Pattern Analysis and Applications
https://doi.org/10.1007/s10044-021-00959-z

SURVEY

Approaches on crowd counting and density estimation: a review

Bo Li[1] · Hongbo Huang[2,3] · Ang Zhang[4] · Peiwen Liu[1] · Cheng Liu[1]

Received: 8 May 2020 / Accepted: 24 January 2021

Abstract
In recent years, urgent needs for counting crowds and vehicles have greatly promoted research of crowd counting and density estimation. Benefiting from the rapid development of deep learning, the counting performance has been greatly improved, and the application scenarios have been further expanded. Aiming to deeply understand the development status of crowd counting and density estimation, we introduce and analyze the typical methods in this field and especially focus on elaborating deep learning-based counting methods. We summarize the existing approaches into four categories, i.e., detection-based, regression-based, convolutional neural network based and video-based. Each category is explicated in great detail. To provide more concrete reference, we compare the performance of typical methods on the popular benchmarks. We further elaborate on the datasets and metrics for the crowd counting community and discuss the work of solving the problem of small-sample-based counting, dataset annotation methods and so on. Finally, we summarize various challenges facing crowd counting and their corresponding solutions and propose a set of development trends in the future.

Keywords Crowd counting · Density estimation · Density map · Convolutional neural network · Deep learning

1 Introduction

Crowd counting and density estimation have been challenging tasks in image and video analysis for many years. Accurate crowd counting is helpful for pedestrian flow analysis and crowd density estimation and has a wide range of applications such as public safety, smart transportation and video surveillance. Early crowd counting and density estimation approaches are mainly based on pedestrian detection [1–3]. In crowded scenes, due to the factors such as mutual occlusions and varying scales, the performance of these methods is difficult to achieve satisfactory results, making it difficult to be adopted in practical applications. An alternative method is to estimate the crowd density in the image and finally give a crowd density level [4–10]. However, the quality of this crowd density classification method is relatively rough, which limits its application in many scenarios. In recent years, with the rapid development of deep learning, the performance of crowd counting has made great progress, and the counting accuracy and speed have been significantly improved under crowded conditions. In order to sort out the research methods and the evolution of crowd counting approaches, this paper reviews the main ideas and methods of crowd counting. Especially, we conduct a detailed review of the latest methods based on deep learning.

This paper divides the development of crowd counting and density estimation into four branches, i.e., detection-based methods, regression-based methods, CNN-based methods and video-based methods. The detection-based methods count the number of objects through an object detector trained on the extracted image features. The methods work well in low-density scenarios, but as the density of the crowd increases, the performance of such methods decreases accordingly. In the case of dense crowds, researchers found that learning a mapping between image features to the number of individuals is helpful and the performance of the methods based on these mappings outperforms the detection-based methods [8]. These methods usually train a

✉ Hongbo Huang
hhb@bistu.edu.cn

1 School of Electromechanical Engineering, Beijing Information Science and Technology University, Beijing 100192, China

2 Computer School, Beijing Information Science and Technology University, Beijing 100192, China

3 Institute of Computing Intelligence, Beijing Information Science and Technology University, Beijing 100192, China

4 School of Information Management, Beijing Information Science and Technology University, Beijing 100192, China

Published online: 20 February 2021

regression model from the learned mappings and are called as regression-based methods. However, the methods rely heavily on the hand-crafted features, and lack of robustness in the scenarios with large changes in light, perspective, crowd distribution, crowd density, etc. Given the powerful feature extraction capabilities of CNNs in deep learning, researchers tried to use them to automatically extract features and trained an end-to-end network to count individuals. The methods can adapt to changes in various factors, predict the number of individuals more accurately and achieved the state of the art on many popular evaluation benchmarks. In this work, we mainly focus on the CNN-based methods.

As just mentioned, crowd counting is still facing many challenges, such as severe occlusions, changing scenes, complex noise, various scales, different perspectives and non-uniform distributions of individuals. In order to provide datasets and benchmarks that are as close to the actual scene as possible, researchers have built many datasets of crowd counting. The most popular datasets are UCSD [11], Mall [12], UCF_CC_50 [13], WorldExpo 10 [14], ShanghaiTech [15], UCF-QNRF [16], the newly proposed GCC [17], etc. These datasets greatly promoted the development of crowd counting. However, compared with the data required for practical applications, the situation of shorting data still exists. Many methods are studying how to use less labeled data to accurately count the number of individuals, among which the typical ones are L2R [18], GWTA-CCNN [19], CAC [20], SL2R [21], SFCN [17], which will be explained in more detail later.

To the best of our knowledge, there are already some reviews and evaluations in the field of crowd counting. Ref. [22–26] are some of the earlier ones released in 2015 and before. Considering the rapid advance of crowd counting, they do not cover the new current research works. Recent review works like Ref. [27–35] review this field from different perspectives and classification methods, such as Ref. [27, 34] sort and analyze existing CNN-based works according to the network structure and inference methodology. In addition, Luo et al. [32] classified them according to the supervision standpoint. Compared with the above works, we put more emphasis on the current difficulties and their corresponding processing strategies. Furthermore, we elaborate on the datasets and metrics of crowd counting and discuss the works of solving the problem of small-sample-based counting, dataset annotation methods and so on. Finally, we discuss various challenges facing crowd counting and their corresponding solutions, as well as propose a set of development trends in the future.

The main contributions of this paper are three aspects:

1. We systematically review almost all popular methods in crowd counting and crowd density estimation in nearly 20 years and discuss their characteristics in great detail.
2. We propose a novel classification criterion to classify the CNN-based crowd counting methods.
3. We discuss the challenges in crowd counting and predict the potential trends and promising directions in the future.

The following paper is organized as follows: The second section analyzes the main works of the four branches in crowd counting. We first briefly review detection-based and regression-based methods and then focus on CNN-based methods. The third part introduces some of the popular datasets and related works, as well as some commonly used evaluation criteria. The fourth section discusses the application and future work of crowd counting.

2 Methodologies

Existing crowd counting methods can be divided into three major categories: detection-based methods, regression-based methods and CNN-based methods. The regression-based methods can be further divided into individual-based and density map-based methods. We also list some typical works in crowd counting in the form of a timeline, as shown in Fig. 1.

2.1 Detection-based methods

Most early crowd counting works are based on detection. They use the features extracted by the elaborately designed detector to count targets. Detection-based methods strongly rely on the characteristics of the targets. Feature extraction methods can be divided into integral-based and parts-based. The integral-based detection methods first extract the features of the entire image, such as edges [57], shapelets [58], textures, HOG [59] and Haar wavelets [60] and then use SVM [61], boosting [58, 62], random forest [37, 63], clustering [37] or other algorithms to detect or classify objects for crowd counting. Most of these methods have achieved good performance when the objects are sparse, but the counting effect will decrease remarkably when facing dense crowds. Therefore, researchers began to explore effectively counting methods in more dense crowd scenarios. It has been observed that in most dense crowd scenarios, using local features can greatly improve the counting performance compared with global features. Many works [64–67] work based on local features. It is worth mentioning that recently Laradji et al. [68] and Liu et al. [69] continue working on detection-based methods. The former does not need to estimate the size and shape of the object but propose a novel loss function that encourages the network to output a single blob per object instance using only point-level annotations. The latter avoids the expensive labeling cost of bounding boxes and

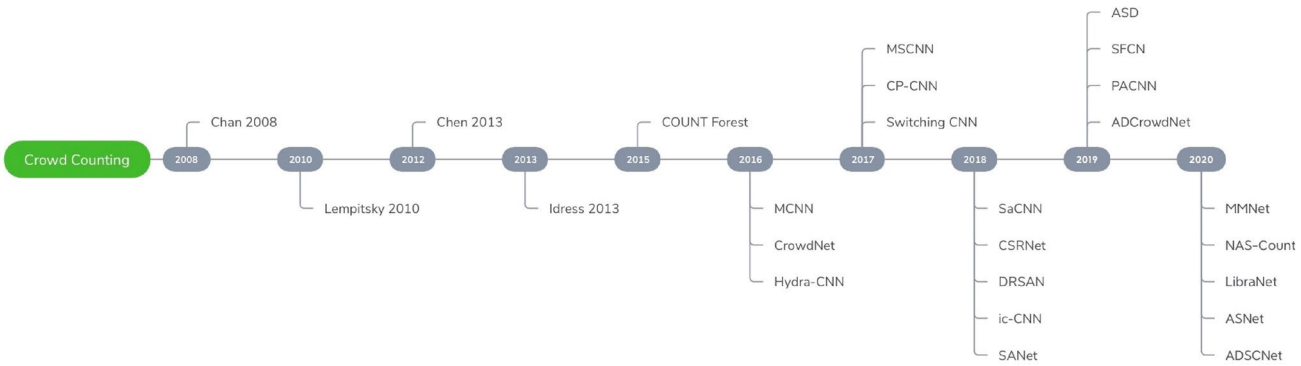

Fig. 1 Typical works in crowd counting in the form of a timeline. The methods listed from left to right are Chan [11], Lemptisky [36], Idress [13], Count Forest [37], MCNN [15], CrowdNet [38], Hydra-CNN [39], MSCNN [40], CP-CNN [41], Switching CNN [42], SaCNN [43] , CSRNet [44], DRSAN [45], ic-CNN [46], SANet [47], ASD [48], SFCN [17], PACNN [49], ADCrowdNet [50], MMNet [51], NAS-Count [52], LibraNet [53], ASNet [54], ADSCNet [55], AMRNet [56]

only uses the supervised information of the points to train the model.

