model: the density maps from each camera image are first predicted and then fused to predict the final density map; The second fusion is the early fusion model: feature maps from each camera image are extracted and then fused to predict the final density map; The multi-view multi-scale (MVMs) early fusion model is the third fusion: on the basis of the early fusion model, the image pyramid is used to extract feature maps of different scales, and the feature maps of different scales are fused to predict the final density map.

3.2.2. Context-aware model

Kasmani et al. [153] proposed an adaptive counting CNN (A-CCNN) based on CCNN [44], which takes the whole image as input and directly outputs corresponding density map. Compared with CCNN, A-CCNN can generate a higher quality density map for crowd counting by using context information. In addition, the authors improved the method of generating density map by CCNN network. CCNN has two important parameters, i.e., the patch size and the covariance of Gaussian function. One disadvantage of this method is that the two parameters are the same for all scales of population, so the generation method of density map is the same. Therefore, when the size of the object in the scene changes greatly, the accuracy of CCNN in density estimation is not high. As shown in Fig. 9, A-CCNN is to detect the size and position of the head in each patch through the head detector. The size and position of the head are transformed into fuzzy information and transferred into the fuzzy inference system (FIS) to output fuzzy language variables in fuzzy form. Finally, according to the fuzzy information generated by each patch, the information is sent to the trained CCNN to get the density map. Compared with CCNN, the main innovation is to change the size of patch and covariance to make the detection more accurate.

Although the method based on regression has made great improvement in crowd counting, how to improve discriminative power of image representation is still a problem to be solved. Sheng et al. [119] proposed a method of integrating semantic information [16,157] into local perceptual feature (LAF) set to perform accurate crowd counting. Inspired by the deep convolution neural network in pixel labeling [122,123], Sheng first built a deep learning model for local feature extraction to learn the semantic feature map at the pixel level, where each dimension of a pixel feature represents the probabilistic strength of a semantic class. Up to now, this is the first time to propose a pixel level semantic feature extraction method based on deep learning. Then, based on the concept of spatial pyramid of adjacent patches, local perceptual feature (LPF) is established to obtain more spatial information and local information. Finally, weighted Vlad (W-VLAD) is used to represent the extracted features as counting result.

Liu et al. [150] considered that multi-layer features are useful for crowd counting, because in the feature extraction network, the convolution layer at the front contains rich information about object position. With the deepening of the convolutional network, some abstract information is extracted, while the information about the position of small objects is gradually lost, which is very important for small object counting. In order to overcome this drawback, they proposed a feature pyramid network (FPN) based on multi-level feature network structure [151], which was used to accurately estimate the counting. They fuse the feature maps extracted from each convolutional layer, combining rich contextual information features with low-level features. As shown in Fig. 10, the network structure removes the full connection layer of VGG16 network and FPN horizontal connection, and combines the feature map of high level with the feature map of low level to ensure that the features contain more context information and small object information. In this figure, P1-P5 represent 5 maximum pooling layers and the density map of model output is 1/4 of the size of input image. This method has achieved good results on several datasets.

Currently, most CNN networks use very limited contextual information in crowd counting. The network looks for crowd features rather than individual features, which easily leads to wrong prediction. Sam et al. [171] proposed a top-down feedback network for counting. The network consists of two parts. The first part is a bottom-up network with two columns and different receptive fields, which is used to predict the crowd density map. The other part is a top-down network which is used to learn how to associate the high-dimensional context information with the low-dimensional features of CNN regression. The top-down network generates masks, which can measure the low-level feature activa-

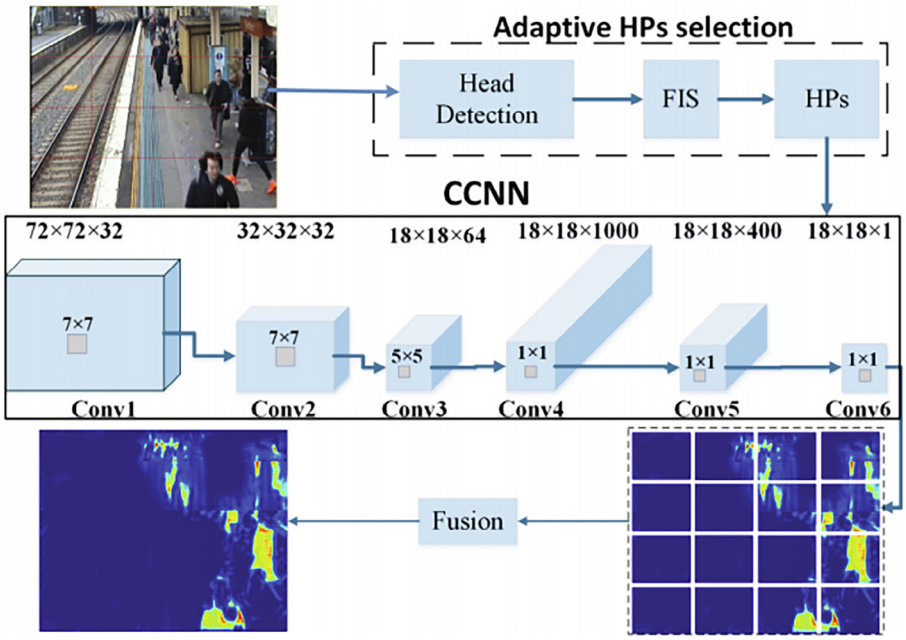


Fig. 9. The overview of A-CCNN crowd counting method proposed Kasmani et al. [153].

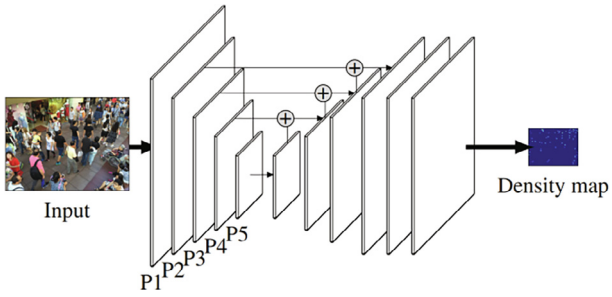


Fig. 10. Network structure proposed by Liu et al. [150].

tion. Liu et al. [182] proposed a network that could adaptively fuse multi-scale features, which are determined by the importance of multi-scale features for dense crowd counting. The network can adaptively encode the range of context information needed to accurately predict the crowd density. It performs better than the most similar crowd counting methods before. Compared with previous crowd counting methods, context information is also used in [59] to explain the scaling effect, but it is only used in the loss function, instead of combining multi-scale context information directly into the crowd counting framework.

The existing methods only deal with the change of crowd size but not the rotation variations. Liu et al. [172] proposed a Deep Recurrent Spatial-Aware Network to address the rotation problem. By using the learnable spatial change module and the regional optimization process, the problem of rotation variations is solved for the first time. This method achieves good detection results. The network consists of two parts: Global Feature Extraction (GFE) and Recurrent Spatial-Aware Refinement (RSAR). The GFE module is used to extract the global features of the input image. RSAR is used to generate high quality density map. It consists of two parts: 1) Spatial Transformer Network (STN) is used to locate the area of real interest in the crowd density map; 2) local refinement network is help for optimizing the density map by residual learning. Finally, a high-quality density map can be obtained. In addition, they conducted an experiment to verify the validity of global context information.

Cheng et al. [200] claim that the existing methods generally use L2 loss to optimize the model, but there are two drawbacks: (1) L2 loss may fail to correctly learn the head position in density maps. 2 When the crowd counting model using L2 loss is affected by noise, the performance will be degraded. In order to solve this problem, Cheng et al. proposed a new network structure called Spatial Awareness Network (SPANet), which can implement crowd counting combined with the spatial environment. The Maximum Excess over Pixels (MEP) loss is proposed in this network. The structure of SPANet is shown in Fig. 11. The input image is represented as a density map D^{pr} through the backbone network. The authors designed a structure including k branches. In each branch k, the network arranges the names of two objectively selected image patches (one image patch and its sub image patch) and generated density map $\tilde{D}_k^{pk}$. S_k represents a value that is significantly different from the true value. The density value S_k is deleted in the next branch, which is used for the second optimization. Finally, the S_k of the k branch are fused to get sub-region s and the MEP loss is calculated by the sub-region. Using MEP loss to optimize the model has achieved good results.

3.2.3. Auxiliary-task model

Marsden [112] used auxiliary task to improve the performance of each task. They proposed a Resnet18 based architecture for crowd counting [86], which is called ResnetCrowd. ResnetCrowd can perform three tasks simultaneously: crowd counting, the detection of violent event and crowd density level classification. There is no method that can carry out these three tasks at the same time prior to ResnetCrowd. Authors found that these tasks could mutually improve the performance. The main modules of the network include the first five convolutional layers of Resnet18, the interleaved batch normalization [113] layer and skip connection. Marsden removed the maximum pooling layer after the first convolutional layer of Resnet18, and retained the larger feature map for pixel level crowd counting. They first used the counting heatmap convolution layer to count the people at the pixel level, and then output the crowd density map. Generally, they fused the feature maps generated by feature extraction network and added

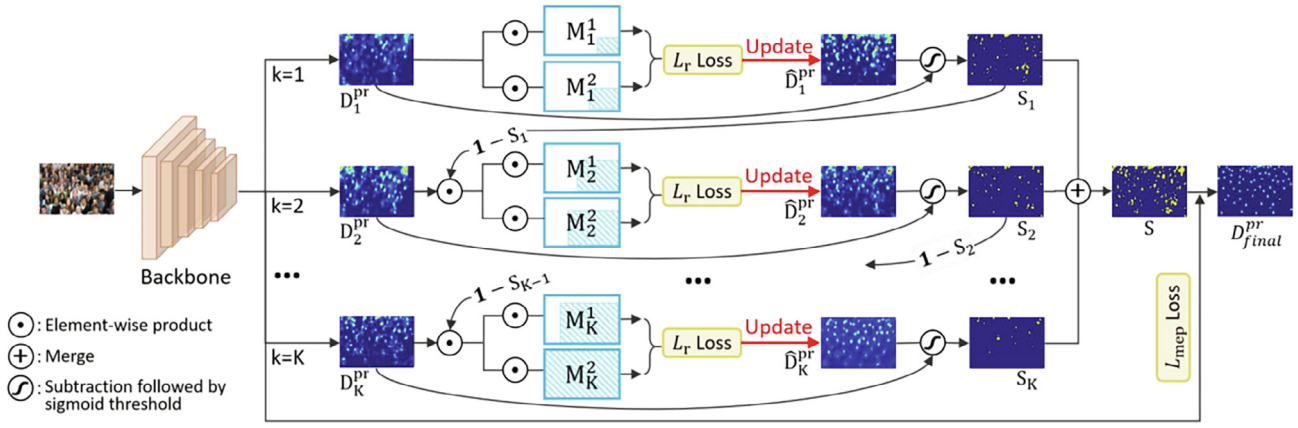


Fig. 11. SPANet structure proposed by Cheng et al. [200].

them to the full connection layer of different tasks. In addition, the author also constructed a dataset for these three tasks. The dataset contains 100 images, which are labeled with the number of people, the location of each head, and the crowd density level of 1 to 5.

Zhao et al. [106] claim that for large public areas with dense population, it is unfeasible to count the population in all areas through surveillance cameras. It is feasible to count it through LOI (line-of-interest) at the exit and entrance. The traditional count of LOI is to extract 2D temporal slices generated on the line corresponding to LOI in the video, but the result is not good in the scene with high pedestrian density and low camera perspective. To this end, they used CNN to learn features from the video sequence. Their method is divided into two stages of training. They first used CNN to learn predicting crowd density map and crowd speed map. These two tasks (crowd density map and crowd speed map) share the bottom neural layers. Then, CNN learns to make end-to-end prediction on the crowd counting map. In addition, the author provides a large dataset for the evaluation of cross line crowd counting method.

The performance of shallow CNN is limited in challenging scenes, such as camera illumination changes, scene occlusion, serious crowd occlusion and angle distortion. Shi et al. [122] proposed a multi-scale multitask model based on VGG16 [92] which adopted a multi-task method to reduce the overfitting caused by small datasets. Shi extracted the multi-scale features of convolution from the whole image. These features were expressed in a compact single vector. And then the Vector of Locally Aggregated Descriptors (VLAD) [127] was applied to accurately count. The strategy of "deeply supervised" is adopted to further improve the performance. The network structure is shown in Fig. 12. They used a network similar to VGG to fuse the features extracted from the third and fourth convolutional layers, which were sent to the NetVLAD [103] layer for processing. After the NetVLAD, there were two regression layers. The output of the two regression layers was averaged to obtain the final predicted density map.

Huang et al. [161] consider that most of the existing methods directly detect the whole body or only the head of the pedestrian, without accurately capturing the semantic structure information of other parts of the body. To solve this problem, they added semantic modeling to the crowd counting model. In order to improve the performance of crowd counting and crowd density estimation, they used the detection of pedestrian body parts as an auxiliary task. The detection adds rich context information to the model. In addition, the authors improved the traditional density map and proposed a structured density map. Then, the deep convolution neural network is applied to a unified pattern learning sub task. In a unified scheme, the problem of feature extraction and multi-task crowd counting is solved.

Most of the previous counting methods used the point supervision algorithm exploited the dataset of point annotation (using points to label head) to train network, which converts the input image into density map by CNN. Shi et al. [173] think that point annotation can be used not only for constructing density map, but also for more monitoring purposes. Firstly, the author supervised the points from the perspective of segmentation, transformed the labeled points into binary maps, and then combined the binary maps with network branches and loss functions to focus on the region of interest. Secondly, the author proposed to monitor points from the global density, and the other branch used the proportion of point annotation in image pixels to regularize the overall density estimation. Both of them are integrated with density estimation to train in the end-to-end single network with multiple losses.

Previous counting methods based on CNN mainly improve the robustness and effectiveness by fusing multi-scale features or add context information. Zhao et al. [186] identified three attributes (geometric, semantic and numeric) that are critical to crowd density estimation, and they used these attributes as multiple auxiliary tasks to improve the performance of the crowd counting task. By adding these auxiliary attributes to the CNN model, more accurate counting results can be obtained. Sam et al. [162] used the Incrementally Growing CNN structure (IG-CNN) for crowd counting, starting from a basic CNN that is used to return the crowd density. In IG-CNN, the basic CNN is copied into two sub-regressors by copying the weight of the parent network. IG-CNN specialized these sub-regressors through differential training [15] (the regressor can better predict the number of people in the image). In the next steps, each sub-regressor is copied to the two networks again. Also, this step performs the difference training. This process is continuously performed to generate a binary tree with two nodes, and the children nodes are more specialized than the parents. When the CNN tree is built, each leaf node is an expert regressor which corresponds to a specific subset of training dataset and the performance is the best in this sub node. In the test, the input image patch is predicted by the best expert regressor to get the best performance, which is a relatively new method. Liu et al. [152] proposed a new framework called Recurrent Attentive Zooming Network (RAZN) that can solve both crowd counting and location. They observed that positioning in high-density areas is often inaccurate, and that improving resolution is an effective and simple method. RAZN can recursively detect the blurred image region, and which is enlarged to high resolution and detect again.