When counting high-density crowd, detectors are hard to train due to more severe occlusions, etc. In this case, the performance of parts-based detection methods is also significantly decreased. The regression-based detection methods avoid the dependence on the detector and often have higher performance and capture growing attentions in crowd counting.

2.2 Regression-based methods

According to the different regression goals, the methods can be divided into individual-based regression [8, 12, 13, 70–72] and density-map-based regression. The former was proposed a little earlier. Compared with the parts-based detection methods, the individual-based methods further improved the counting performance. For examples, Ke et al. [12] first normalized the foreground of the image and then used the extracted local foreground, edge and texture features to learn multiple regression to get the number of individuals in the image. Compared with earlier regression models, this method enhanced the robustness of the model by learning a low-dimensional feature and a multi-structured output function, making it applicable to more practical scenarios. With further research, the concept of density map proposed by Lempitsky and Zisserman [36] has attracted widespread attention from researchers. It avoids the dependence on the detector by learning the mapping of images to density maps. Rodriguez et al. [73] confirmed that counting using density map can improve counting performance immensely. By reason of the density map not only reflects the spatial distribution information of the crowd but also increases the counting accuracy, density-map-based regression gradually becomes a popular category.

Similar to the detection-based methods, the methods of regression can be divided into integral-based [8, 70, 74] and patch-based [12, 75–77] categories. The integral-based regression methods invariably have difficulties in dealing with large scale and density changes while the patch-based regression methods contain more local information of the image and are less affected by changes in scale and density. Therefore, the performance of the patch-based regression methods is often better than the integral-based. Pham et al. [37] divided an image into multiple patches and used random forest to classify features, so that the leaf nodes of each tree contained only similar features. Meanwhile, the author also proposed crowdedness prior to make the prediction of the density of the next patch more accurate, which improved the model performance considerably.

Although the regression-based methods alleviate the dependency on the detector, they still rely heavily on handcrafted features. Consequently, feature extraction algorithm became a crucial bottleneck for the regression-based methods. With the rapid development of deep learning, the powerful feature extraction capabilities of convolutional neural networks (CNNs) have a tremendous fascination on the researchers, and the CNN-based crowd counting methods and crowd density estimation methods are developing rapidly.

2.3 CNN-based methods

In recent years, deep learning has increasingly attracted attention of researchers. CNNs have shown strong learning capabilities in image processing, inspiring plenty of CNN-based crowd counting works. Min et al. [78] is the first approach that applied CNN to crowd counting. However, it only estimated the crowd density level and not the specific number of the crowd. Since then, counting work based on CNN has progressed rapidly. The CNN-based methods

have better performance in scenarios such as the large span of human head scale, non-uniform density distributions and large changes in perspective and scene, which makes CNN-based approaches dominate the current crowd counting research. For the current crowd counting challenges, researchers have adopted different methods to deal with them. In order to enable researchers to comprehensively understand the current difficulties and their corresponding processing strategies, we divide the existing works into the following seven categories, each of which represents a mainstream counting strategy, as shown in Fig. 2. The main method and its motivation in each category will be discussed in detail in the following subsections.

Compared with the above works, we put more emphasis on the current difficulties and their corresponding processing strategies. Furthermore, we elaborate datasets and metrics of crowd counting and discuss the works of solving the problem of small-sample-based counting, dataset annotation methods and so on.

2.3.1 Multi-scale fusion

The multi-scale feature fusion methods attempt to solve the problem of large varying scales of human heads and the size of crowds by fusing multiple different levels of features, which is the main challenge of crowd counting.

1. MCNN: In order to extract features of different scales, Zhang et al. [15] proposed a multi-column CNN network model (As shown as MCNN in Fig. 3). The model consists of three columns of full convolutional networks. The only difference in each column is the number and size of the convolution kernels. During model training, each column of the network is trained independently and then fine-tuned after merging. The model finally uses 1X1 convolution kernels to fuse features of three different scales to obtain a density map. Because the architecture does not rely on any fully connected layer, the size of the input image can be arbitrary. It is worth

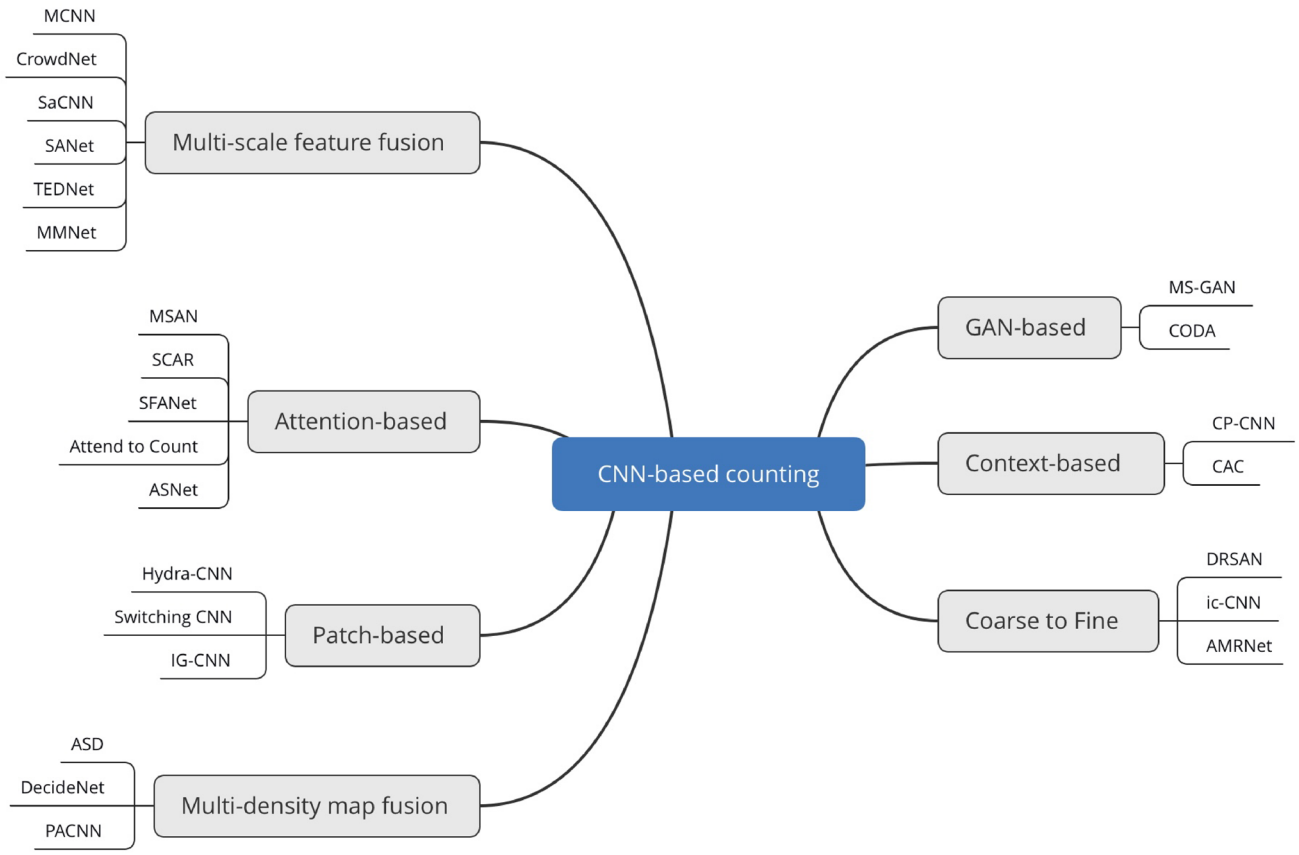

Fig. 2 Subcategories of CNN-based counting approaches, there are multi-scale fusion (MCNN [15], CrowdNet [38], SaCNN [43], SANet [47], TEDnet [79], MMNet [51]), attention-based (MSAN [80], SCAR [81], SFANet [82], Attend To Count [83], ASNet [54]), patch-based (Hydra-CNN [39], Switching CNN [42], IG-CNN [84]), multi-density map fusion(ASD [48], DecideNet [85], PACNN [49]), GAN-based (MS-GAN [86], CODA [87]), context-based (CP-CNN [41], CAC [88]), coarse-to-fine (DRSAN [45], ic-CNN [46], AMRNet [56])

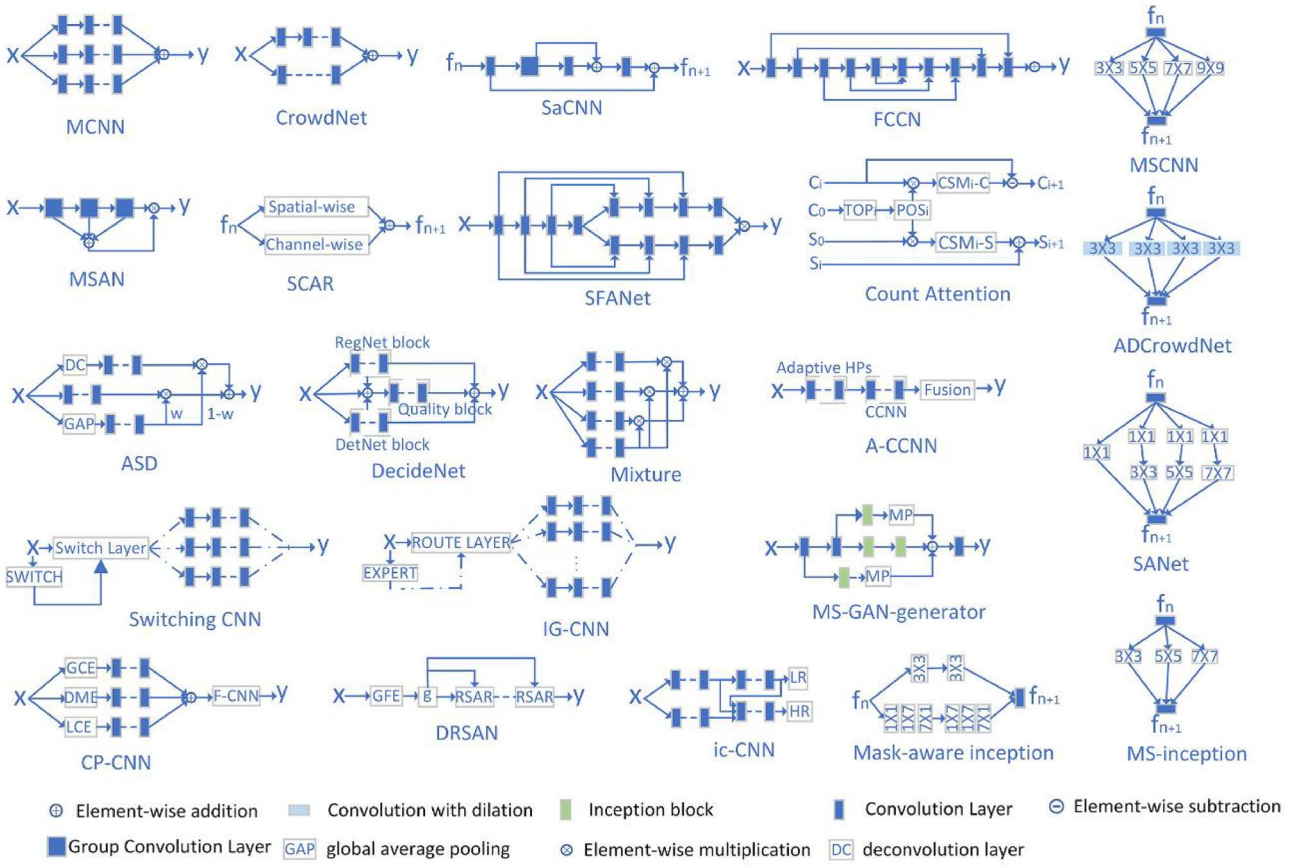

Fig. 3 A glimpse of diverse range of network architectures used for crowd counting using deep networks

mentioning that they also proposed a widely used dataset, the ShanghaiTech dataset, which contributed a large dataset and a benchmark to the crowd counting research community.

The later proposed CSRNET [44] proved through experiments that the features learned from three separate columns in MCNN were similar, which went against the original intention of learning different features for each column. In contrast, many inefficient branch structures cause the model to be too computationally expensive to perform real-time crowd counting. However, the intuition of multi-column network is very natural and instructive and many methods later followed or extended the idea of this design.

2. CrowdNet: Similar to MCNN, CrowdNet [38] also used a multi-column convolutional network to generate density maps and then predicted the corresponding crowd population. This method first performs multi-scale data enhancement on the input image and then uses a combination of deep and shallow fully convolutional networks to extract different levels of feature information, so that the network can simultaneously capture high-level features and low-level features to achieve multi-scale feature fusion. The CrowdNet is shown in Fig. 3.
3. SaCNN: To solve the problem of varying head scales, most methods use multi-column convolution to extract features of different levels for fusion. However, these methods always have high computing complexity, which limits the real-time performance of crowd counting. Zhang et al. [43] proposed a scale-adaptive convolutional neural network: SaCNN, as shown in Fig. 3. The network makes the model automatically adapt to the variation in scales and viewpoints in the input images by fusing multiple layers of features. Because the parameters and feature representations are shared between the layers of the network, the model has fewer parameters and is easy to train.

Sang et al. [89] further improved SaCNN network by optimizing the geometry-adaptive Gaussian kernel to obtain a high-quality ground truth density map. Combining absolute count loss and density map loss helps to improve the performance of the model in sparse crowd scenarios. It is noteworthy that the data are augmented using a random cropping method to improve the generalization ability of the model.

Zou et al. [90] also made some improvements to the network by using the deformation aggregation network (DA-Net) to generate fine-grained density maps. They used the network's geometric transformation capabilities to solve the multi-scale problem of the human head, etc. This is the first time that deformation aggregation network has been applied to the field of crowd counting, which effectively enhances the robustness of the model to processing varying head scales. Finally, they used adaptive learning weights to fuse the features of multiple branches in the network.

4. SANet: SANet [47] used a network of codecs, which they call a large-scale aggregation network. The encoder used modules similar to the Inception architecture proposed in GoogLeNet [91] to extract multi-scale header information. As shown in SANet in Fig. 3, the decoder uses a set of transposed convolutions to generate a high-resolution density map. In addition, the author combined Euclidean loss and local pattern consistency loss to train the model.
5. TEDnet: The codec architecture network proposed in TEDnet [79] obtains high-resolution density maps by fusing multi-level features. The architecture uses the entire image as input, continuously encodes and decodes feature maps at different stages and introduces skip connections for hierarchical fusion. The multi-path decoder in this approach aggregates features at multiple levels, fuses low-level detail maps with rich spatial information and high-level semantic maps with deeper semantic information and then restores the resolution of the density map by upsampling.
6. MMNet: This paper proposes an end-to-end scale-aware network (MMNet [51]). Compared with most existing scale perception works, the proposed MMNet not only captures the multi-scale features generated by filters of different sizes but also integrates the multi-scale features generated at different stages to deal with the scale of the human head variety.