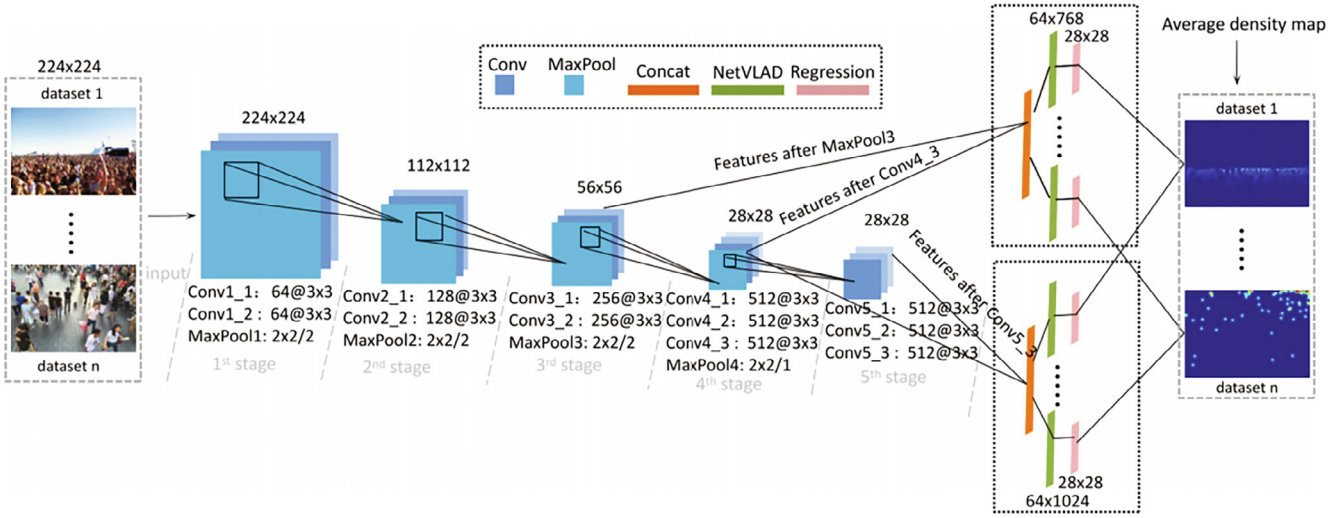


Fig. 12. The multi-scale multi-task depth NetVLAD network structure proposed by Shi et al. [122].

3.2.4. Models dealing with the lack of labeled data

In the task of crowd counting, there is very little data with labels, and it takes a lot of effort to label such data. If there is an unsupervised method to get the feature representation we need, it will be very helpful to the task of population counting. Lu et al. [52] redefined the counting problem as the object matching problem based on the image self-similarity, and established a counting model suitable for any object category. For example, only one image patch of interest needs to be specified. If a small patch in the original image is similar to this image in some way, then they can be considered as self-similarity. For this reason, Lu developed a general matching network (GMN), which learned a discrimination classifier to match a given image patch. GMN consists of three parts: 1) embedding module, a two-stream network is used to encode the input image into advanced semantic features; 2) matching module, which learns a discriminator to match the sample image patch to the image instance; 3) adaptive module, which freezes all parameters of pre-training GMN and trains only adapters and batch normalization layers to speed up training. Lu demonstrated the flexibility of this method on different counting tasks (especially cell, car and crowd counting tasks). The model achieves desirable performance on the cell and crowd counting datasets.

a. Semi-supervised model

Olmschenk et al. [74] proposed a semi-supervised model for crowd counting, which extended the semi-supervised generative adversarial networks (GANs) from classification problem to regression for dense crowd counting. As shown in Fig. 13, they designed a semi-supervised dual objective GAN structure, which requires the discriminator to provide two separate outputs: the expected regression value and a tag used to determine if the input sample is real. The combination of supervised regression and unsupervised classification forces the discriminator to learn more robust features of the crowd image, so that it can perform well even with few labeled data. The structure of the frequency discriminator is based on DenseNet201. Compared with the full supervision network, the network has higher predictive ability.

b. Weakly Supervised model

Sam et al. [183] proposed an almost unsupervised method for crowd counting. They proposed an autoencoder called Grid Winner-Take-All (GWTA) to learn useful features from unlabeled images. GWTA divides the convolution layer into grid cells in space. In each cell, only the most active neurons are allowed to update the filter. So, the parameters in response propagation are sparsely updated. Unsupervised training is done in stages. Each layer is updated by reconstructing its own regularized input through GWTA sparsity. The last two levels are trained under supervision. Almost 99.9% of the model parameters of the network are trained without any labeled data, and the remaining 0.1% of the model parameters are fine-tuned under supervision. Up to now, this is the first 99.9% of the model parameters are from the unlabeled data crowd counting system.

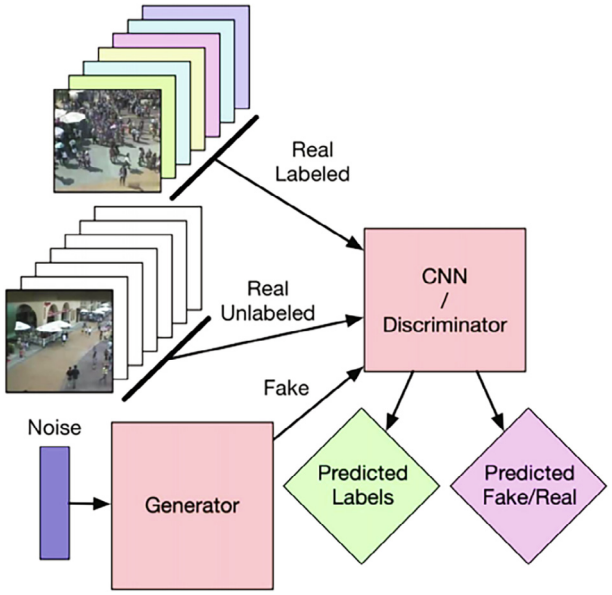


Fig. 13. The structure of a semi-supervised dual-goal GAN [74].

Many existing approaches use point annotation dataset (mark a dot in the center of the head) to perform the counting, which is simpler than bounding-box annotation, but it is still very difficult in highly crowded scenes. Therefore, it is necessary to develop a model with weakly supervised model. Lei et al. [94] observed that it is not sufficient to solve the integration of the density map to the direct solution of the object counting. They used multiple auxiliary tasks to train the crowd-counting network and added a stronger regularization, which was useful for predicting the density of weakly annotated images.

c. Self-supervised model

Because the size of the existing crowd counting dataset is very limited, many crowd counting methods have some overfitting problems. To this end, Liu et al. [25] proposed a method for crowd counting by using more available unlabeled images. They observe that the total number of people in any sub-image of an image is less or equal than that of people in its image. Liu used this constraint to assist in training the network. They provide two ways to collect datasets. One is keyword query: search for images that may be keywords on Google, and then delete images that are not related to the problem. Another way is query by-example image retrieval: use the existing dataset as the training image to search on Google, select the top ten similar images and delete the irrelevant images. Then the sorting image dataset is generated according to the rules and is embedded into the crowd density estimation network. This work proposes three embedding methods: (1) Ranking plus fine-tuning: the network first trains on large, unlabeled and ordered datasets, then fine-tunes them using small and labeled datasets which is the self-supervised (Training unlabeled datasets in a supervised manner) method used in most literatures [80–84]; (2) Alternating-task training; (3) Multi-task training. The model is shown in Fig. 14, in which VGG conv indicates the convolution layer of VGG16. This work carried out experiments on three kinds of training methods respectively, among which the alternative-task training method has the lowest mean square error and the multi-task training mode has the lowest average absolute error.

Liu et al. [20] extended the work in [25], proved that the ranking task can be used as a self-supervised auxiliary task to improve the performance of the counting task. The author uses the following algorithm to generate the ranking dataset. 1) In the center of the original image, cut out a sub-image that is the same as the original image, and the size of the sub-image is 1/r of the original image. 2) Cut the sub-image into the largest square, centered around the center of the sub-image. 3) Get other k–1 square patches, whose sizes are iteratively reduced according to the scale factor s, and let the center of all image patches be at anchor. 4) Resize these k sub-images to the size of the network input. 5) Sort the sub-images according to the number of people in the sub-image to get the final ranking dataset. The method of collecting the image dataset is similar to the method in [25]. And the ranking dataset is embedded into the crowd density estimation network for training.

3.2.5. Domain adaptation model

Due to the variant environments of the real scene, the existing crowd counting methods cannot address well the high density of the population. Wang et al. [28] solved the problem from two aspects. One is data. They developed a data collector that generates synthesized data and a data labeler, which is used to automatically label the synthesized data without any manual work. Another aspect is methodology. The author proposed two kinds of methods. One is that the network trains on the synthesized large dataset and then fine-tunes them using the labeled dataset. The other crowd counting method is based on the domain adaptation, which can avoid a lot of data annotation works. In [28], the GCC dataset collected in "Grand Theft Auto V "(GTA5) by the collector and labeler, is different from the literature [87–89]. Because there is no crowded scene in GTA5 (256 people at most in the same scene), Wang et al. designed a strategy to construct a crowded scene. Specifically, they segmented several non-overlapping regions, placed people in each region, and finally spliced multiple scenes into one scene. GCC dataset consists of 15,212 images with a resolution of 1080 * 1920, including 7,625,843 people. This large dataset can be used to effectively evaluate the performance of crowd counting methods. Because FCN based methods [12,18,90,91] perform well in crowd counting, Wang designs a spatial full convolution network (SFCN) to directly regress the density map, and its structure is shown in Fig. 15. SFCN can use VGG-16 [92] or ResNet101 [86] as the backbone network. It adds a spatial encoder at the top of the backbone network and a regression layer after the spatial encoder. The density map of SFCN output is 1/8 of the input image.

Another method proposed by Wang et al., is the domain adaptive crowd counting, which can migrate and learn effective features in different domain datasets. Based on the Cycle Gan [91], SSIM [93] is embedded in the network, to achieve the migration from synthetic image to real image. After that, the migrated data is put into SFCN for training, and the specific process is shown in Fig. 16.

3.2.6. Perspective map model

In practice, when a trained model is used in a new scene, the performance will be significantly degraded. In order to solve the problem, Zhang et al. [19] suggested learning a mapping that transforms low dimensional features extracted from images into crowd density images. They first alternately trained two kinds of related loss functions of crowd counting and density map estimation, and then optimized these two functions alternately to get better local optimization. They used training samples tightly related to the target scene. They chose some special scenes that were similar to the trained scene dataset and put them into the training dataset to fine-tune the model, so as to adapt to the new target scene. They also proposed a method which generating density map based on perspective information, which enables the network to perform perspective normalization, and then improve the robustness of the network to scene scale and perspective changes. In addition,

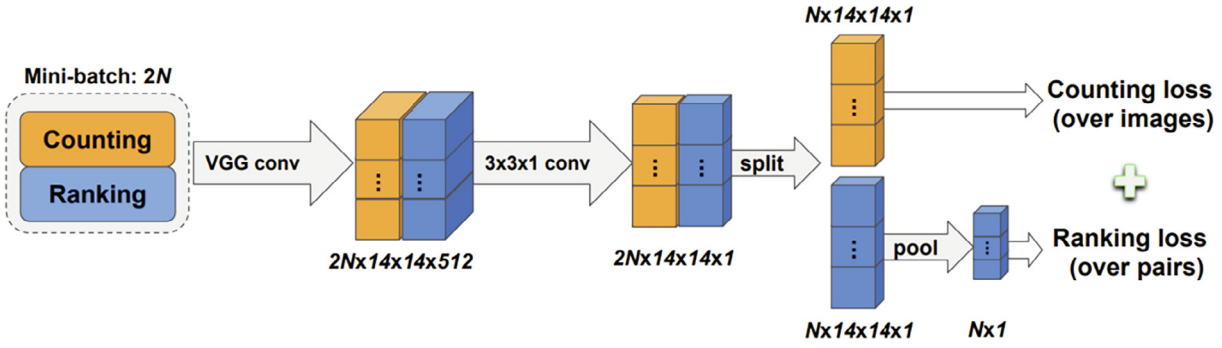


Fig. 14. Multi-task model combine counting and ranking proposed by Liu et al. [25].

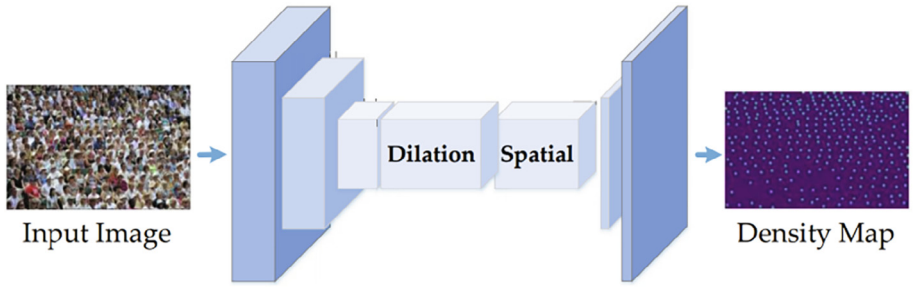


Fig. 15. SFCN architecture proposed by Wang et al. [28].

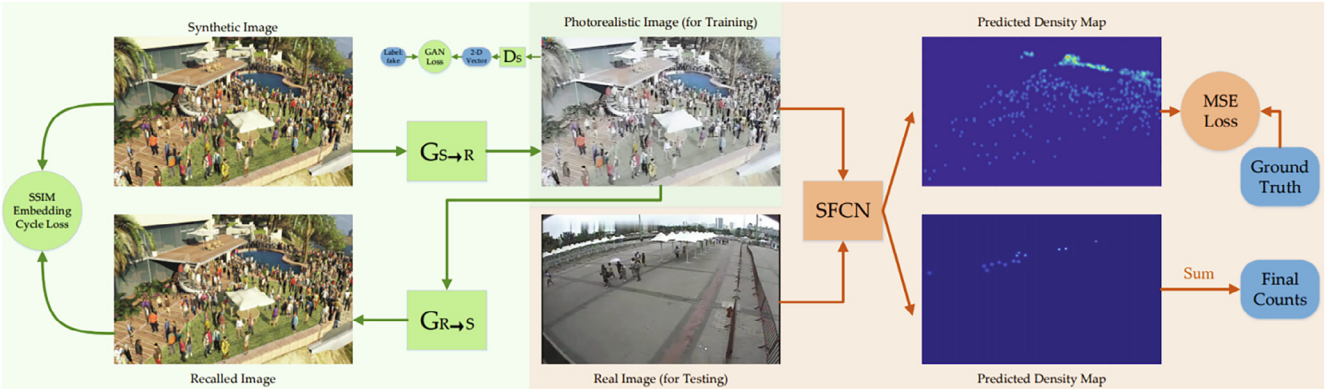


Fig. 16. The flowchart of crowd counting by domain adaptation proposed by Wang et al. [28].

they provide a large dataset containing 108 crowd scenes and nearly 200,000 labeling people, which bring about a greater challenge to the cross-scene crowd counting model.

Shi et al. [177] proposed a Perspective-Aware CNN (PACNN), which used perspective map when predicting density map. The perspective map provides scale change information. It is important for the detection of smaller pedestrians. Shi first used a new method to generate ground truth perspective for training, which can predict the perspective and density map during the test phase. Perspective is encoded as two weight layers of perceptual perspective to adaptively combine the different scale information used to predict the density map. This method has achieved advanced results in UCSD dataset. Yan et al. [188] proposed a crowd counting algorithm (PGCNet), which was used to overcome the scale variation of people in scenes caused by visual effects. PGCNet uses the perspective information to guide the smooth spatial variation of feature maps, and then inputs it into subsequent convolution. An effective perspective estimation branch is introduced in PGCNet, which can be trained under either supervised or weakly supervised settings. In addition, they also provided a dataset called Crowd-Surveillance. Crowd-surveillance contains 13,945 images that are similar in density to Shanghai Tech B.

3.2.7. Attention mechanism model

The most direct way to solve the scale change of crowd counting task is to fuse the features of different scales. In the feature extraction network, the earlier convolutional layer captures the original features. With the deepening of the depth, the later convolutional layer extracts more abstract features. However, it may not be the most effective way to directly integrate these multi-scale feature maps. Sindagi et al. [174] proposed a Hierarchical Attention-based Crowd Counting Network (HA-CCN). They added attention mechanism to the network to improve crowd counting performance. By using the attentional mechanism on different convolutional layers, it can selectively enhance some of the important features that help with the crowd counting task. HA-CCN consists of VGG16 network [93], Spatial Attention Model (SAM) and a series of Global Attention Models (GAM). SAM is used to enhance low-level features in the network, and GAM is more focused on high-level information, which selectively enhances important features while suppressing unnecessary features. The network structure is shown in Fig. 17, in which VGG16 is the basic network, and the feature map extracted by conv3 is sent to SAM. SAM integrates the pixel level segmentation information into the feature, and the feature map extracted by higher levels (conv4, conv5) is sent to GAM to enhance the feature map.

Liu et al. [175] developed an Attention-injective Deformable Convolutional network (ADCrowdNet). They added attention mechanism and multi-scale deformable convolution to ADCrowdNet to solve the problem of decreasing the counting accuracy in high-congestion noise scenes. The ability of visual attention mechanism used to reduce the impact of various noises in input has been proved by [176], and multi-scale deformable convolutional network is used to deal with dense population. The structure of ADCrowdNet including two connection networks: 1) the Attention Map Generator (AMG) is used to detect parts of the crowd area in a dense scene, and then calculates the degree of crowding in those areas. 2) the Density Map Estimator (DME) takes the crowd area detected by AMG as input, based on the multi-scale deformable network, and then generates high mass density diagram. Thanks to attention aware training and multi-scale deformable convolution, ADCrowdNet can capture the features of the crowd more effectively and resist all kinds of noises.

For a single image, Hossain et al. [196] introduced a novel scale-aware attention network (SAANet) to address the challenges brought by scale changes. The network can extract the global scale and local scale features suitable for counting tasks, and implement the dense crowd count evaluation through the global and local

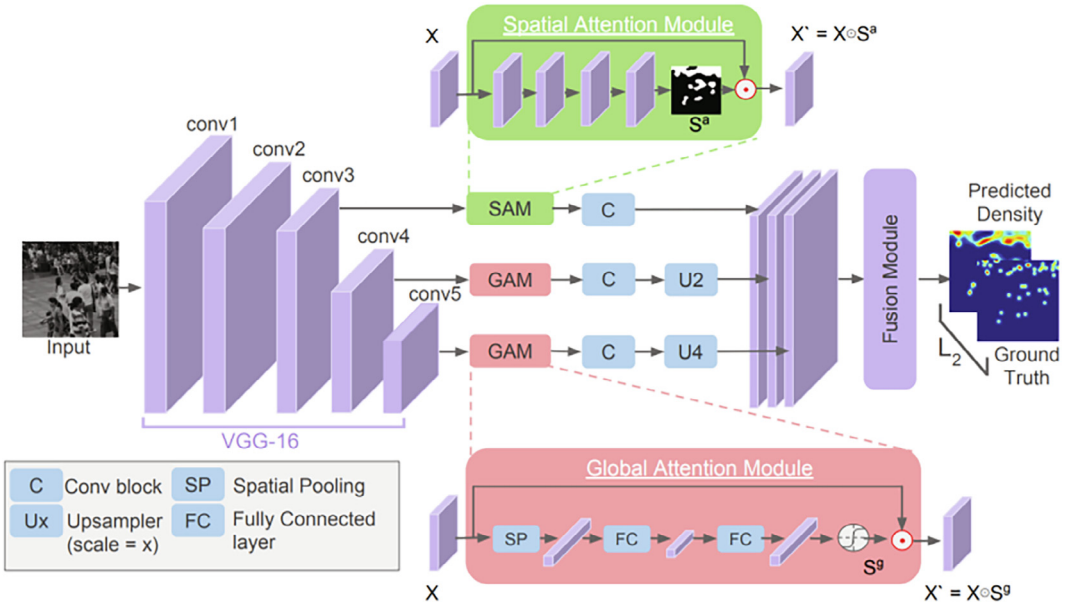


Fig. 17. Crowd counting network based on hierarchical attention proposed by Sindagi et al. [174].

attention network. Their network structure is shown in Fig. 18. Their model consists of three parts: (1) multi-scale feature extractor is used to extract three different scale feature maps. (2) global scale attention network (GSA), which is used to predict three global score. (3) local scale attention network (LSA) that is applied to generate a local attention map. Then the multi-scale information is weighted according to the global score and local attention map, and the weighted multi-scale information is used to predict the crowd density map. Finally, counting result is obtained by integrating the density map. This approach has got the best result on the dataset of MALL.

3.2.8. Network architecture search model

At present, most of the crowd counting and crowd density estimation methods using CNN are based on the hand-designed density estimation network. In this kind of network, the multi-scale features are generally used to deal with the scale changing, which usually requires careful design to extract the multi-scale features of the basics network. Hu et al. [71] used neural architecture search (NAS) to automatically build the crowd counting model, and introduced an end-to-end automatic search Multi-Scale Network (AMSNet). As shown in Fig. 19, NAS-Count effectively searches for a multi-scale encoder-decoder network, i.e., AMSNet. The encoder-decoder network is composed of different units, each of which can automatically extract and fuse multi-scale features. In addition, Hu introduced a novel Scale Pyramid Pooling Loss (SPPLoss) to optimize AMSNet, and the pyramid structure was used to perform multi-scale structural supervision. NAS-Count is the first network that applies network architecture search to crowd counting tasks. By automatically implementing the multi-scale model in less

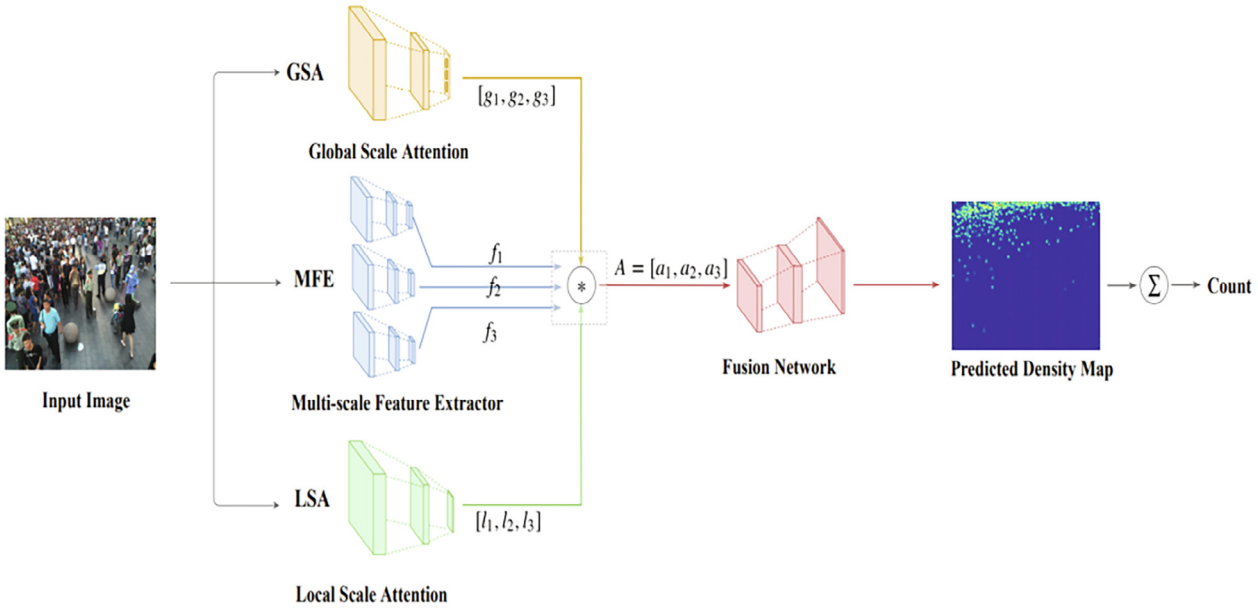


Fig. 18. Scale-aware attention network proposed by Hossain et al. [196].

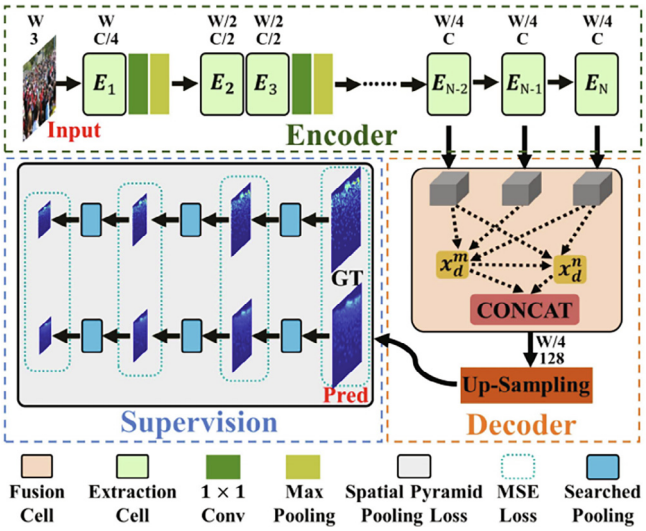


Fig. 19. AMSNet architecture and SPPLoss monitoring. (all searched cells are shown in black).

than one GPU day, NAS-Count surpasses tedious hand-designing efforts and shows good overall performance on four challenging datasets.

3.3. Combination of detection and regression method

In the case of low crowd density, the method based on detection is more effective than those based on regression. When the crowd density is high, the regression-based method is better than the detection-based method. In order to make use of the advantage of these two methods, Liu et al. [85] proposed a method named DecideNet (Density Estimation Network) combining detection and regression method. DecideNet can adaptively adjust the weights of detection and regression methods according to the changes of crowd density. As shown in Fig. 20. DecideNet consists of three modules: RegNet, DetNet and QualityNet. The RegNet module uses the regression-based method to calculate the network density map. In order to obtain more scene information, Liu et al. used a relatively large (7×7, 5×5) receptive field in RegNet. The DetNet module uses the detection-based method to detect the human head, which can accurately locate everyone in the scene. QualityNet: the essence of this module is to assign weights

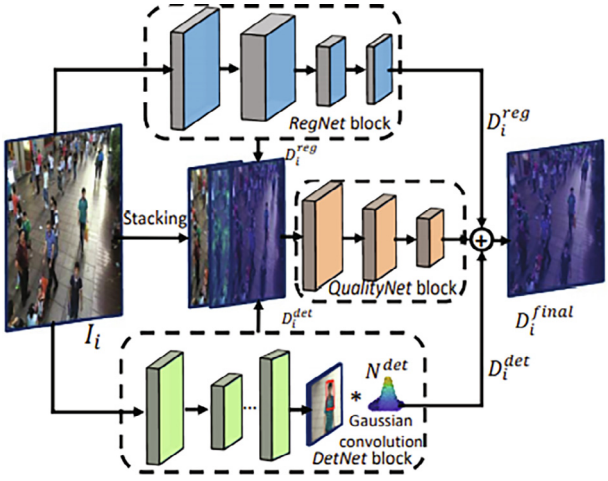


Fig. 20. DecideNet network structure proposed by [85].

to two networks. This approach has achieved good results on several datasets with low crowd density.

The work [85] is only to carry out experiments on datasets with low crowd density, and when the scene is very crowded, the cost of boundary box annotation on the population is very large. To this end, Liu et al [29] proposed a Point-Supervised Deep Detection Network (PSDDN), which only needs points rather than the bounding box to label the head. It can count the crowd in the scene and locate each person. Liu et al. find that the size of head is related to the distance of a number of heads around when the crowd density is high. In a crowd scene, the head size on the same horizontal line is usually unchanged. The head tends to be small when it is close to the top of the image. Based on this fact, a method was proposed to initialize the point annotation as the real annotation of the initial box and update it in real-time during training.

In order to solve the problem of crowd counting and locating in dense scenes, Lian et al. [192] proposed a Regression Guided Network (RDNet) for the crowd counting in RGB-D scenarios. They used the density map to improve the performance of the detection network. At the same time, they proposed a depth-adaptive kernel considering the variance of head size to improve the quality of density maps. In addition, in order to better initialize the anchor size, the depth-aware anchor is also used in the detection framework. RDNet structure is shown in Fig. 21, it consists of two parts. One is the density map regression module and the other is the head detection module. In the first part, they introduced an adaptive kernel to improve the quality of the generated density maps. In the head detection module, Lian et al. used RetinaNet [193] which has advantages in detection speed and accuracy. At the same time, Lian also proposed a depth-aware anchor strategy to initialize the appropriate anchor size, which is also helpful to improve the detection performance. This model can be extended to crowd counting based on RGB image. In addition, because the existing RGB-D dataset is too small, Lian introduced a large RGB-D crowd counting dataset named ShanghaiTechRGBD, which contains 2193 images and 144,512 human heads.

Sam et al. [142] proposed a dense detection framework LSC-CNN (Locate, Size and Count) for crowd counting. LSC-CNN only uses the dataset of point annotation for training. It can locate the size and position of each head in the scene. LSC-CNN adopts a multi-column architecture with top-down feature modulation to better detect people in the scene and generate accurate predictions at multiple resolutions, which combine to form the final detection. LSC-CNN consists of three modules. The first module is the feature extraction module, which sends features extracted from images of multiple resolutions into the module, i.e., Top-down Feature Modulator (TFM) module. The second module TFM fuses multi-scale feature graphs, predicts detection boxes, and then uses NMS to select effective detection from multiple resolutions. The third module is Grid Winner-Take-All (GWTA) module. GWTA is mainly applied to address data imbalance during training. GWTA divides the prediction graph into a series of grids with fixed sizes, and then only calculates loss for one grid. Since there is only a small region of loss in prediction graph, the problem of gradient average is avoided and the influence of local minimum can be reduced. Experiments show that this model not only has better population counting performance than the existing regression method, but also has superior positioning performance and has all the advantages of the detection system.