In addition to the above works, there are many studies that try to solve the problem of varying scales. Chen et al. [92] used a multi-column convolutional network architecture and gradient fusion for crowd counting. Deb and Ventura [93] used a multi-column dilated convolutional network aggregation to fuse features at different levels. However, as mentioned earlier, some inherent disadvantages of multi-column networks still exist, such as the large amount of calculations and the difficulties in real-time counting. Therefore, some researchers started to study how to use a single network to fuse multi-scale features. Liu et al. [94] used laterally connected feature pyramid network to fuse high-level features with low-level features, as shown in FCNN in Fig. 3. Wang et al. [95] also proposed a single-column counting network, which is comprised of several special purpose modules, four residual fusion modules for multi-scale feature extraction, one pyramid pooling module for information fusion and one sub-pixel convolutional module for resolution restoring. The combination of these modules enables SCNet to effectively fuse multi-scale features in a compact single-column architecture. Dai et al. [96] used dense dilated convolutional blocks to extract information of continuously varying scales. Kang and Chan [97] used image pyramid method for multi-scale sampling. Gao et al. [98] constrained the density map by introducing fore/background segmentation. Some researchers also used modules similar to Inception to extract density map, such as Zeng et al. [40] introduced a multi-scale blob of different kernel sizes to extract features at different levels, as shown in MSCNN in Fig. 3. Existing counting frameworks widely use the static density map supervision method proposed in MCNN [15], but this method cannot tolerate labeling errors and failed to reflect changes in the crowd scale. In order to solve this difficulty, an adaptive dilated convolution and a novel supervised learning framework named self-correction (SC) supervision is proposed in ADSCNet [55].

The huge span of the human head scales in images has always been a major problem for crowd counting. Most current solutions are based on feature fusion on multiple scales. These methods mentioned in this section simply stack features together without using weight information. A lot of works recently introduced the attention mechanism to crowd counting, which includes the works of weighted fusion of features.

2.3.2 Attention-based

1. MSAN: Varior et al. [80] used a multi-branch scale-aware attention to solve the problem of large changes in the head scales in the image. The network guides branches at different levels to predict the corresponding density maps at multiple scales and finally uses a soft attention mechanism to fuse the previously predicted multi-scale density maps, as shown in MSAN in Fig. 3. Moreover, they also introduced a scale-aware loss function to guide the network training at different stages, which has a significant improvement on scenes with large-scale changes.
2. SCAR: Gao et al. [81] noticed that most existing crowd counting methods only focus on the local appearance features of the crowd but ignored a lot of contextual information and attention information. Therefore, the authors proposed a SCAR framework (Spatial-Channel-wise Attention Regression Network), which includes a SAM (Spatial-wise Attention Model) and a CAM (Channel-wise Attention Model), as shown in Fig. 3. SAM encodes the entire input image to obtain a wide range

of context information to predict the density map more accurately. CAM extracts the most discriminative features from the channel, making the network model more robust to noisy backgrounds. Finally, the information of the two attention networks is integrated to obtain a fused density map.

3. SFANet: Aiming to overcome the problems of varying head scales and strong background noise in the scenes, SFANet [82] proposed a dual path multi-scale fusion networks with attention for crowd counting. They used the VGG-16 network as the front end for feature extraction, and the dual path multi-scale fusion networks as the back end to generate density maps, as shown in SFANet in Fig. 3. One of them highlights the crowd area in the image to generate an attention density map. The other branch fuses the features of different levels extracted by the VGG-16 network and finally combines with the generated attention density map to generate a high-quality density map with high-resolution. Besides, this paper used a combination of the Euclidean loss and the attention map loss as the final loss function. Minimizing the former loss helps reduce pixel-level errors, while minimizing the latter loss can locate crowd areas more accurately.
4. Attend To Count: Zou et al. [83] proposed an adaptive capacity model in crowd counting. The model makes better use of multiple branches for prediction: Coarse network, Fine network and Smooth network. The Coarse network takes the original image as input and outputs a rough density map after passing through the multi-column network. The Fine network obtains a fine-tuned density map area through continuous fusion between layers. Finally, the Smooth network combines the two density maps to obtain the final density map. The authors proposed an attention mechanism called count attention, as shown in Fig. 3. It continuously uses the coarse density map generated by the Coarse network to locate the dense regions and then uses the Fine network to fine-tune the area.
5. ASNet: In order to overcome the problem of uneven crowd density in the image, X. Jiang et al. [54] proposed a two-branch method: One branch outputs intermediate density maps and scale factors, and the other branch provides a corresponding attention mask. The first branch multiplies intermediate density maps and scaling factors by attention masks to generate attention-based density maps, which are then summed to give the final density map.

In addition, Hossain et al. [99] enhanced the performance of the counting model by focusing on both global and local information. Similarly, Sindagi and Patel [100] effectively used spatial segmentation information and high-level channel information through the attention mechanism. Ranjan et al. [101] introduced encoder attention that performed well in NLP to fuse local and non-local information to obtain a density map. Sindagi and Patel [102] used segmentation information for the counting network through the inverse attention mechanism. Liu et al. [50] proposed a scheme in which two sub-networks are connected in series: The attention map generator at the front end generates an attention density map, and the density map estimator at the back end obtains the final density map. The author also combined inception and dilated convolution for multi-scale feature fusion, as shown by ADCrowdNet in Fig. 3.

The attention-based methods are inspired by the human brain cognitive mechanism and have been proved to be effective in many artificial intelligence fields. The attention mechanism in crowd counting can remarkably improve the counting performance of the models in complex scenarios such as varying scales, complicated intensities and changing perspectives. Further research is expected in this field in the future.

2.3.3 Patch-based

Patch-based counting methods divide the image into multiple patches, count them separately and fuse them at the final step. These approaches, to a certain extent, solved the problem of uneven crowd density of the input image.

1. Hydra-CNN: To address the problems of mutual occlusion and scene perspective in crowd counting, Onoro-Rubio and L*ó*pez-Sastre [39] proposed two deep learning network models: CCNN and Hydra-CNN. CCNN is an efficient fully convolutional network model, which is dedicated to the accurate regression of the patch density maps. Hydra-CNN proposes a solution by changing the scale: Resize patches of different sizes in the original image to a normalized size so that the model can perceive the changes in different scales.

   Kasmani et al. [103] further improved the model by reconsidering two parameters in the CCNN network: the size of the patch and the covariance of the Gaussian function. In the CCNN, these two parameters are the same for each patches of all scales, which leads to inaccurate regression. Therefore, the author proposed an adaptive CCNN model to remedy this issue. The author first used a head detector to detect the size and position of the human head in each patch and then fed them into a FIS (Fuzzy Inference Engine) to output the corresponding fuzzy information. Finally, the fuzzy information generated by each patch is sent to the corresponding network of the CCNN, which is used to direct the counting process and make the result more accurate. The network architecture of A- CCNN is shown in Fig. 3.

2. Switching CNN: Sam et al. [42] focused on solving the problem of uneven crowd density distribution in the input image. They used three branch networks similar to MCNN [15] for counting, as shown in Switch-CNN in Fig. 3. Unlike other approaches, it uses a classification network called Switch-CNN to determine which network branch each patch should be forwarded to. In this way, each patch can get the most appropriate processing and then can get more accurate predictions. By this means, it improves the problem of uneven density distribution in the image to some extent.

   In contrast, PaDNet [104] tried to solve this problem by weighting each branches of the network and achieved better performance in estimating of extremely high or low local density areas.
3. IG-CNN: Sam et al. [84] proposed an incremental network model IG-CNN to explain the diversity of the crowd in the scene. After pre-training, a CNN-based tree model is gradually established, where each node represents a fine-tuned regressior tained on one subdataset. After that, the above process is repeated recursively, and finally formed a CNN tree, with the leaf nodes of the tree are more specialized than the parent nodes. Each patch in the image will be sent to an appropriate leaf node, and a dedicated sub-network is used to get more accurate predictions.

   Zhang et al. [105] used the appearance of crowd as an auxiliary mechanism to filter out most of the background, so that the model pays more attention to the human heads. The author divided the crowd image into multiple patches and used the spatial position of each patch to deal with the problem of uneven distribution of crowd density. Han et al. [106] first divided the image into multiple overlapping patches and then used Markov Random Field to constrain the counting error.

2.3.4 Multi-density map fusion

Considering the problem of varying conditions of the input image, one of the solutions is to fuse density maps of multiple scales for crowd counting.

In order to overcome the difficulties of camera perspective changes and obstacle occlusion in crowd counting, ASD [48] proposed a network model that adaptively recalibrates path response by implicitly discovering and modeling dynamic scenes. As shown by ASD in Fig. 3, the framework first uses a convolutional neural network to extract features and then uses three branches to generate a density map. Two of them are similar parallel channels with different receptive fields, which generate density maps with different granularities. The third branch generates weight information for the density maps generated by the two branches and then weights the density maps to obtain the final density map. Experiments show that this branch greatly improved the generalization performance of the model.

Shen et al. [85] proposed a framework that combines detection and regression: DecideNet. As shown in Fig. 3, it includes three parts: the regression network RegNet, the detection network DetNet and the QualityNet. The former two parts generated two density maps, respectively, and the latter network weights and fuses the generated density maps to form a final map. This model is more robust to the varying sizes of the crowd and more suitable for a wider range of scenarios. Liu et al. [49] proposed a perspective-aware network for counting. The author took the perspective information as auxiliary information for the crowd scale changes and weighted the multi-level density maps and fused them. Shi et al. [107] took the entire image as input, used three branches to predict the number of individuals in the image and then used a branch to weight the previous prediction results and fuse them all, as shown in Fig. 3.

2.3.5 GAN-based

Generative Adversarial Networks (GAN) is a deep learning model proposed in Ref. [108]. In recent years, it has been one of the most promising methods for unsupervised learning of complex distributions. GAN includes two modules, namely a generative model and a discriminative model, which compete with each other to understand the distribution of actual data as much as possible. In related works on crowd counting, some researchers used a generator to obtain a density map and then used a discriminator to distinguish the density map from ground truth. This competition with each other ultimately makes the resulting density maps more accurate.

Similar to GAN, the model of MS-GAN proposed by Yang et al. [86] includes a generator and a discriminator. The generator is a multi-scale full convolutional network, which combines the features of different convolutional layers to generate a density map, as shown in MS-GAN-generator in Fig. 3. The density map generated by the generator is used as negative samples and trained in combination with ground truth by the discriminator. In this way, the performance of the generator in the adversarial network is improved iteratively, and a better density map is obtained. The two branches of the generator use the combination of the inception module and the max pooling layer proposed by Szegedy et al. [91] to fuse the features of different levels. The inception module is shown in MS-inception in Fig. 3.

Wang et al. [87] proposed an unsupervised adaptive learning method designed to enhance the performance of the model in a unseen scene. The author used multi-scale pyramid patches in the source and target domains for adversarial training to handle different crowd scales and density distributions. Shen et al. [85] designed a crowd counting

framework based on an adversarial generative network model: A U-shaped network was used as the model's generative network, and a high-resolution density map was screened out using a discriminator. Olmschenk et al. [109] was dedicated to studying how to train a crowd counting network model using only a small amount of data.

2.3.6 Context-based

Some crowd counting works use the contextual semantics of images to guide the counting procedure. This method mainly uses the context and semantic information of the crowd scene to constrain the density map to achieve better performance.

CP-CNN [41] proposed a context pyramid network to make full use of contextual information to generate high-precision density maps. The network consists of four parts, i.e., GCE (Global Context Estimator), LCE (Local Context Estimator), DME (Density Map Estimator) and F-CNN (Fusion-CNN), as shown in Fig. 3. The GCE encodes global information and extracts high-level semantic features, while classifying the entire input image into different density levels; the LCE encodes local information and extracts local features, while classifying each patch into different density levels; the DME is used to directly generate density map; finally, the outputs of the three parts are fused by the F-CNN to obtain a high-quality density map. Considering that using only Euclidean loss will cause the density map to be blurred, a weighted combination of pixel-level Euclidean loss and adversarial loss is used as the loss function.

Chong et al. [110] did not directly count the crowd number based on the entire image but calculated the final number of individuals by using the shared computations over overlapping areas. Liu et al. [88] combined features of multiple receptive field sizes and each image location and then trained them using an end-to-end trainable network. Finally, the network outputs a high-quality density map.

2.3.7 Coarse-to-fine

Most of coarse-to-fine works first obtain a coarse density map and then optimize or fine-tune it to obtain the final fine-grained density map.

In order to solve the problems of varying rotations, scales and perspectives caused by changing views of cameras, DRSAN [45] proposed a deep recurrent space-aware network. The network uses the Global Feature Embedding model based on VGG-16 as the front end to generate the primary density map and then uses the Recurrent Spatial-Aware Refinement model to optimize the generated density map. To be specific, the Recurrent Spatial-Aware Refinement model consists of two parts: a spatial transformation network, which is used to dynamically locate a region from the density map and then convert it to a suitable size by bilinear interpolation. Finally, residual density learning is used to optimize the density map of the selected area, and then a high-quality density map is thus obtained, as shown in DRSAN in Fig. 3.

Previous works such as L2SM [111] and S-DCNet [112] merge the feature maps of different convolution layers and obtain multi-scale information through the network structure of the feature pyramid. In contrast, [56] only implements multi-scale information enhancement on a single-layer feature map and repeats this operation on different convolution layers to bring rich information into the subsequent regression module.

ic-CNN [46] proposed a two-stage crowd counting model. As shown by IC-CNN in Fig. 3, the LR branch generates a low-resolution density map, and the HR branch incorporates the feature map and the low-resolution prediction to generate a high-resolution density map. The model can also be extended to a multi-stage model, that is, iterative fusion is used to improve the performance of the model, thereby obtaining a high-quality density map. Xu et al. [111] simulated human behavior when counting: First counted the sparse area, then zoomed in the dense area for a more accurate counting. The network consists of two parts: Scale preserving network (SPN) and Learning to scale module (L2SM). SPN uses multi-scale feature fusion to generate the initial density map. L2SM divides the image into multiple non-overlapping regions and selects some denser regions for re-prediction to improve the counting accuracy. This work is somewhat similar to Ref. [113], which uses a Recurrent Attentive Zooming Network to continuously detect blurry areas in an image then zoom in and recheck.

2.4 Video-based crowd counting

Crowd counting is mostly based on a still image. In addition, considering that video sequence contains timing information that is beneficial to counting, some researchers are currently working on counting the number of individuals in the video. We will briefly describe some of these typical tasks as follows.

1. ConvLSTM: Most video-based counting methods only consider single-frame of the video separately and ignore the temporal correlation between video frames, which is informative and beneficial for the counting context. ConvLSTM [114] effectively used the time correlation to assist the task of counting. The ConvLSTM model is an extension of FC-LSTM [115]. In the input-to-state and state-to-state, the fully connected structure is replaced with a convolution structure to perform feature transformation, and a 3D tensor fused with spatiotemporal information is used as an information representation for

information transmission and control gate. Unlike CNN-based methods that only consider spatial information, ConvLSTM model pours more attention to the temporal correlation between adjacent frames of the video and thus can effectively use the time domain information. This method more adequately captures the relationship between space and time in the video, thereby improving the accuracy of counting in complex scenes.

2. LSTN: Unlike LSTM-based implicit modeling method, LSTN [116] used a Locality-constrained Spatial Transformer module to explicitly capture the spatiotemporal dependencies in the video. The model is mainly composed of two modules: the density map regression module and the position constraint-based spatial transformer (LST) module. The density map regression module directly estimates the density map of a single frame and then uses the LST module to associate the density map of adjacent frames to output a more accurate density map.

   Similar to LSTN, Wu et al. [117] used crowd images and predicted density maps to explicitly model time information. They used a set of dilated residual blocks to model the relationship between features of adjacent frames of the video. At each stage, an expanded set of convolutions over time is used to generate the initial density map that is used to optimize the subsequent density map iteratively in the next stage.
3. E3D: Considering the superior performance of 3D convolution in motion recognition, Zou et al. [118] attempted to use of 3D convolution to encode the spatiotemporal features in the video. It encodes global context information into modulation weights while rescaling the characteristic response of each channel, adaptively highlighting useful features and using short skip connections to simplify model training.