3.4. Crowd counting method based on video

In video surveillance, there are usually two kinds of views: oblique angle and vertical angle. It is difficult to judge a person in different scenes by a unified detection model. Li et al. [95] proposed a model based on Faster R-CNN [96] combining the head and shoul-

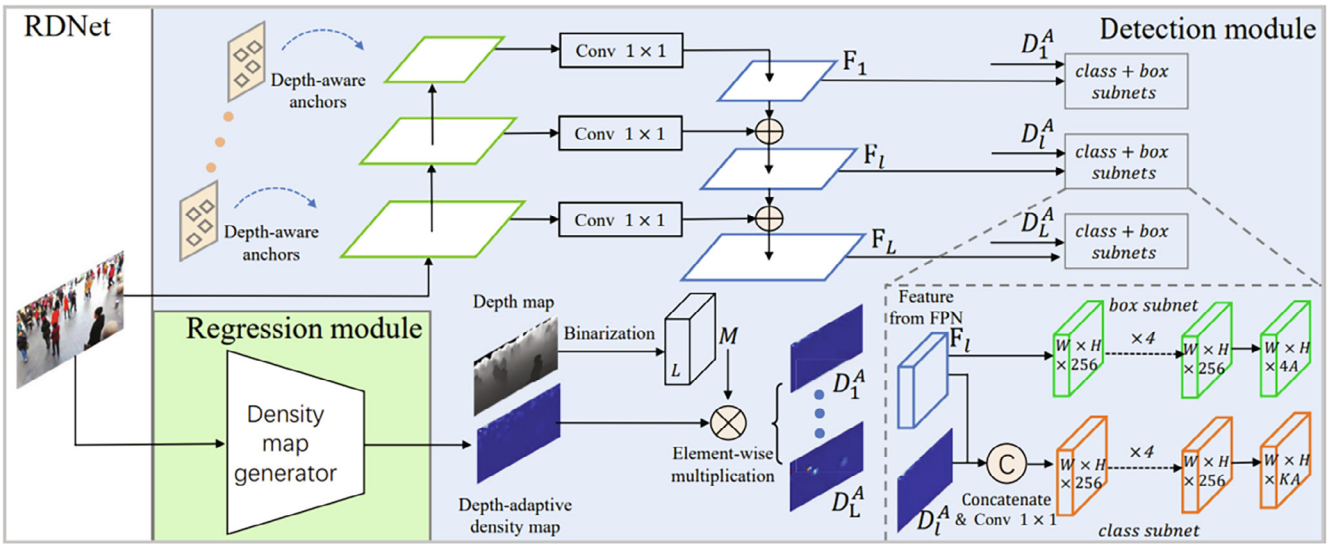


Fig. 21. RDNet structure proposed by Lian et al. [192].

der detection and tracking, in which each tracking bounding box is matched with one detection in following-up frames. They first trained a head shoulder detector using Faster R-CNN model, and added online hard example mining (OHEM) [98] to reduce dection error during the training. Because people who are not detected in one frame may be detected in the following-up frame, they used kernelized correlation filter (KCF) [97] to track pedestrians and got their tracks, and then merged the tracks obtained by detection and tracking together to get a continuous and stable pedestrian counting track. The network structure is shown in Fig. 22. This figure first inputs a video, and uses Faster R-CNN to detect the crowd bounding box, and then initializes a KCF tracker for a pedestrian who has not been detected before. Finally, we can extract the motion curve according to the results of the KCF tracker. When a tracking box does not match continuously in the next 10 frames, it can be considered that the initialization box is an incorrect detection.

Although the CNN-based methods had achieved good results in crowd counting in recent years, they still consider each frame independently when processing video data, and ignore the time information between adjacent frames. In order to utilize the very useful time information in video dataset, Xiong et al. [30] proposed a deep learning model based on convolutional LSTM (ConvLSTM) [116], which can capture the correlation between space and time at the same time. Inspired by [117,118], Xiong used a bidirectional ConvLSTM model to obtain bidirectional long-distance information. In addition, in order to verify the validity of the time information, they proposed ConvLSTM-nt (i.e. no time information). Compared with ConvLSTM and bidirectional ConvLSTM, the validity of time information is proved. The results of this method are better than those of the existing similar methods.

The research of crowd counting is mainly focused on the single image. Since the faster moving crowd counting is very important for urban public security management, Wei et al. [110] proposed a faster moving crowd counting method. In this method, support vector regression and spatial–temporal multi-features are used to enhance the ability of deep attribute learning. Firstly, a new spatial–temporal multi-feature is developed by combining multi-appearance feature and multi motion feature based on super-pixel. Secondly, they proposed a new depth attribute cumulative learning architecture (CA-VGG) based on VGG16 to solve the problem of human posture deformation when people move rapidly. Finally, a deep attribute learning method based on support vector regression method is proposed to improve the performance for crowd counting.

Zou et al. [197] used enhanced 3D convolution network (E3D) to solve crowd counting task in a video scene. As we know, this is the first time to apply 3D convolution to crowd counting. By introducing 3D kernel, the model can capture temporal and spatial information simultaneously, thus improve the counting performance on video dataset. In the video scene, the translation of the human will cause the head density map to change between neighboring frames. At the same time, people entering or leaving the dynamic scene will lead to changes in the head counts, which increases the difficulty of processing video. To solve these problems, Fang et al. [70] proposed a Locality-constrained Spatial Transformer Network (LSTN). They used LST module instead of using LSTM or ConvLSTM to extract the time information between adjacent

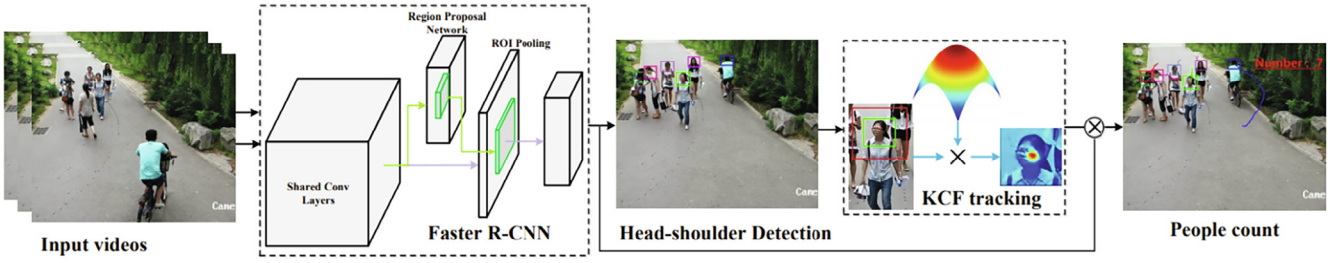


Fig. 22. Network structure proposed by Li et al. [95].

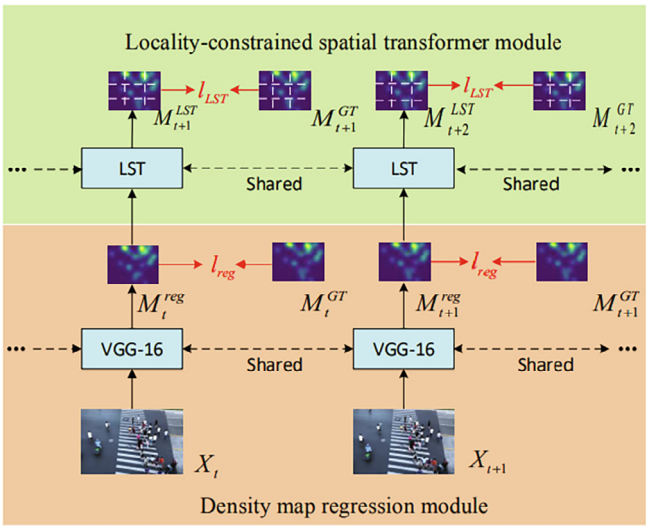


Fig. 23. The structure of LSTN module for video crowd counting.

frames in the video as in literature [30]. The structure of LSTN is shown in Fig. 23, which includes two modules: the density map regression module and the LST module. The density map regression module predicts the density of each frame, and inputs the generated density map to the LST module, which predicts the density map of the next frame. In another work, Fang et al. [156] proposed a Multi-Level Feature Fusion Based Locality-Constrained Spatial Transformer Network (MLSTN). They fused the features extracted from each layer of the feature extraction network, which helped to detect people of different scales. To facilitate performance evaluation, Fang collected a large video crowd count dataset consisting of 15,000 frames that captured approximately 394,000 annotation headers from 13 different scenarios. This is by far the largest dataset in terms of the number of images (see Fig. 24).

3.5. Analysis

In Subsections 3.1 to 3.4, we introduced the crowd counting methods based on CNN in recent years. In this subsection, we analyzed the advantages and disadvantages of various methods and discussed some novel approaches.

In order to solve the problem of perspective and crowd scale, many methods [12,15,42] use multi-column network structure to capture multi-scale information, which improves the performance of the network to a certain extent. However, extracting related features through multi-column or multi-network models requires a lot of computation and is difficult to optimize. To this end, literatures [18,90,132] gave up the multi-column network structure. Zhang et al. [18] proved that the multi-channel convolutional network (MCNN) had the limitation of structural redundancy, various parameters of structure and the difficulty in training. To solve this problem, Zeng et al. [90] extracted multi-scale features from a single column structure by introducing multi-scale blobs similar to naive Inception module [108], which has better performance of crowd counting in practical applications. In another different method, Kang et al. [51] used image pyramid to deal with the change of scale. Images of different scales were sent into FCN to obtain the density map of corresponding scale. All scales of FCN and attention mechanism shared parameters. Finally, all density maps were fused by convolution of 1*1. The results are better than MCNN and the training parameters are less. However, these methods only consider discrete representative scales and cannot model continuous scale changes. We can see that single-column networks have fewer network parameters, higher speed, and better performance than multi-column networks.

Literatures [20,21,25,112,122,161] use multitask to improve crowd counting performance. Sindagi et al. [21] used multitasking learning to add the prior knowledge to the network. They first classified crowd images into 10 groups according to crowd density classification, and learned a crowd counting group classifier. They proved that crowd density classification could improve the count-

Fig. 24. Sample images of the BRAINWASH dataset [17].

ing performance. It should be noted that the number of such classifier is determined based on the density distribution of the dataset. The methods in [121,150,155,172,184] add contextual information to the network to reduce detection error. But these methods only consider the spatial information of a single image, and the relevant knowledge and semantic priori between samples have not been fully utilized. To this end, Wang et al. [190] proposed a method, which uses the relevant information between images to perform crowd counting. The existing research shows that mining relevant knowledge by comparing samples can learn more intrinsic features [191], and consequently improve the generalization ability in unfamiliar scenes. In addition, Wang also shows how to effectively use semantic information before improving the performance of crowd counting. The quality of predicted density map is improved by using antagonistic loss to further improve the performance of crowd counting. First, residual regression is applied to predict the difference between the input image and the supporting image. Residual regression is used to predict the residual map (the difference between the density images). Then the final density map is generated by combining all residual predictions with the predicted density map of the original image. They took MCNN [12] and CSRNet [18] as the backbone network to experiments, and got good results. In our analysis, we found that the performance of the crowd counting task could be improved by using the relevant information between samples, or by optimizing the model jointly with the auxiliary task and the crowd counting task.

As we all known, the training of neural network needs a lot of annotated data. but in some scenarios, the training data are very difficult to obtain or annotate. Therefore, it is obviously very important to use unlabeled datasets to assist the network in training in these scenarios. To this end, [20,25] generated a dataset by sorting the unlabeled data. Then, these datasets were embedded into the crowd density estimation network to perform self-supervised learning. In addition to self-supervised learning, weakly supervised learning and semi-supervised learning are also solutions to the lack of data. Sam et al. [183] presented an unsupervised learning method for crowd counting. Grid Winner-Take-All (GWTA) was used to learn several layers of useful filters from unlabeled crowd images. The training procedure is carried out in stages. GWTA was applied in every layer except the last two convolutional layers. Therefore, 99.9% of the parameters in the network were obtained without supervision, which fully reduced the network's dependence on labeled data. Since the GANs [142,143] has been proven that it can improve the accuracy of deep networks and allow a smaller amount of data to be trained to achieve a higher accuracy [111]. In [43], by redesigning GAN objective formula, semi-supervised generated antagonistic network GANs was applied in crowd counting, which significantly reduced the amount of annotated data required by training in some scenarios. Olmschenk et al. [74] proposed a dual-goal GAN for crowd counting. In addition to getting the number of people in a dense scene, dual-goal GAN was used to distinguish between real images and generated images. Since dual-goal GAN can be used to train for unlabeled data, this method can significantly reduce the amount of labeled data. In the future work of CNN, the design of semi-supervised model, weak supervised model, or unsupervised model will be a research focus.

It is worth mentioning that most methods based on CNN rely on density map to estimate the number of people, but these methods cannot determine the location of each person in the scenario, which is required in some application scenarios (such as counting the number of students in the class and positioning each student). Based on this fact, the works [17,67] adopted the method based on detection to count the dense crowd. In order to solve the problem of serious crowd occlusion in dense scenes, they adopted the detection method based on human head to detect the crowd, so as to accurately detect the human head in the scene. But these methods require bounding boxes to mark heads, which is very expensive for dense scenes in the training. In addition, many methods try to solve problems from different perspectives. Ranjan et al. [169] proposed an iterative CNN structure to generate high-resolution density maps. The network predicts low-resolution density map through a low-resolution CNN branch, and then fuses low-resolution density map with feature map to predict higher-quality density map in a high-resolution CNN branch. In addition, Ranjan thinks that the method can be extended iteratively, and each stage can use the density graph generated in the previous stage to predict the more refined density graph. The method has achieved good results on multiple datasets. Shi et al. [164] believed that many methods based on CNN had the overfitting problem. Ranjan proposed D-ConvNet for crowd counting. Through deep negative correlation (NCL) learning strategy, the generalization ability of the model can be improved by generating more generalizable features. Ma et al. [199] proposed a new loss function, Bayesian loss, which can contribute to the density distribution probability model from point annotation. Their method gets a better result than previous methods on the UCF-QNRF dataset.

From what has been discussed above, we can draw the following conclusions:

1) Scale information aware: Multi-scale information can be extracted through multi-column network, pyramid structure or multi-scale block. Generally, pyramid structure and multi-scale block can get better result than multi-column network with fewer network parameters.
2) Context information aware: The effectiveness of context information has been proved in [172]. Sending the whole image into the network for training can better obtain the global context information.
3) The auxiliary task can speed up the training speed and improve the detection accuracy.

In Table 1, we summarize of crowd counting methods, which are divided into basic network, single-column network and multi-column network according to the attributes of the network. In terms of the mode of supervision, it can be divided into fully-supervised, weakly-supervised, self-supervised and semi-supervised model. According to the training data type, the model can be divided into point-supervision and anchor-supervision. Anchor-supervision means that the dataset annotated by bounding box is used for supervised training, which is generally based on detection method. Point-supervision means that the dataset annotated by point is used for supervised training, which is generally based on regression method. In terms of the training mode, it can be divided into whole image-based model and patch-based model.

It can be seen from Table 1 that most networks adopt a single-column network structure, which has the advantage of fewer network parameters. In recent years the attention mechanism has been gradually added to the crowd counting task, and point-based supervision and fully-supervision methods are still the mainstream.

4. Datasets and results

4.1. Dataset and evaluation criteria for detection based method

BRAINWASH dataset [17]: The BRAINWASH is a dense head detection dataset that contains a group of people who appear in a cafe and then label the group. The dataset consists of three parts:

Table 1
Summary of crowd counting method based on CNN.

Methods	Year/Periodical	Network Properties	Supervision form	Data type	Attention based	Training patterns
Wang [41]	2015/ACMMM	Single-column	Fully-Supervised	Point	No	Patch-based
Zhang [19]	2015/CVPR	Single-column	Fully-Supervised	Point	No	Patch-based
ReInspect [17]	2016/CVPR	Single-column	Fully-Supervised	Anchor	No	Whole image-based
MCNN [12]	2016/CVPR	Multi-column	Fully-Supervised	Point	No	Whole image-based
CNN-Boosting [43]	2016/ECCV	Basic	Fully-Supervised	Point	No	Patch-based
FCNCC [48]	2017/VISAPP	Single- column	Fully-Supervised	Point	No	Whole image-based
CMTL [21]	2017/AVSS	Multi-column	Fully-Supervised	Point	No	Whole image-based
Switching-CNN [15]	2017/CVPR	Multi-column	Fully-Supervised	Point	No	Patch-based
MSCNN [90]	2017/ICIP	Single- column	Fully-Supervised	Point	No	Patch-based
NetVLAD [122]	2018/TII	Single- column	Fully-Supervised	Point	No	Whole image-based
SaCNN [132]	2018/WACV	Single- column	Fully-Supervised	Point	No	Whole image-based
FCNN [150]	2018/ICIP	Single- column	Fully-Supervised	Point	No	Patch-based
BSAD [161]	2018/TIP	Multi-column	Fully-Supervised	Point	No	Patch-based
CSRNet [18]	2018/CVPR	Single- column	Fully-Supervised	Point	No	Whole image-based
IG-CNN [162]	2018/CVPR	Multi-column	Fully-Supervised	Point	No	Patch-based
IC-CNN [169]	2018/ECCV	Multi-column	Fully-Supervised	Point	No	Whole image-based
SANet [170]	2018/ECCV	Single- column	Fully-Supervised	Point	No	Whole image-based
DRSAN [172]	2018/IJCAI	Multi-column	Fully-Supervised	Point	No	Whole image-based
L2R [25]	2018/CVPR	Basic	Self-Supervised	Point	No	Whole image-based
SL2R [20]	2019/T-PAMI	Basic	Self-Supervised	Point	No	Whole image-based
HA-CCN [174]	2019/TIP	Single- column	Fully-Supervised	Point	Yes	Whole image-based
PACNN [177]	2019/CVPR	Single- column	Fully-Supervised	Point	No	Whole image-based
CANet [182]	2019/CVPR	Single- column	Fully-Supervised	Point	No	Whole image-based
L2SM [194]	2019/ICCV	Single- column	Fully-Supervised	Point	No	Patch-based
SAANet [196]	2019/WACV	Multi-column	Fully-Supervised	Point	Yes	Whole image-based
E3DNet [197]	2019/BMCV	Single- column	Fully-Supervised	Point	No	Whole image-based
DSSINet [198]	2019/ICCV	Multi-column	Fully-Supervised	Point	No	Patch-based
SPANet [200]	2019/ICCV	Single- column	Fully-Supervised	Point	No	Patch-based
GWTA-CCNN [183]	2019/AAAI	Single- column	Weakly-Supervised	Point	No	Patch-based
AMSNet [71]	2020/ECCV	Single- column	Fully-Supervised	Point	No	Whole image-based
LSC-CNN [142]	2020/T-PAMI	Multi-column	Fully-Supervised	Point	No	Patch-based

training set (10769 images, 81,975 heads); validation set (500 images, 3318 heads); test set (500 images, 5007 heads).

A comparison of various counting methods based on the detection is described in Table 2, and the evaluation criterion is MAP (average precision).

4.2. Common dataset and evaluation criteria based on regression method

UCSD dataset. the UCSD dataset [26] (2008) is the first dataset created for crowd counting, which was collected from a one-hour video taken from a camera on the sidewalk. The size of each frame is 740 × 480 and the FPS is 30. UCSD dataset selects 2000 frames of the video sequence for labeling and down samples the frame size to 238 × 158. UCSD contains 49,885 labeled pedestrians. The training set contains 800 frames whose indexes are from 600 to 1399, and the rest of the images make up the test set. UCSD has a low crowd density, with a maximum of 46 people in the scene. Moreover, as the dataset is collected from a single location, the scene in the dataset is single and does not have diversity.

Mall dataset. the Mall dataset [10] (2012) was collected by the camera in a shopping Mall. In addition to having different densities of crowd, from sparse to crowded, it also has different lighting conditions. Compared with the UCSD dataset, the scenes in Mall dataset have serious perspective deformity, and the size and appearance of people have larger scale changes. Due to more shielding objects in the Mall, the shielding problem is more serious. Similar to UCSD, Mall dataset is composed of the first 2,000 frames in the video sequence with the size of 320 × 240. The training set and testing set of the Mall data set are divided in the same way (The first 800 frames are used for training, the rest for evaluation). The total number of labeled people in this dataset is 62325, and the maximum number is 53. The crowd density is still relatively low and has a simple scene.

Table 2
Comparison of detection-based methods on the BRAINWASH dataset

Methods	MAP
ReInspect, Lfix [17]	0.60
ReInspect, Lfirstk [17]	0.63
ReInspect, Lhungarian [17]	0.78
Overfeat – AlexNet [168]	0.62
HeadNet [67]	0.91

UCF_CC_50 dataset. UCF_CC_50 dataset is the first dataset that is really challenging in terms of crowd counting [23] (2013). Since the dataset is searched from the web, UCF_CC_50 includes different crowd densities and many different types of dense crowd scenarios. Although the UCF_CC_50 dataset contains only 50 images, each image has a high crowd density, with an average of 1,280 people per image. The only drawback of this dataset is that there are too few images to train and test. Most methods recently based on CNN can hardly achieve desirable performance on the dataset.

WorldExpo'10 dataset. Zhang et al. [11] (2015) provided a crowd counting dataset containing the largest number of scenes, which was obtained from videos that were taken by 108 cameras. And most cameras are shot from the air, so the data set contains a wide variety of scenes. The dataset consists of 3,980 images of size 576 × 720 and 199,923 pedestrians were labeled. The training set and validation set consist of 103 crowd scenes, and the test set consists of 5 crowd scenes. Each test scene contains 120 labeled frames, with an interval of 30 s between the two frames, and the number of people ranges from 1 to 220.

Shanghai Tech (SHTech) dataset. The dataset of Shanghai Tech dataset [12] (2016) is divided into two parts, A and B. It contains a total of 1198 images and 330,165 labeled heads. Considering the limitations of other datasets in terms of perspective, each image in the Shanghai Tech dataset is taken from a different angle. Therefore, the dataset contains a variety of crowded scenes, which brings a great challenge for the complex crowd counting model based on CNN. Shanghai tech A contains 482 pictures, including 300 for

training and 182 for testing. Shanghai tech B contains 716 pictures, including 400 for the training and 316 for the testing. Compared with part B, part A has a larger crowd density.

UCF-QNRF dataset. As shown in Fig. 25, due to the low resolution of the existing dataset image, it is difficult to distinguish each labeled head, and some areas of some images are not labeled or labeled incorrectly. The large-scale dataset UCF-QNRF can overcome the above drawbacks of the other datasets [166]. The UCF-QNRF dataset contains 1535 images, in which 1201 were used for training and 334 for testing. This dataset labeled a total of 1251,642 heads (see Fig. 26).

GCC dataset [28] **(2019).** GCC dataset collected in "Grand Theft Auto V "(GTA5) by the collector and labeler. The dataset consists of 15,212 images with a total of 2,133,375 heads and the density ranges from 0 to 3995. Since the dataset is composed of synthetic data, it is quite different from the real scene. In terms of the number of people labeled, this is by far the most intensive dataset, and it is very suitable for the complex CNN model for training.

City Street dataset [187] **(2019).** City Street dataset is a multi-view dataset which consisting of three perspectives and contains 500 images (300 of which are training and 200 of which are testing). Each image contains 70 to 150 people.

JHU-CROWD++ dataset [58] **(2020).** JHU-CROWD++ dataset consists of 4,372 images, with a total of 1,515,005 heads annotated. The dataset was collected under different scenarios, including some images under different weather conditions and different lighting conditions. It was a very challenging dataset.

NWPU-Crowd dataset [65] **(2020).** The dataset consists of 5,109 images with a total of 2,133,375 heads labeled (using points and boxes). NWPU-Crowd dataset contains some negative samples, such as various density images and different lighting scenes, which can be used to evaluate the robustness of the model.

Table 3 shows the basic information for each dataset, UCF_CC_50 (44 MB); WorldExpo'10 (325 MB); ShanghaiTech A (67 MB); and the proposed UCF-QNRF dataset (4.33 GB), where Min, Ave, Max represents the minimum, average and maximum number of people in the images respectively.

4.2.1. Analysis of results

Tables 4 and 5 show the results of various methods on different datasets, and all experimental data are from the original paper. The bolded number represents the minimum error on the dataset (Black, red, and blue are the top three with the lowest error on this data set), "-" means there is no experimental data. Through the analysis of the results we found that:

1) Compared with multi-column network, single-column network has better performance with less network parameters. Using multiple branch network structure to extract different scale features of the crowd is not optimal. We can believe that deeper network can get better result, instead of wider.
2) The methods combining the multi-scale information and context information can get lower error in each dataset.
3) In recent years, attention mechanism has been applied to the crowd counting, which can make the network focus more on the area of interest rather than treating all features equally. The scale-aware attention network (SAANet) has achieved the best results on the Mall dataset.
4) Due to the limitation of dataset, more and more methods adopt semi-supervised, weakly supervised and unsupervised methods to count the crowd. These methods are not as good as the supervised network in performance, but they are applicable in some scenarios with fewer marked data.
5) Adding perspective information to the network can better obtain multi-scale information in the image.
6) Spatial Transformer Network (STN) helps to solve the problem of crowd rotation and deformation and is suitable for crowded and noisy scenes.

4.2.2. Evaluation metrics

The evaluation metrics are MAE (mean absolute error) and MSE (mean square error), which are as follows:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - y_i| \tag{1}$$

$$MSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}|y_i - y_i|^2} \tag{2}$$

where N indicates the number of test images, y_i denotes the correct counting of image i, and y_i is the estimated counting of image i.

5. Summary and prospect

This paper comprehensively reviewed and analyzed the recent research progress of crowd counting and crowd density estimation based on CNN and briefly reviewed the traditional methods of crowd counting. We divided the existing methods based on CNN into three categories: 1) detection-based CNN methods; 2) regression-based CNN methods; 3) combining detection and regression based CNN methods. Furthermore, in Section 3, we discuss the crowd counting method based on regression from eight perspectives. For some representative methods, we compare their results on several datasets in Section 4. It can be seen that the detection-based method can be applied well when the crowd is sparse and the person in the scene needs to be located. For large-scale dense scenes, the method based on regression can work well. The method with multi-scale information and context information can achieve better results when occlusion is serious. For video datasets, capturing the time information between adjacent frames can reduce the detection error. However, it is still very difficult to

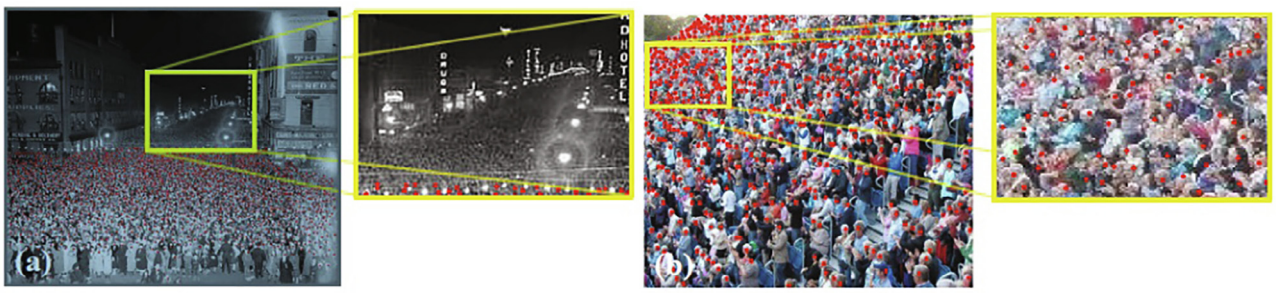


Fig. 25. (a) shows an example in which parts of the image are not annotated because it is almost impossible to distinguish between adjacent people's head, (b) showing certain location/counting errors, so it is not suitable for localization.

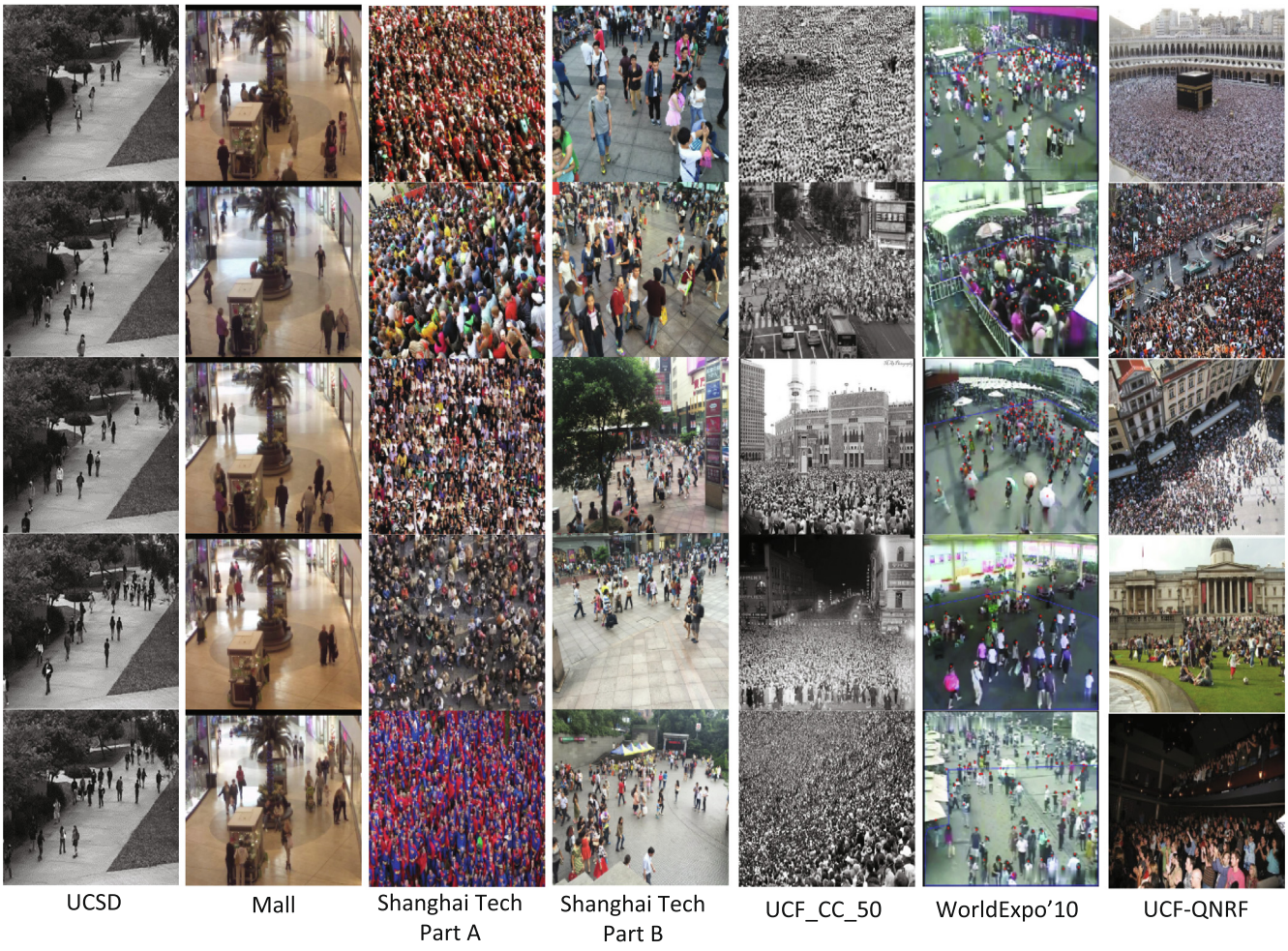


Fig. 26. Sample images on each dataset, it can be seen that Shanghai Tech Part A, UCF_CC_50, UCF-QNRF dataset has the highest density of people and is the most challenging.

Table 3
Summary of datasets

Dataset based on density graph regression (labeled as point)						
Dataset	No. of images	Ave. size	Min	Ave	Max	Total number
UCSD	2000	158*238	11	25	46	49,885
Mall	2000	320*240	13	–	53	62,325
UCF_CC_50	50	–	94	1280	4543	63,974
WorldExpo'10	3980	576*720	1	50	253	199,923
ShanghaiTech Part A	482	–	33	501	3139	241,677
ShanghaiTech Part B	716	768*1024	9	123	578	88,488
UCF-QNRF	1535	–	49	815	12,865	1,251,642
NWPU-Crowd	5190	2191*3209-	0	418	20,033	2,133,375
GCC	15,212	1080*1920	0	501	3995	7,625,843
JHU-CROWD++	4372	910*1430	0	346	25,791	1,515,005
City Street	500	1520*2704	70	–	150	63,974
Detection dataset (labeled as boding box)						
BRAINWASH	11,769	480*640	–	–	–	90,330

achieve a robust, high-performance and real-time prediction model. The challenge of crowd counting and crowd density estimation is still very huge. Through the analysis of the crowd counting methods, we can find that the deep CNN based crowd counting method is the mainstream method. These methods will further promote the development of the area of crowd counting and crowd density estimation.