   The novel architecture temporal channel-aware (TCA) constructed in this paper can not only capture the time dependence of video sequences effectively but also fuse local and global spatiotemporal information. The author stacked multiple TCAs together to obtain a deeper enhanced network which achieved better performance.
4. Cross-Line Pedestrian Counting: Zheng et al. [119] proposed a scalable crowd counting method, designed to count pedestrians crossing virtual lines when the crowd is highly dynamic and dense. The method includes two parts: local crowd density estimation and cross-line pedestrian counting. In order to accurately estimate the local crowd density, they divided the neighborhood on the virtual line into several blocks and enhanced the spatial consistency between the local count and the closed area count to ensure the consistency of the local crowd density estimation. To address the problem of uneven density, they proposed a two-stage solution: First, divide the sample into multiple density levels and then train an expert regressor with overlapping operating ranges for each density level to offset the error caused by the first density classification.
5. Dynamic region division: Considering that the straight-line double region pedestrian counting [120] method may divide a head into two parts, thereby introducing counting errors, He et al. [121] proposed a dynamic region partitioning algorithm to ensure the integrity of the counting object. On the premise of maintaining the integrity of the head, they used the bounding box and scene segmentation lines of the objects obtained by YOLOV3 [122] to segment the distal and proximal regions. For the near-end area, YOLOV3 [122] is used for direct pedestrian detection; for the far-end area, the receptive field is expanded by introducing dilated convolution, and an Inception module is designed to automatically select the dilation rate. Finally, the two results are fused to obtain global distribution information.

2.5 Extensions and related issues

In the field of crowd counting and density estimation, besides the aforementioned seven categories and video crowd counting, there are many other works worth mentioning.

In order to locate a person's head when counting crowd, Lian et al. [123] proposed an RGBD-based network architecture called RDNet. The architecture is improved from Ref. [124] and consists of a regression model and a detection model. The density map generated by the regression model is used to enhance the robustness of the detection model when encountering small heads. They introduced a depth adaptive kernel that takes the changes of head size into account, which makes the regression of the density map more robust. In addition, a depth-aware anchor is also designed to help anchor initialization and improve the model's detection performance for small targets. Similar to RDNet, Sam et al. [125] aimed to locate each individual in the image while counting. They made full use of density maps to enhance the robustness of face detection models to small targets. The effect of their detection model named LSC-CNN is illustrated in Fig. 4.

CSRNet [44] (Congested Scene Recognition Network) is a very popular job in the field of crowd counting. Its advantages include good performance, simple network architecture, easy training and outstanding generalization performance. The network is mainly composed of two parts: The first 13 layers of VGG16 are used as the front end for feature extraction, and the back end uses dilated convolutional layers to deepen the network while expanding the receptive field. In addition, they confirmed through experiments that the multi-column network designed in MCNN [15] failed to achieve the expected effect but instead introduced extra

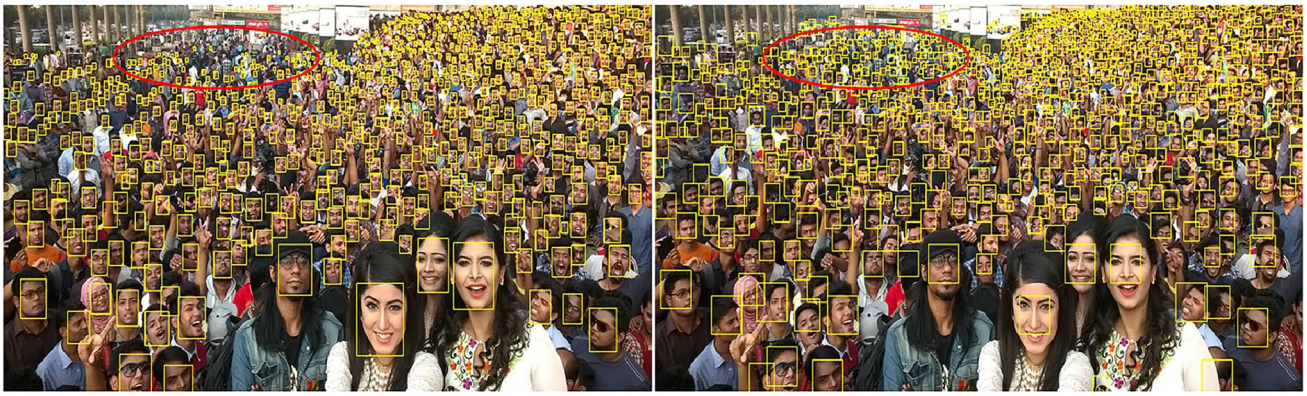

Fig. 4 The left picture shows the detection effect of Tiny Face detector [126]. Only 731 individuals were detected among 1151 people, and the main errors were concentrated in the crowded areas (such as in the red ellipses). In contrast, the right side shows the detection effect of LSC-CNN [125], and 999 people were detected. Reproduced with permission of Ref. [125], Copyright of 2020 IEEE

calculations, resulting in reduced real-time performance of counting. This raises debates about the advantages and disadvantages of single-column and multi-column network design, which we discussed in Sect. 4.2.

Aiming at the problem of high background noise in crowd counting, researchers use a branch to divide the background and foreground of the image, thereby improving the accuracy of the density map. More typical works in this area are Ref. [127] and [98]. The former uses a network of inception structure, as shown in MA-inception in Fig. 3.

For the perspective problem caused by camera angle changes in crowd counting, Kang et al. [128] used the camera's angle, height and vertical field of view as extra information to assist the counting. Marsden et al. [129] fed multiple scales of the input image into the network model and then used the average of the number of individuals as the predicted value. W-net [130] introduced independent decoding enhancement branches to U-net [131] to speed up model convergence. Ding et al. [132] used ResNet-based deep recursive network for crowd counting. The recursive structure makes this model capable of capturing statistical patterns in a crowd environment without increasing the number of parameters. Zhao et al. [133] used spatial, semantic and numeric attributes to assist in training the model. Wan et al. [134] counted the crowd using relevant semantic information between samples. Huang et al. [135] counted from the perspective of semantic modeling, making full use of composite body-part semantic structure information. Y. Yang et al. [136] proposed a reverse perspective network to deal with the scale change of the input image. The network can explicitly evaluate the perspective distortion and effectively correct the distortion by warping the input image uniformly.

In addition, hwan Oh et al. [137] proposed a scalable neural network framework that uses the ensemble strategy of bootstrap to quantify the uncertainty after decomposition. This is the first work to quantify the uncertainty of crowd counting. Trying to avoid over-fitting problems, Shi et al. [138] took the combined dataset as the model's input and utilized multitask training to learn a generalizable representation across similar domains. Wei et al. [139] were dedicated to counting fast moving crowds. Chen et al. [72] used the method of cumulative attribute space learning to deal with the sparseness and imbalance of data. Shi et al. [140] used Deep Negative Correlation Learning to generate a generalizable feature learning strategy that transforms regression counting problems into integration problems. Chan et al. [11] were committed to counting individuals while protecting privacy. Oncel [141] used a covariance matrix as a target descriptor to detect pedestrians in still images. Arteta et al. [142] used foreground and background segmentation and local uncertainty estimation to enhance density map estimation. The author used penguin counting as an example to design a deep multitasking structure to effectively utilize the mutual assistance between tasks. Experiments showed that the multitask density estimation method greatly improved accuracy compared with the single-task density estimation method. The deep residual structure ResnetCrowd proposed by ResnetCrowd [143] can be used for crowd counting, violent behavior detection and density classes classification. To train and evaluate the proposed multi-objective technology, they created a new dataset, called Multi Task Crowd, which has large scene differences, reasonable distribution of violent and non-violent images, and significant differences in crowd size. Most of the existing crowd counting works manually design the network to learn the density estimation map. Y. Hu et al. [52] used Neural Architecture Search (NAS) to search the encoder–decoder structure in nine extraction and fusion cells, which uses dilated convolution to capture multi-scale information. Yang et al. [136] proposed a lightweight and effective structured knowledge transfer method which transfers the structured knowledge of the crowd counting model to a lightweight model. Meanwhile, it increases the

efficiency by several times while maintaining the accuracy of the original model. Reinforcement learning is used to transform the crowd counting problem into a sequential decision-making problem. Different from the existing counting models that directly output the number of individuals in one step, Liu et al. [53] divided the one-step estimation into a series of sub-decision problems. Liu et al. [144] proposed a semi-supervised learning method that uses unlabeled data to train a general feature extractor. Considering that the location information of people in the image is costly, Yang et al. [145] proposed a weakly supervised network that does not require location supervision, which mainly uses the number of individuals of the image for counting.

3 Datasets and metrics

Datasets are of great significance for training and evaluating crowd counting models. To train more generalized models, researchers have built a wide variety of datasets. Early datasets are mostly images or video frames with low crowd density and similar scenarios. Most of the later datasets often have large sample size and more accurate labels. In terms of sample size, these datasets mostly cover a wide range of factors: diverse scenarios, varying crowd densities, large head scale span, etc., which are much closer to the data distribution in real applications. From the perspective of labeling accuracy, the density maps generated by these datasets are more accurate and reasonable. We will elaborate and analyze these datasets and perform a performance comparison of popular crowd counting methods on the datasets. We also summarized some works on target domains with less labeled data. Finally, we explained and evaluated the metrics commonly used in crowd counting.

3.1 Datasets

The most popular datasets available in crowd counting are UCSD [11], Mall [12], UCF_CC_50 [13], WorldExpo 10 [14], ShanghaiTech [15], UCF-QNRF [16] and the newly proposed virtual synthetic dataset GCC [17]. We will give a brief introduction of these datasets and evaluate the performance of popular models based on these benchmarks.

A. UCSD

The UCSD dataset [11] consists of 2000 frames in a video sequence, with every five frames corresponding to one ground truth. It was acquired by a fixed camera mounted above a sidewalk, so the scenes relatively lack of variation. In addition, the density of crowd on the sidewalks varies from sparse to crowded. This dataset is the first dataset created in crowd counting. Since the dataset was released earlier, there are many limitations with the dataset, such as images were collected from a single fixed position and the scenes were inevitably single. The data distribution does not match with many real scenes, which makes it unsuitable for more general applications.

B. Mall

The Mall dataset [12] consists of 2000 320 $\times$ 230 video frames with 6000 labeled pedestrians. The labeled individuals were provided by marking the pedestrian head of all frames. Compared with the USCD dataset, the Mall dataset has a higher crowd density and more diverse scenes.

C. UCF_CC_50

The UCF_CC_50 dataset [13] is composed of 50 different resolution images. Each image contains an average of 1280 people. The entire dataset includes 63,075 people totally. The number of individuals in each image is between 94 and 4543, and some images contain very dense crowds. This dataset also contains much more diverse scenes, such as concert hall, protest rally and gymnasium. Considering the dataset is relatively small for large capacity models, Idrees et al. [13] defined a cross-validation protocol for training and validating models.

D. WorldExpo 10

The WorldExpo 10 dataset [14] consists of 3980 576 $\times$ 720 video frames, with a total of 199,923 labeled pedestrians. Its training set is derived from 1127 one-minute video sequences in 103 scenes, and its test set is derived from five one-hour video sequences in five different scenes. Each test scene contains 120 frames of images, and the number of individuals in each frame is between 1 and 220.

E. ShanghaiTech

The ShanghaiTech dataset [15] contains a total of 1198 labeled images and 330,165 labeled heads, which are divided into A and B parts. Part A contains 482 images, of which the training set and the test set have 300 and 182 images, respectively. These images are collected from the Internet. Part B contains 400 training images and 316 test images which are taken in Shanghai's urban streets. Compared with Part A, the crowd density in Part B is relatively less. This dataset covers multiple scenarios and different density levels. The ShanghaiTech dataset is a very challenging dataset and most of recently crowd counting works are based on this dataset for comparison.

F. GCC

The images in GCC were taken from Grand Theft Auto V (GTA5). Wang et al. [17] designed a data collector and labeler to capture stable images and their head labels in the game. The dataset covers 400 types of scenes, and the individuals in the scenes have different skin colors, genders, appearances, etc. The author also used a step-by-step approach to break the limit of the maximum number of individuals in the image. This is the largest dataset in crowd counting, both in terms of sample size and the scenarios covered. Using it to pre-train the model and then fine-tune

the model with actual data usually can get better counting performance.

G. NWPU-Crowd

Considering that the CNN-based method requires a huge dataset to support, but the existing datasets are too small, so Wang et al. [146] proposed the NWPU-Crowd. It is currently the largest dataset in crowd counting, with 5109 images and 2,133,238 labeled entities. Furthermore, the dataset also contains some negative samples, which help to enhance the robustness of the model. In addition, it contains various illumination scenes and has the largest density range [0, 20033].

3.2 Annotation Methods of Datasets

There is no doubt that how to give accurate ground truth of crowd counting annotation is of critical importance. Different datasets are quite different in density map generation. The main density map generation methods are as follows.

A. Point-wise convolution of human head

Lempitsky and Zisserman [36] first introduced the concept of density map in crowd counting, which led this field into a new stage and had a great influence on the subsequent works. To avoid the difficulties of detecting and locating objects, the counting problem is regarded as a mapping problem between the input and the density map. Finally, the density map is used by a regress method to obtain the total number of the individuals. To be specific, they labeled the human head in the image as a point, and then performed 2D Gaussian convolution on each point to obtain the corresponding ground truth.

B. MCNN

Due to the large variation in scales of head in most images, MCNN [15] avoided directly using the static two-dimensional Gaussian convolution methods as in Ref. [36], they introduced a geometric adaptive convolution kernel to promote the accuracy of the density map. The size of the Gaussian convolution kernel is determined by the k heads near the central position. Therefore, the generated density map is closer to the actual distribution of the crowd. Moreover, in order to avoid the irrationality of the density map due to the sparseness of the human head, the size of the Gaussian convolution kernel is generally limited to 100 pixels.

C. Content-aware density map

Oghaz et al. [147] divided the previous density map generation methods into two categories: static two-dimensional Gaussian method, such as Ref. [36], and dynamic two-dimensional Gaussian method [15]. The static method does not consider the size changes of the human head, which prevent the density map to be more accurate. Dynamic two-dimensional Gaussian method try to complement this shortcoming in Ref. [36]. However, it does not take into account the content information of the crowd in the image, which may introduce a lot of noise and has a negative impact on the counting accuracy. The author combined the Chan-Vese segmentation strategy, two-dimensional Gaussian convolution kernel and brute-force nearest-point search to improve the performance, and used content-aware technology to make up for the lack of accuracy of previous methods. Experiments show that models using the density maps generated by this method can achieve much higher accuracy.

D. IKNN map

Olmschenk et al. [148] noticed that the previous density map generation methods still had two aspects that can be improved. Firstly, consider an extreme case, each pedestrian is completely residing on a single pixel in the density map, the network predicting density 1 pixel away from the correct labeling is considered just as incorrect as 10 pixels away from the correct labeling. Obviously, it is not desired because a discontinuous training gradient is generated. Another aspect is that some localities have very large Gaussian distributions, which also leads to inaccurate spatial information of density locations. Therefore, the author proposed IKNN (Inverse K-Nearest Neighbor) map, which provides a substantial spatial gradient. They experimentally demonstrated that using the IKNN map to train existing crowd counting models can improve the performance notably.

E. Point supervised Bayesian estimation

Existing density map generation methods mainly use pixel-wise supervised method, which uses Gaussian kernel convolution to transform labeled points into the according density map. Ma et al. [149] proposed a Bayesian loss function based on point supervision, which constructs a density contribution probability model for supervised training from the point annotations perspective.

3.3 Evaluation metrics

We briefly analyze and evaluate the existing evaluation criteria of the crowd counting model from the two aspects of objects quantity and the density map. Following are the most widely used evaluation criteria.

3.3.1 For evaluating objects quantity

MAE (mean absolute error) is a commonly used evaluation criteria in regression models and represents the sum of the absolute values of the differences between the predicted values and ground truth, which is defined as follows:

$$\mathrm{MAE} = \frac{1}{m}\sum_{i=1}^{m}|y_i - \hat{y}_i| \tag{1}$$

where m represents the number of objects, y_i is the ground truth and $\hat{y}_i$ is the predicted value.

Mean square error (MSE) is another commonly used regression evaluation criteria, which represents the sum of

the squares of the distances between predicted values and ground truth, which is defined as follows:

$$\mathrm{MSE} = \frac{1}{m}\sum_{i=1}^{m}(y_i - \hat{y}_i)^2 \tag{2}$$

where the variables have same meaning as Eq. 1. Compared with MAE, MSE gives greater weight to the outliers, which better reflects the robustness of the model.

3.3.2 For evaluating the Density map

Peak signal-to-noise ratio (PSNR) is a commonly used measure of the similarity of two images. It is based on the error between corresponding pixels. A larger value indicates less image distortion.

$$\mathrm{PSNR} = 10\lg\frac{(2^n-1)^2}{\mathrm{MSE}} \tag{3}$$

The calculation formula is shown in Eq. 3, where MSE is shown in Eq. 2. It is an objective evaluation standard that does not take into account the visual characteristics of the human eye. Therefore, in some cases, the evaluation results are inconsistent with human subjective perception.