Next, we will point out the development directions of crowd counting as follows.

(1) Most counting methods use LSTM and RNN to capture time series information. Due to the success of Temporal Convolutional Network (TCN) [57] in NLP, we suggest adding TCN to the counting model, which may achieve better results.
(2) Most of popular methods currently cannot achieve real-time detection, and cannot perform real-time detection on edge devices such as cameras. Therefore, it is an important research direction to study a lightweight network with high detection efficiency.

Table 4
Comparison of various methods on UCF-QNRF dataset

Dataset	UCF-QNRF	
Evaluation criterion	MAE	MSE
FHSc + MRF [23]	315	508
MCNN [12]	277	426
CMTL [21]	252	514
Switching-CNN [15]	288	445
Resnet101 [86]	190	277
CL [166]	132	191
L2SM [194]	104.7	173.6
DSSINet [198]	99.1	159.2
BL [199]	**88.7**	**154.8**

(3) NAS-FPN [53] has proven to have good performance, and is a good choice for detection-based methods to add NAS.
(4) In some scenarios, it is very difficult to label the data and consequently the dataset containing lots of labeled data can hardly be built. So many works add unlabeled data to the network, which enables the network to self-supervise or semi-supervise the training, and achieves preferable results. Therefore, it is necessary to develop a small sample size network model with high detection performance.
(5) New and elaborated loss functions are needed. The existing methods generally use L2 loss to optimize the model, but L2 loss often fails to learn spatial awareness (i.e., head position), and it is very sensitive to various noises in the image. Therefore, designing a new optimization loss method is also an important research direction.
(6) Beyond counting. Some of the recent approaches combine detection-based and regression-based methods, allowing networks to locate people in scenarios where only point-annotated dataset are used for training.
(7) Attention mechanism. The attention mechanism has been widely applied in various fields since it was proposed in 2015, and the crowd counting model based on attention mechanism has also achieved good results.
(8) Interpretability. At present, the counting methods based on CNN are unexplainable. These methods may perform well under specific scene, but their performance will degrade when the scene changes. In the future work, we hope to add interpretability such as rule reasoning to the counting model.
(9) Most crowd counting methods are aimed at a single camera perspective, and the crowd counting based on multi-view fusion is also an important research direction.

CRediT authorship contribution statement

Zizhu Fan: Writing – original draft. **Hong Zhang:** Methodology, Software. **Zheng Zhang:** Conceptualization, Methodology, Validation. **Guangming Lu:** Visualization. **Yudong Zhang:** Investigation, Methodology. **Yaowei Wang:** Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

This research was supported in part by the Natural Science Foundation of China (Nos. 61991401, 62002085), Jiangxi Provincial Natural Science Foundation of China (20192ACBL20010), and Shenzhen Fundamental Research Fund under Grant (GXWD20201230155427003-20200824103320001, JCYJ20210324132212030).

References

[1] P. Dollar, C. Wojek, B. Schiele, P. Perona, Pedestrian detection: an evaluation of the state of the art, IEEE Trans. Pattern Anal. Mach. Intell. 34 (2011) 743–761.

Table 5
Comparison of various methods on different datasets

Dataset	UCSD		MAll		UCF_CC_50		WorldExpo'10		SHTech partA		SHTech part B	
Evaluation Criteria	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE
CSCC [19]	1.60	3.31	–	–	467.0	498.5	12.9	–	181.8	277.7	32.0	49.8
MCNN [12]	1.07	1.35	–	–	377.6	509.1	11.6	–	110.2	173.2	26.4	41.3
CNN-Boosting [43]	1.10	–	2.01	–	364.4	–	–	–	–	–	–	–
CSRNet [18]	1.16	1.47	–	–	266.1	397.5	8.6	1.47	68.2	115.0	10.6	16.0
CNN-pixel [49]	1.12	2.06	–	–	406.2	404.0	13.4	–	–	–	–	–
CMTL [21]	–	–	–	–	332.8	341.4	–	–	101.3	152.4	20.0	31.1
Switching-CNN [15]	1.62	2.10	–	–	318.1	439.2	9.4	–	90.4	135.0	21.6	33.4
MSCNN [90]	–	–	–	–	363.7	468.4	11.7	–	83.8	127.4	17.7	30.2
NetVLAD [122]	–	–	–	–	311.3	401.8	10.3	–	107.6	169.3	21.4	33.9
SaCNN [132]	–	–	–	–	314.9	424.8	8.5	–	86.8	139.2	16.2	25.8
DA-Net [128]	1.03	1.31	–	–	290.8	326.5	–	–	71.6	104.9	15.0	21.9
FCNN [150]	–	–	–	–	253.1	356.4	–	–	67.6	110.6	10.1	18.8
BSAD [161]	1.00	1.40	–	–	409.5	563.7	10.5	–	–	–	20.2	35.6
IG-CNN [162]	–	–	–	–	291.4	349.4	11.3	–	72.5	118.2	13.6	21.1
IC-CNN [169]	–	–	–	–	260.9	365.5	10.3	–	68.5	116.2	10.7	16.0
SANet [170]	1.02	1.29	–	–	258.4	334.9	8.2	–	67.0	104.5	8.4	13.6
DRSAN [172]	–	–	**1.72**	**2.1**	219.2	250.2	7.76	–	69.3	96.4	11.1	18.2
HA-CCN [174]	–	–	–	**–**	256.2	348.4	–	–	62.9	94.9	8.1	13.4
PACNN [177]	**0.89**	**1.18**	–	**–**	241.7	320.7	–	–	62.4	102.2	7.6	11.8
CANet [182]	–	–	–	**–**	**212.2**	**243.7**	**7.2**	–	62.3	100.0	7.8	12.2
L2SM [194]	–	–	–	**–**	**188.4**	315.3	**–**	–	64.2	98.4	7.2	11.1
SAANet [196]	–	–	**1.28**	**1.68**	271.6	391.0	–	–	–	–	16.86	28.4
E3DNet [197]	**0.93**	**1.17**	**1.64**	**2.13**	–	–	8.32	–	–	–	–	–
DSSINet [198]	–	**–**	–	–	216.9	302.4	**6.67**	–	60.63	**96.04**	**6.85**	**10.34**
SPANet [200]	**1.00**	**1.28**	–	–	232.6	311.7	**–**	–	**59.4**	**92.5**	**6.5**	**9.9**
PGCNet [188]	–	–	–	–	244.6	361.2	8.1	–	**57.0**	**86.0**	8.8	13.7
AMSNet [71]	–	–	–	–	**208.6**	**296.3**	**6.8**	–	**58.0**	96.2	**7.1**	**10.4**

[2] J. Gall, A. Yao, N. Razavi, L. Van Gool, V. Lempitsky, Hough forests for object detection, tracking, and action recognition, IEEE Trans. Pattern Anal. Mach. Intell. 33 (2011) 2188–2202.
[3] B. Leibe, E. Seemann, B. Schiele, Pedestrian detection in crowded scenes, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 878–885.
[4] M. Enzweiler, D.M. Gavrila, Monocular pedestrian detection: Survey and experiments, IEEE Trans. Pattern Anal. Mach. Intell. 31 (12) (2009) 2179–2195.
[5] O. Tuzel, F. Porikli, P. Meer, Pedestrian detection via classification on Riemannian manifolds, IEEE Trans. Pattern Anal. Mach. Intell. 30 (2008) 1713–1727.
[6] P.F. Felzenszwalb, R.B. Girshick, D. McAllester, D. Ramanan, Object detection with discriminatively trained part-based models, IEEE Trans. Pattern Anal. Mach. Intell. 32 (2009) 1627–1645.
[7] B. Wu, R. Nevatia, Detection and tracking of multiple, partially occluded humans by Bayesian combination of edgelet based part detectors, Int. J. Comput. Vision 75 (2007) 247–266.
[8] A.B. Chan, N. Vasconcelos, Bayesian poisson regression for crowd counting, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 545–551.
[9] F. Xia, J. Wang, X. Kong, Z. Wang, J. Li, C. Liu, Exploring human mobility patterns in urban scenarios: a trajectory data perspective, IEEE Commun. Mag. 56 (3) (2018) 142–149.
[10] K. Chen, C.C. Loy, S. Gong, T. Xiang, Feature mining for localised crowd counting, in: Proceedings of the British Machine Vision Conference, 2012, pp. 1–11.
[11] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput. 9 (8) (1997) 1735–1780.
[12] Y. Zhang, D. Zhou, S. Chen, S. Gao, Y. Ma, 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.
[13] V.Q. Pham, T. Kozakaya, O. Yamaguchi, R. Okada, Count forest: co-voting uncertain number of targets using random forest for crowd density estimation, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 3253–3261.
[14] Y. Wang, Y. Zou, Fast visual object counting via example-based density estimation, in: Proceedings of the IEEE International Conference on Image Processing, 2016, pp. 3653–3657.
[15] D.B. Sam, S. Surya, R.V. Babu, Switching convolutional neural network for crowd counting, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 4031–4039.
[16] Z. Zhang, Z. Lai, Z. Huang, W.K. Wong, G.-S. Xie, L. Liu, L. Shao, Scalable supervised asymmetric hashing with semantic and latent factor embedding, IEEE Trans. Image Process. 28 (2019) 4803–4818.
[17] R. Stewart, M. Andriluka, A.Y. Ng, End-to-end people detection in crowded scenes, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2325–2333.
[18] Y. Li, X. Zhang, D. Chen, CSRNet, Dilated convolutional neural networks for understanding the highly congested scenes, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 1091–1100.
[19] C. Zhang, H. Li, X. Wang, X. Yang, Cross-scene crowd counting via deep convolutional neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 833–841.
[20] X. Liu, J. Van De Weijer, A.D. Bagdanov, Exploiting unlabeled data in CNNs by self-supervised learning to rank, IEEE Trans. Pattern Anal. Mach. Intell. 41 (2019) 1862–1878.
[21] V.A. Sindagi, V.M. Patel, Cnn-based cascaded multi-task learning of high-level prior and density estimation for crowd counting, in: Proceedings of the IEEE International Conference on Advanced Video and Signal Based Surveillance, 2017, pp. 1–6.
[22] V.A. Sindagi, V.M. Patel, Generating high-quality crowd density maps using contextual pyramid CNNs, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 1861–1870.
[23] H. Idrees, I. Saleemi, C. Seibert, M. Shah, Multi-source multi-scale counting in extremely dense crowd images, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 2547–2554.
[24] S. Zhang, G. Wu, J.P. Costeira, J.M. Moura, Fcn-rlstm: Deep spatio-temporal neural networks for vehicle counting in city cameras, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 3667–3676.
[25] X. Liu, J. Van De Weijer, A.D. Bagdanov, Leveraging unlabeled data for crowd counting by learning to rank, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7661–7669.
[26] A.B. Chan, Z.-S.-J. Liang, N. Vasconcelos, Privacy preserving crowd monitoring: Counting people without people models or tracking, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2008, pp. 1–7.
[27] S. Sun, N. Akhtar, H. Song, C. Zhang, J. Li, A. Mian, Benchmark data and method for real-time people counting in cluttered scenes using depth sensors, IEEE Trans. Intell. Transp. Syst. 20 (2019) 3599–3612.
[28] Q. Wang, J. Gao, W. Lin, Y. Yuan, Learning from synthetic data for crowd counting in the wild, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 8198–8207.
[29] Y. Liu, M. Shi, Q. Zhao, X. Wang, Point in, box out: Beyond counting persons in crowds, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 6469–6478.
[30] F. Xiong, X. Shi, D.-Y. Yeung, Spatiotemporal modeling for crowd counting in videos, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 5151–5159.
[31] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, Commun. ACM 60 (6) (2017) 84–90.
[32] V.A. Sindagi, V.M. Patel, A survey of recent advances in CNN-based single image crowd counting and density estimation, Pattern Recogn. Lett. 107 (2018) 3–16.
[33] A.C. Davies, J.H. Yin, S.A. Velastin, Crowd monitoring using image processing, Electron. Commun. Eng. J. 7 (1995) 37–47.
[34] C.C. Loy, K. Chen, S. Gong, T. Xiang, in: Crowd Counting and Profiling: Methodology and Evaluation, Modeling, Simulation and Visual Analysis of Crowds, Springer, 2013, pp. 347–382.
[35] P. Viola, M.J. Jones, Robust real-time face detection, Int. J. Comput. Vision 57 (2) (2004) 137–154.
[36] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005, pp. 886–893.
[37] B. Wu, R. Nevatia, Detection of multiple, partially occluded humans in a single image by Bayesian combination of edgelet part detectors, in: Proceedings of the IEEE International Conference on Computer Vision, 2005, pp. 90–97.
[38] P. Sabzmeydani, G. Mori, Detecting pedestrians by learning shapelet features, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2007, pp. 1–8.
[39] S.-F. Lin, J.-Y. Chen, H.-X. Chao, Estimation of number of people in crowded scenes using perspective transformation, IEEE Trans. Syst., Man, Cybern. -Part A: Syst. Hum. 31 (2001) 645–654.
[40] N. Paragios, V. Ramesh, A MRF-based approach for real-time subway monitoring, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2001, pp. 1034–1040.
[41] C. Wang, H. Zhang, L. Yang, S. Liu, X. Cao, Deep people counting in extremely dense crowds, in: Proceedings of the 23rd ACM International Conference on Multimedia, 2015, pp. 1299–1302.
[42] L. Boominathan, S.S. Kruthiventi, R.V. Babu, Crowdnet, A deep convolutional network for dense crowd counting, in: Proceedings of the 24th ACM International Conference on Multimedia, 2016, pp. 640–644.
[43] E. Walach, L. Wolf, Learning to count with CNN boosting, in: Proceedings of the European Conference on Computer Vision,, Springer, 2016, pp. 660–676.
[44] D. Onoro-Rubio, R.J. López-Sastre, Towards perspective-free object counting with deep learning, in: Proceedings of the European Conference on Computer Vision, 2016, pp. 615–629.
[45] G. Li, Y. Yu, Visual saliency based on multiscale deep features, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 5455–5463.
[46] C. Shang, H. Ai, B. Bai, End-to-end crowd counting via joint learning local and global count, in: Proceedings of the IEEE International Conference on Image Processing, 2016, pp. 1215–1219.
[47] G.-S. Xie, Z. Zhang, L. Liu, F. Zhu, X.-Y. Zhang, L. Shao, X. Li, SRSC: selective, robust, and supervised constrained feature representation for image classification, IEEE Trans. Neural Networks Learn. Syst. 31 (10) (2020) 4290–4302.
[48] K. Sirinukunwattana, S.E.A. Raza, Y.-W. Tsang, D.R. Snead, I.A. Cree, N.M. Rajpoot, Locality sensitive deep learning for detection and classification of nuclei in routine colon cancer histology images, IEEE Trans. Med. Imaging 35 (2016) 1196–1206.
[49] D. Kang, Z. Ma, A.B. Chan, Beyond counting: comparisons of density maps for crowd analysis tasks—counting, detection, and tracking, IEEE Trans. Circuits Syst. Video Technol. 29 (2018) 1408–1422.
[50] V. Lempitsky, A. Zisserman, Learning to count objects in images, Adv. Neural Inf. Process. Syst. 23 (2010) 1324–1332.
[51] D. Kang, A. Chan, Crowd counting by adaptively fusing predictions from an image pyramid, Proceedings of the British Machine Vision Conference, 2018.
[52] E. Lu, W. Xie, A. Zisserman, Class-agnostic counting, in: Proceedings of the Asian Conference on Computer Vision, 2018, pp. 669–684.
[53] G. Ghiasi, T.-Y. Lin, Q.V. Le, Nas-fpn, Learning scalable feature pyramid architecture for object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 7036–7045.
[54] J. Shao, K. Kang, C. Change Loy, X. Wang, Deeply learned attributes for crowded scene understanding, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 4657–4666.
[55] B. Zhou, X. Wang, X. Tang, Understanding collective crowd behaviors: learning a mixture model of dynamic pedestrian-agents, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2012, pp. 2871–2878.
[56] Y. Sun, Z. Zhang, W. Jiang, Z. Zhang, L. Zhang, S. Yan, M. Wang, Discriminative local sparse representation by robust adaptive dictionary pair learning, IEEE Trans. Neural Networks Learn. Syst. 31 (10) (2020) 4303–4317.
[57] C. Lea, M.D. Flynn, R. Vidal, A. Reiter, G.D. Hager, Temporal convolutional networks for action segmentation and detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 156–165.
[58] V.A. Sindagi, R. Yasarla, V.M. Patel, JHU-CROWD++: Large-Scale crowd counting dataset and a benchmark method, arXiv preprint arXiv:2004.03597, (2020).
[59] Z. Shen, Y. Xu, B. Ni, M. Wang, J. Hu, X. Yang, Crowd counting via adversarial cross-scale consistency pursuit, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 5245–5254.

[60] B. Zhou, X. Tang, X. Wang, Learning collective crowd behaviors with dynamic pedestrian-agents, Int. J. Comput. Vision 111 (2015) 50–68.
[61] W. Li, V. Mahadevan, N. Vasconcelos, Anomaly detection and localization in crowded scenes, IEEE Trans. Pattern Anal. Mach. Intell. 36 (2013) 18–32.
[62] R. Chaker, Z. Al Aghbari, I.N. Junejo, Social network model for crowd anomaly detection and localization, Pattern Recogn. 61 (2017) 266–281.
[63] A. Finogeev, A. Finogeev, L. Fionova, A. Lyapin, K.A. Lychagin, Intelligent monitoring system for smart road environment, Journal of Industrial Information, Integration 15 (2019) 15–20.
[64] A. Abdelghany, K. Abdelghany, H. Mahmassani, W. Alhalabi, Modeling framework for optimal evacuation of large-scale crowded pedestrian facilities, Eur. J. Oper. Res. 237 (2014) 1105–1118.
[65] Q. Wang, J. Gao, W. Lin, X. Li, Nwpu-crowd: A large-scale benchmark for crowd counting, arXiv preprint arXiv:2001.03360, (2020).
[66] W.K. Chow, C.M. Ng, Waiting time in emergency evacuation of crowded public transport terminals, Saf. Sci. 46 (2008) 844–857.
[67] W. Li, H. Li, Q. Wu, F. Meng, L. Xu, K.N. Ngan, Headnet: An end-to-end adaptive relational network for head detection, IEEE Trans. Circuits Syst. Video Technol. 30 (2) (2020) 482–494.
[68] L. Lu, C.-Y. Chan, J. Wang, W. Wang, A study of pedestrian group behaviors in crowd evacuation based on an extended floor field cellular automaton model, Transp. Res. Part C: Emerg. Technol. 81 (2017) 317–329.
[69] Z. Zhang, L. Liu, F. Shen, H.T. Shen, L. Shao, Binary multi-view clustering, IEEE Trans. Pattern Anal. Mach. Intell. 41 (2018) 1774–1782.
[70] Y. Fang, B. Zhan, W. Cai, S. Gao, B. Hu, Locality-constrained spatial transformer network for video crowd counting, in: Proceedings of the IEEE International Conference on Multimedia and Expo, 2019, pp. 814–819.
[71] Y. Hu, X. Jiang, X. Liu, B. Zhang, J. Han, X. Cao, D. Doermann, NAS-Count: Counting-by-density with neural architecture search, arXiv preprint arXiv:2003.00217, (2020).
[72] T.M. Lawrence, M.-C. Boudreau, L. Helsen, G. Henze, J. Mohammadpour, D. Noonan, D. Patteeuw, S. Pless, R.T. Watson, Ten questions concerning integrating smart buildings into the smart grid, Build. Environ. 108 (2016) 273–283.
[73] S.A.M. Saleh, S.A. Suandi, H. Ibrahim, Recent survey on crowd density estimation and counting for visual surveillance, Eng. Appl. Artif. Intell. 41 (2015) 103–114.
[74] G. Olmschenk, J. Chen, H. Tang, Z. Zhu, Dense crowd counting convolutional neural networks with minimal data using semi-supervised dual-goal generative adversarial networks, CVPR Workshops, 2019.
[75] T. Li, H. Chang, M. Wang, B. Ni, R. Hong, S. Yan, Crowded scene analysis: a survey, IEEE Trans. Circuits Syst. Video Technol. 25 (3) (2015) 367–386.
[76] J.R. Barr, K.W. Bowyer, P.J. Flynn, The effectiveness of face detection algorithms in unconstrained crowd scenes, in: Proceedings of the IEEE Winter Conference on Applications of Computer Vision, 2014, pp. 1020–1027.
[77] G. French, M. Fisher, M. Mackiewicz, C. Needle, Convolutional neural networks for counting fish in fisheries surveillance video, Proceedings of the Machine Vision of Animals and their Behaviour, 2015.
[78] B. Zhan, D.N. Monekosso, P. Remagnino, S.A. Velastin, L.-Q. Xu, Crowd analysis: a survey, Mach. Vis. Appl. 19 (5-6) (2008) 345–357.
[79] M. Fu, P. Xu, X. Li, Q. Liu, M. Ye, C. Zhu, Fast crowd density estimation with convolutional neural networks, Eng. Appl. Artif. Intell. 43 (2015) 81–88.
[80] C. Doersch, A. Gupta, A.A. Efros, Unsupervised visual representation learning by context prediction, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1422–1430.
[81] X. Liu, J. van de Weijer, A.D. Bagdanov, Rankiqa, Learning from rankings for no-reference image quality assessment, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 1040–1049.
[82] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, A.A. Efros, Context encoders: Feature learning by inpainting, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2536–2544.
[83] R. Zhang, P. Isola, A.A. Efros, Colorful image colorization, in: Proceedings of the European Conference on Computer Vision, 2016, pp. 649–666.
[84] M. Noroozi, H. Pirsiavash, P. Favaro, Representation learning by learning to count, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 5898–5906.
[85] J. Liu, C. Gao, D. Meng, A.G. Hauptmann, Decidenet, Counting varying density crowds through attention guided detection and density estimation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 5197–5206.
[86] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.
[87] M.S. Zitouni, H. Bhaskar, J. Dias, M.E. Al-Mualla, Advances and trends in visual crowd analysis: a systematic survey and evaluation of crowd modelling techniques, Neurocomputing 186 (2016) 139–159.
[88] S.R. Richter, Z. Hayder, V. Koltun, Playing for benchmarks, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2213–2222.
[89] S.R. Richter, V. Vineet, S. Roth, V. Koltun, Playing for data: Ground truth from computer games, in: Proceedings of the European Conference on Computer Vision, 2016, pp. 102–118.
[90] L. Zeng, X. Xu, B. Cai, S. Qiu, T. Zhang, Multi-scale convolutional neural networks for crowd counting, in: Proceedings of the IEEE International Conference on Image Processing, 2017, pp. 465–469.
[91] J.-Y. Zhu, T. Park, P. Isola, A.A. Efros, Unpaired image-to-image translation using cycle-consistent adversarial networks, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2223–2232.
[92] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, Computer Science (2014).
[93] Z. Wang, A.C. Bovik, H.R. Sheikh, E.P. Simoncelli, Image quality assessment: from error visibility to structural similarity, IEEE Trans. Image Process. 13 (4) (2004) 600–612.
[94] Y. Lei, Y. Liu, P. Zhang, L. Liu, Towards using count-level weak supervision for crowd counting, arXiv preprint arXiv:2003.00164, (2020).
[95] Z. Li, L. Zhang, Y. Fang, J. Wang, H. Xu, B. Yin, H. Lu, Deep people counting with faster r-cnn and correlation tracking, in: Proceedings of the International Conference on Internet Multimedia Computing and Service, 2016, pp. 57–60.
[96] S. Ren, K. He, R. Girshick, J. Sun, Faster r-cnn: Towards real-time object detection with region proposal networks, IEEE Trans. Pattern Anal. Mach. Intell. 39 (6) (2017) 1137–1149.
[97] J.F. Henriques, R. Caseiro, P. Martins, J. Batista, High-speed tracking with kernelized correlation filters, IEEE Trans. Pattern Anal. Mach. Intell. 37 (2014) 583–596.
[98] A. Shrivastava, A. Gupta, R. Girshick, Training region-based object detectors with online hard example mining, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 761–769.
[99] X. Huang, Y. Zou, Y. Wang, Cost-sensitive sparse linear regression for crowd counting with imbalanced training data, in: Proceedings of the IEEE International Conference on Multimedia and Expo, 2016, pp. 1–6.
[100] Z. Zhang, M. Wang, X. Geng, Crowd counting in public video surveillance by label distribution learning, Neurocomputing 166 (2015) 151–163.
[101] B. Xu, G. Qiu, Crowd density estimation based on rich features and random projection forest, in: Proceedings of the IEEE Winter Conference on Applications of Computer Vision WACV, 2016, pp. 1–8.
[102] J. Wen, Z. Zhong, Z. Zhang, L. Fei, Z. Lai, R. Chen, Adaptive locality preserving regression, IEEE Trans. Circuits Syst. Video Technol. 30 (2020) 75–88.
[103] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, J. Sivic, NetVLAD: CNN architecture for weakly supervised place recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 5297–5307.
[104] A. Marana, L.d.F. Costa, R. Lotufo, S. Velastin, On the efficacy of texture analysis for crowd monitoring, in: Proceedings SIBGRAPI'98. International Symposium on Computer Graphics, Image Processing, and Vision, 1998, pp. 354–361.
[105] C.S. Regazzoni, A. Tesei, Distributed data fusion for real-time crowding estimation, Signal Process. 53 (1996) 47–63.
[106] Z. Zhao, H. Li, R. Zhao, X. Wang, Crossing-line crowd counting with two-phase deep neural networks, in: Proceedings of the European Conference on Computer Vision, 2016, pp. 712–726.
[107] L. Fiaschi, U. Köthe, R. Nair, F.A. Hamprecht, Learning to count with regression forest and structured labels, in: Proceedings of the 21st International Conference on Pattern Recognition, 2012, pp. 2685–2688.
[108] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 1–9.
[109] M. Lin, Q. Chen, S. Yan, Network in network, Proceedings of the International Conference on Learning Representations, 2014.
[110] X. Wei, J. Du, M. Liang, L. Ye, Boosting deep attribute learning via support vector regression for fast moving crowd counting, Pattern Recogn. Lett. 119 (2019) 12–23.
[111] A. Radford, L. Metz, S. Chintala, Unsupervised representation learning with deep convolutional generative adversarial networks, Proceedings of the International Conference on Learning Representations, 2016.
[112] M. Marsden, K. McGuinness, S. Little, N.E. O'Connor, ResnetCrowd, A residual deep learning architecture for crowd counting, violent behaviour detection and crowd density level classification, in: Proceedings of the IEEE International Conference on Advanced Video and Signal Based Surveillanc, 2017, pp. 1–7.
[113] S. Ioffe, Batch renormalization: towards reducing minibatch dependence in batch-normalized models, in: Advances in Neural Information Processing Systems, 2017, pp. 1945–1953.
[114] T.N. Mundhenk, G. Konjevod, W.A. Sakla, K. Boakye, A large contextual dataset for classification, detection and counting of cars with deep learning, in: Proceedings of the European Conference on Computer Vision, 2016, pp. 785–800.
[115] S. Zhang, G. Wu, J.P. Costeira, J.M. Moura, Understanding traffic density from large-scale web camera data, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 5898–5907.
[116] X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, W.-C. Woo, Convolutional LSTM network: a machine learning approach for precipitation nowcasting, Adv. Neural Inf. Process. Syst. 28 (2015) 802–810.
[117] Y. Huang, W. Wang, L. Wang, Bidirectional recurrent convolutional networks for multi-frame super-resolution, Adv. Neural Inf. Process. Syst. 28 (2015) 235–243.
[118] Y. Zhang, W. Chan, N. Jaitly, Very deep convolutional networks for end-to-end speech recognition, in: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, 2017, pp. 4845–4849.

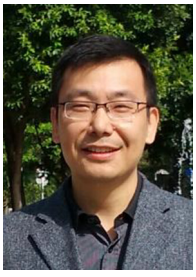
[119] B. Sheng, C. Shen, G. Lin, J. Li, W. Yang, C. Sun, Crowd counting via weighted VLAD on a dense attribute feature map, IEEE Trans. Circuits Syst. Video Technol. 28 (2016) 1788–1797.
[120] J. Wen, Z. Zhang, Z. Zhang, L. Fei, M. Wang, Generalized incomplete multi-view clustering with flexible locality structure diffusion, IEEE Trans. Cybern. 51 (2021) 101–114.
[121] G. Lin, C. Shen, A. Van Den Hengel, I. Reid, Efficient piecewise training of deep structured models for semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 3194–3203.
[122] Z. Shi, L. Zhang, Y. Sun, Y. Ye, Multiscale multitask deep NetVLAD for crowd counting, IEEE Trans. Ind. Inf. 14 (2018) 4953–4962.
[123] M.S. Parwez, D.B. Rawat, M. Garuba, Big data analytics for user-activity analysis and user-anomaly detection in mobile wireless network, IEEE Trans. Ind. Inf. 13 (2017) 2058–2065.
[124] K. Chen, S. Gong, T. Xiang, C. Change Loy, 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.
[125] D. Ryan, S. Denman, S. Sridharan, C. Fookes, An evaluation of crowd counting methods, features and regression models, Comput. Vis. Image Underst. 130 (2015) 1–17.
[126] Z. Shi, Y. Ye, Y. Wu, Rank-based pooling for deep convolutional neural networks, Neural Networks 83 (2016) 21–31.
[127] R. Arandjelovic, A. Zisserman, All about VLAD, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 1578–1585.
[128] Z. Zou, X. Su, X. Qu, P. Zhou, Da-net: Learning the fine-grained density distribution with deformation aggregation network, IEEE Access 6 (2018) 60745–60756.
[129] P. Viola, M.J. Jones, D. Snow, Detecting pedestrians using patterns of motion and appearance, Int. J. Comput. Vision 63 (2) (2005) 153–161.
[130] Y. Cai, Z. Liu, H. Wang, X. Sun, Saliency-based pedestrian detection in far infrared images, IEEE Access 5 (2017) 5013–5019.
[131] B. Yang, J.-M. Cao, N. Wang, Y.-Y. Zhang, G.-Z. Cui, Cross-scene counting based on domain adaptation-extreme learning machine, IEEE Access 6 (2018) 17029–17038.
[132] L. Zhang, M. Shi, Q. Chen, Crowd counting via scale-adaptive convolutional neural network, in: Proceedings of the IEEE Winter Conference on Applications of Computer Vision, 2018, pp. 1113–1121.
[133] H. Idrees, K. Soomro, M. Shah, Detecting humans in dense crowds using locally-consistent scale prior and global occlusion reasoning, IEEE Trans. Pattern Anal. Mach. Intell. 37 (10) (2015) 1986–1998.
[134] M. Li, Z. Zhang, K. Huang, T. Tan, Estimating the number of people in crowded scenes by mid based foreground segmentation and head-shoulder detection, in: Proceedings of the International Conference on Pattern Recognition, 2018, pp. 1–4.
[135] B. Liu, N. Vasconcelos, Bayesian model adaptation for crowd counts, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 4175–4183.
[136] J. Sang, W. Wu, H. Luo, H. Xiang, Q. Zhang, H. Hu, X. Xia, Improved crowd counting method based on scale-adaptive convolutional neural network, IEEE Access 7 (2019) 24411–24419.
[137] J. Shao, C. Change Loy, X. Wang, Scene-independent group profiling in crowd, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 2219–2226.
[138] V. Mahadevan, W. Li, V. Bhalodia, N. Vasconcelos, Anomaly detection in crowded scenes, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2010, pp. 1975–1981.
[139] F. Zhu, X. Wang, N. Yu, Crowd tracking with dynamic evolution of group structures, in: Proceedings of the European Conference on Computer Vision, 2014, pp. 139–154.
[140] G. Olmschenk, H. Tang, Z. Zhu, Crowd counting with minimal data using generative adversarial networks for multiple target regression, in: Proceedings of the IEEE Winter Conference on Applications of Computer Vision, 2018, pp. 1151–1159.
[141] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, Adv. Neural Inf. Process. Syst. 27 (2014) 2672–2680.
[142] D.B. Sam, S.V. Peri, M.N. Sundararaman, A. Kamath, V.B. Radhakrishnan, Locate, size and count: accurately resolving people in dense crowds via detection, IEEE Trans. Pattern Anal. Mach. Intell. (2020).
[143] X. Ding, Z. Lin, F. He, Y. Wang, Y. Huang, A deeply-recursive convolutional network for crowd counting, in: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, 2018, pp. 1942–1946.
[144] D. Eigen, J. Rolfe, R. Fergus, Y. LeCun, Understanding deep architectures using a recursive convolutional network, Proceedings of the International Conference on Learning Representations, 2013.
[145] D. Kong, D. Gray, H. Tao, A viewpoint invariant approach for crowd counting, in: Proceedings of the International Conference on Pattern Recognition, 2006, pp. 1187–1190.
[146] J. Yang, Y. Zhou, S.-Y. Kung, Multi-scale generative adversarial networks for crowd counting, in: Proceedings of the International Conference on Pattern Recognition, 2018, pp. 3244–3249.
[147] S. Wang, E. Zhu, J. Yin, F. Porikli, Anomaly detection in crowded scenes by SL-HOF descriptor and foreground classification, in: Proceedings of the International Conference on Pattern Recognition, 2016, pp. 3398–3403.
[148] K. Nakamura, T. Ono, N. Babaguchi, Detection of groups in crowd considering their activity state, in: Proceedings of the International Conference on Pattern Recognition, 2016, pp. 277–282.
[149] Z. Lin, L.S. Davis, Shape-based human detection and segmentation via hierarchical part-template matching, IEEE Trans. Pattern Anal. Mach. Intell. 32 (2010) 604–618.
[150] M. Liu, J. Jiang, Z. Guo, Z. Wang, Y. Liu, Crowd counting with fully convolutional neural network, in: Proceedings of the IEEE International Conference on Image Processing, 2018, pp. 953–957.
[151] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, S. Belongie, Feature pyramid networks for object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 2117–2125.
[152] C. Liu, X. Weng, Y. Mu, Recurrent attentive zooming for joint crowd counting and precise localization, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 1217–1226.
[153] S. Amirgholipour, X. He, W. Jia, D. Wang, M. Zeibots, A-CCNN: Adaptive CCNN for density estimation and crowd counting, in: Proceedings of the IEEE International Conference on Image Processing, 2018, pp. 948–952.
[154] D. Deb, J. Ventura, An aggregated multicolumn dilated convolution network for perspective-free counting, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 195–204.
[155] J. Selinummi, J. Seppälä, O. Yli-Harja, J.A. Puhakka, Software for quantification of labeled bacteria from digital microscope images by automated image analysis, Biotechniques 39 (2005) 859–863.
[156] Y. Fang, S. Gao, J. Li, W. Luo, L. He, B. Hu, Multi-level feature fusion based locality-constrained spatial transformer network for video crowd counting, Neurocomputing 392 (2020) 98–107.
[157] Z. Zhang, L. Liu, Y. Luo, Z. Huang, F. Shen, H.T. Shen, G. Lu, Inductive structure consistent hashing via flexible semantic calibration, IEEE Trans. Neural Networks Learn. Syst. 32 (10) (2021) 4514–4528.
[158] J.C.S.J. Junior, S.R. Musse, C.R. Jung, Crowd analysis using computer vision techniques, IEEE Signal Process Mag. 27 (2010) 66–77.
[159] P. Isola, J.-Y. Zhu, T. Zhou, A.A. Efros, Image-to-image translation with conditional adversarial networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1125–1134.
[160] O. Ronneberger, P. Fischer, T. Brox, U-net, Convolutional networks for biomedical image segmentation, in: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, 2015, pp. 234–241.
[161] S. Huang, X. Li, Z. Zhang, F. Wu, S. Gao, R. Ji, J. Han, Body structure aware deep crowd counting, IEEE Trans. Image Process. 27 (2017) 1049–1059.
[162] D. Babu Sam, N.N. Sajjan, R. Venkatesh Babu, M. Srinivasan, Divide and grow: Capturing huge diversity in crowd images with incrementally growing cnn, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 3618–3626.
[163] M. Wang, X. Wang, Automatic adaptation of a generic pedestrian detector to a specific traffic scene, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2011, pp. 3401–3408.
[164] Z. Shi, L. Zhang, Y. Liu, X. Cao, Y. Ye, M.-M. Cheng, G. Zheng, Crowd counting with deep negative correlation learning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 5382–5390.
[165] T. Zhao, R. Nevatia, B. Wu, Segmentation and tracking of multiple humans in crowded environments, IEEE Trans. Pattern Anal. Machi. Intell. 30 (2008) 1198–1211.
[166] H. Idrees, M. Tayyab, K. Athrey, D. Zhang, S. Al-Maadeed, N. Rajpoot, M. Shah, Composition loss for counting, density map estimation and localization in dense crowds, in: Proceedings of the European Conference on Computer Vision, 2018, pp. 532–546.
[167] Z. Li, Z. Zhang, J. Qin, Z. Zhang, L. Shao, Discriminative fisher embedding dictionary learning algorithm for object recognition, IEEE Trans. Neural Networks Learn. Syst. 31 (2019) 786–800.
[168] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, Y. LeCun, Overfeat: Integrated recognition, localization and detection using convolutional networks, arXiv preprint arXiv:1312.6229, (2013).
[169] V. Ranjan, H. Le, M. Hoai, Iterative crowd counting, in: Proceedings of the European Conference on Computer Vision, 2018, pp. 270–285.
[170] X. Cao, Z. Wang, Y. Zhao, F. Su, Scale aggregation network for accurate and efficient crowd counting, in: Proceedings of the European Conference on Computer Vision, 2018, pp. 734–750.
[171] D.B. Sam, R.V. Babu, Top-down feedback for crowd counting convolutional neural network, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2018, pp. 7323–7330.
[172] L. Liu, H. Wang, G. Li, W. Ouyang, L. Lin, Crowd counting using deep recurrent spatial-aware network, in: Proceedings of the International Joint Conference on Artificial Intelligence, 2018, pp. 849–855.
[173] Z. Shi, P. Mettes, C.G. Snoek, Counting with focus for free, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 4200–4209.
[174] V.A. Sindagi, V.M. Patel, Ha-CCN: Hierarchical attention-based crowd counting network, IEEE Trans. Image Process. 29 (2020) 323–335.
[175] N. Liu, Y. Long, C. Zou, Q. Niu, L. Pan, H. Wu, Adcrowdnet: an attention-injective deformable convolutional network for crowd understanding, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 3225–3234.

[176] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, Computer Vision and Pattern Recognition, IEEE, Piscataway, 2018.
[177] M. Shi, Z. Yang, C. Xu, Q. Chen, Revisiting perspective information for efficient crowd counting, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 7279–7288.
[178] G.J. Brostow, R. Cipolla, Unsupervised Bayesian detection of independent motion in crowds, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2006, pp. 594–601.
[179] V. Rabaud, S. Belongie, Counting crowded moving objects, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2006, pp. 705–711.
[180] P. Viola, M. Jones, Rapid object detection using a boosted cascade of simple features, in: Proceedings of the IEEE computer society conference on computer vision and pattern recognition, 2001, pp. 511–518.
[181] A.B. Chan, N. Vasconcelos, Counting people with low-level features and Bayesian regression, IEEE Trans. Image Process. 21 (2011) 2160–2177.
[182] W. Liu, M. Salzmann, P. Fua, Context-aware crowd counting, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 5099–5108.
[183] D.B. Sam, N.N. Sajjan, H. Maurya, R.V. Babu, Almost unsupervised learning for dense crowd counting, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2019, pp. 8868–8875.
[184] X. Jiang, Z. Xiao, B. Zhang, X. Zhen, X. Cao, D. Doermann, L. Shao, Crowd counting and density estimation by trellis encoder-decoder networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 6133–6142.
[185] W. Ge, R.T. Collins, Marked point processes for crowd counting, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 2913–2920.
[186] M. Zhao, J. Zhang, C. Zhang, W. Zhang, Leveraging heterogeneous auxiliary tasks to assist crowd counting, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 12736–12745.
[187] Q. Zhang, A.B. Chan, Wide-area crowd counting via ground-plane density maps and multi-view fusion CNNs, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 8297–8306.
[188] Z. Yan, Y. Yuan, W. Zuo, X. Tan, Y. Wang, S. Wen, E. Ding, Perspective-guided convolution networks for crowd counting, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 952–961.
[189] Z. Ma, L. Yu, A.B. Chan, Small instance detection by integer programming on object density maps, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3689–3697.
[190] J. Wan, W. Luo, B. Wu, A.B. Chan, W. Liu, Residual regression with semantic prior for crowd counting, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 4036–4045.
[191] F. Sung, Y. Yang, L. Zhang, T. Xiang, P.H. Torr, T.M. Hospedales, Learning to compare: Relation network for few-shot learning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 1199–1208.
[192] D. Lian, J. Li, J. Zheng, W. Luo, S. Gao, Density map regression guided detection network for rgb-d crowd counting and localization, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 1821–1830.
[193] T.-Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, Focal loss for dense object detection, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2980–2988.
[194] C. Xu, K. Qiu, J. Fu, S. Bai, Y. Xu, X. Bai, Learn to scale: Generating multipolar normalized density maps for crowd counting, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 8382–8390.
[195] J. Yao, J.-M. Odobez, Multi-layer background subtraction based on color and texture, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2007, pp. 1–8.
[196] M. Hossain, M. Hosseinzadeh, O. Chanda, Y. Wang, Crowd counting using scale-aware attention networks, in: Proceedings of the IEEE Winter Conference on Applications of Computer Vision, 2019, pp. 1280–1288.
[197] Z. Zou, H. Shao, X. Qu, W. Wei, P. Zhou, Enhanced 3D convolutional networks for crowd counting, arXiv preprint arXiv:1908.04121, (2019).
[198] L. Liu, Z. Qiu, G. Li, S. Liu, W. Ouyang, L. Lin, Crowd counting with deep structured scale integration network, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 1774–1783.
[199] Z. Ma, X. Wei, X. Hong, Y. Gong, in: Bayesian loss for crowd count estimation with point supervision, in, 2019, pp. 6142–6151.
[200] Z.-Q. Cheng, J.-X. Li, Q. Dai, X. Wu, A.G. Hauptmann, Learning spatial awareness to improve crowd counting, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 6152–6161.
[201] S. Xu, R. Zhang, W. Cheng, J. Xu, MTLM: a multi-task learning model for travel time estimation, GeoInformatica (2020) 1–17.

Zizhu Fan received the Ph.D. degree in computer science in computer science & technology at Shenzhen graduate school, Harbin Institute of Technology (HIT), China, in 2014. Now he is a professor at School of Basic Science in East China Jiaotong University. His current interests include pattern recognition and image processing. He has published more than 50 journal papers.

Hong Zhang received the Bachelor degree and now is postgraduate student in computer science in computer science & technology at School of Basic Science in East China Jiaotong University. His current interests include pattern recognition and deep learning.

Zheng Zhang received his M.S. degree in Computer Science and Ph.D. degree in Computer Applied Technology from the Harbin Institute of Technology, China, in 2014 and 2018, respectively. Dr. Zhang was a Postdoctoral Research Fellow at The University of Queensland, Australia. He is currently an Assistant Professor at Harbin Institute of Technology, Shenzhen, China. He has published over 80 technical papers at prestigious international journals and conferences, including IEEE TPAMI, TIP, TNNLS, CVPR, ECCV, AAAI, ACMM, SIGIR, IJCAI,etc. He is an Editorial Board Member of Information Processing & Management Journal. He serves/served as an SPC/PC member of several top conferences. His current research interests include machine learning, computer vision and multimedia analytics. He is an IEEE Senior Member.



Guangming Lu received the B.S. degree in electrical engineering, the M.S. degree in control theory and control engineering, and the Ph.D. degree in computer science and engineering from the Harbin Institute of Technology (HIT), Harbin, China, in 1998, 2000, and 2005, respectively. He is currently a Professor with Biocomputing Research Center, Shenzhen Graduate School, HIT, Shenzhen, China. His current research interests include pattern recognition, image processing, and automated biometric technologies and applications.

Yu-Dong Zhang received his PhD degree in Signal and Information Processing from Southeast University in 2010. He worked as a postdoc from 2010 to 2012 with Columbia University, USA; and as an assistant research scientist from 2012 to 2013 with Research Foundation of Mental Hygiene (RFMH), USA. He served as a Full Professor from 2013 to 2017 with Nanjing Normal University. Now he serves as Professor with Department of Informatics, University of Leicester, UK. His research interests include deep learning and medical image analysis. Prof. Zhang is the Fellow of IET (FIET), and Senior Members of IEEE and ACM. He was included in "Most Cited Chinese researchers (Computer Science)" by Elsevier from 2014 to 2018. He was the 2019 recipient of "Highly Cited Researcher" by Web of Science. He won "Emerald Citation of Excellence 2017" and "MDPI Top 10 Most Cited Papers 2015". He was included in "Top Scientist" in Guide2Research. He is the author of over 250 peer-reviewed articles, including more than 30 "ESI Highly Cited Papers", and 3 "ESI Hot Papers". His citation reached 13118 in Google Scholar with h-index of 63, and 7779 in Web of Science with h-index of 50. He has conducted many successful industrial projects and academic grants from NSFC, NIH, Royal Society, EPSRC, MRC, and British Council.

Yaowei Wang received the Ph.D. degree in computer science from Graduate University, Chinese Academy of Sciences, in 2005.He is currently an associate researcher at Peng Cheng Laboratory, Shenzhen, China. He was an Assistant Professor with the School of Information and Electronics, Beijing Institute of Technology, and also was a Guest Assistant Professor with the National Engineering Laboratory for Video Technology, Peking University, China. He has been the author or co-author of over 50 refereed journals and conference papers. His research interests include machine learning and multimedia content analysis and understanding. His team was ranked as one of the best performers in the TRECVID CCD/SED tasks from 2009 to 2012 and PETS 2012. He is a member of CIE.