SSIM [150] (structural similarity) is also a measure of the similarity of two images. As an implementation of structural similarity theory, SSIM models the distortion of an image as the product of three different factors: brightness, contrast and structure. The mean is used as an estimate of brightness, the standard deviation is used as an estimate of contrast and covariance is used as a measure of structural similarity. The value of SSIM is in the range of [0, 1], and the larger the value, the smaller the image distortion. SSIM can make up for the defect that MSE cannot measure the similarity of image structure. The calculation formula is shown as follows:

$$\mathrm{SSIM}(x,y) = \frac{(2\mu_x\mu_y + c_1)(2\delta_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\delta_x^2 + \delta_y^2 + c_2)} \tag{4}$$

where μ_x is the average of x, μ_y is the average of y, δ_x^2 is the variance of x, δ_y^2 is the variance of y, δ_{xy} is the covariance of x and y. $c_1 = (k_1L)^2$,$c_2 = (k_2L)^2$ are constants used to maintain stability, L is the dynamic range of pixel values, $k_1 = 0.01$, $k_2 = 0.03$.

3.4 Performance comparison

We compare the performance of off-the-shelf crowd counting models based on several popular benchmarks, as shown in table 1 and 2. Specifically, the datasets include ShanghaiTech [15], UCF_CC_50 [13], WorldExpo'10 [14], and the evaluation metrics are MAE and MSE.

We made a brief summary of the table: The best performing work in ShanghaiTech A is PGCNet [157], the best performing work in ShanghaiTech B is S-DCNet [112], the best performing work in UCF_CC_50 is PaDNet [104] and the best performer in WorldExpo'10 is DSSINet [156].

3.5 Crowd counting with small sample data

Except for GCC [17], the above datasets need to spend a lot of human and financial resources for labeling. Undoubtedly, the quality of the labels has a major influence on the performance of the model. Moreover, as the number of network parameters increases, the model's demand for large datasets becomes more urgent. As a result, researchers have begun to explore ways to avoid annotating too many data manually.

A. L2R and SL2R

L2R [18] used the learning-to-rank framework to sort unlabeled images and facilitate the training of the counting model. They cropped the image into multiple small patches and then ranked them according to the observation that the number of individuals in the sub-image is less than or equal to the number of individuals in the original image. Besides, they proposed two network models: a density estimation network with multi-scale input patches and a network that ranks unlabeled data. They verified three following training methods using the labeled data and sorted data: training with the sorted data and then fine-tuning with the labeled data; alternate training with two types of data; multitask training. The results hint that the performance of the three training models is better than the model trained only by the labeled data. Among them, the performance of the multitask training model is usually the best.

SL2R [21] is a further work based on L2R. The author took ranking as a proxy task and proposed a backpropagation technique for Siamese networks, thereby avoiding redundant calculations caused by multi-branch network structures.

B. SFCN

Wang et al. [17] proposed a virtual synthetic dataset called GCC. The images in the dataset are derived from GTA5, with higher resolution and more diverse scenes. Due to the setting of GTA5, an image contains up to 256 people, so they used integration between images to break through the limitation of the number of individuals. The domain adaptive method proposed in the paper can effectively learn the domain invariant features between synthetic data and real data. More specifically, GAN-related technologies are used to make synthetic images more realistic, and

Table 1 Comparison of crowd counting methods on shanghaiTech and UCF_CC_50 datasets

Year-journal/conference	Methods	ShanghaiTechPart_A		Part_B		UCF_CC_50	
		MAE	MSE	MAE	MSE	MAE	MSE
2016-CVPR	MCNN [15]	110.2	173.2	26.4	41.3	377.6	509.1
2017-AVSS	CMTL [151]	101.3	152.4	20	31.1	322.8	397.9
2017-CVPR	Switching CNN [42]	90.4	135	21.6	33.4	318.1	439.2
2017-ICIP	MSCNN [40]	83.8	127.4	17.7	30.2	363.7	468.4
2017-ICCV	CP-CNN [41]	73.6	106.4	20.1	30.1	–	–
2018-TIP	BSAD [135]	–	–	20.2	35.6	409.5	563.7
2018-AAAI	TDF-CNN [152]	97.5	145.1	20.7	32.8	354.7	491.4
2018-WACV	SaCNN [43]	86.8	139.2	16.2	25.8	314.9	424.8
2018-CVPR	ACSCP [85]	75.7	102.7	17.4	27.4	291	404.6
2018-CVPR	D-ConvNet-v1 [153]	73.5	112.3	18.7	26	–	–
2018-CVPR	IG-CNN [84]	72.5	118.2	13.6	21.1	291.4	349.4
2018-CVPR	L2R [18] (Query-by-example)	72	106.6	14.4	23.8	291.5	397.6
2018-CVPR	L2R [18] (Keyword)	73.6	112	13.7	21.4	279.6	388.9
2018-CVPR	DecideNet [85]	–	–	21.53	31.98	–	–
2018-IJCAI	DRSAN [45]	69.3	96.4	11.1	18.2	219.2	250.2
2018-ECCV	ic-CNN [46] (one stage)	69.8	117.3	10.4	16.7	–	–
2018-ECCV	ic-CNN [46] (two stages)	68.5	116.2	10.7	16	260.9	365.5
2018-CVPR	CSRNet [44]	68.2	115	10.6	16	–	–
2018-ECCV	SANet [47]	67	104.5	8.4	13.6	258.4	334.9
2019-AAAI	GWTA-CCNN [19]	154.7	229.4	–	- -	–	–
2019-ICASSP	ASD [48]	65.6	98	8.5	13.7	196.2	270.9
2019-ICCV	CFF [154]	65.2	109.4	7.2	12.2	–	–
2019-CVPR	SFCN [17]	64.8	107.5	7.6	13	214.2	318.2
2019-ICCV	SPN+L2SM [111]	64.2	98.4	7.2	11.1	188.4	315.3
2019-CVPR	ADCrowdNet [50] (AMG-bAttn-DME)	63.2	98.9	7.7	12.9	273.6	362.0
2019-CVPR	ADCrowdNet [50] (AMG-DME)	66.1	102.1	7.6	13.9	257.9	357.7
2019-CVPR	PACNN [49]	66.3	106.4	8.9	13.5	267.9	357.8
2019-CVPR	TEDnet [79]	64.2	104.5	8.2	12.8	249.4	354.5
2019-CVPR	PACNN [49]+CSRNet [72]	62.4	102	7.6	11.8	241.7	320.7
2019-CVPR	CAN [88]	62.3	100	7.8	12.2	212.2	243.7
2019-TIP	HA-CCN [100]	62.9	94.9	8.1	13.4	–	–
2019-ICCV	BL [149]	62.8	101.8	7.7	12.7	229.3	308.2
2019-WACV	SPN [155]	61.7	99.5	9.4	14.4	–	–
2019-ICCV	DSSINet [156]	60.63	96.04	6.85	10.34	216.9	302.4
2019-TIP	PaDNet [104]	59.2	98.1	8.1	12.2	185.8	278.3
2019-ICCV	S-DCNet [112]	58.3	95	6.7	10.7	204.2	301.3
2019-ICCV	PGCNet [157]	57.0	86.0	8.4	13.6	244.6	361.2
2020-CVPR	ADSCNet [55]	55.4	97.7	6.4	11.3	–	–
2020-CVPR	ASNet [54]	57.78	90.13	–	–	174.84	251.63
2020-ECCV	AMRNet [56]	61.59	98.36	7.02	11.00	184.0	265.8
2020-ECCV	LibraNet [53]	55.9	97.1	7.3	11.3	181.2	262.2
2020-ECCV	AMSNet [52]	56.7	93.4	6.7	10.2	208.4	297.4

the resulting images can be directly used for model training. Besides, they experimentally confirmed that training existing models on this dataset can indeed improve model counting performance. Compared with traditional datasets, GCC has richer scenes, more accurate labels, larger sample capacity and higher resolution. For model training, two strategies can

Table 2 Comparison of crowd counting methods on WorldExpo'10 datasets

Year-journal/conference	Method	WorldExpo'10					
		S1	S2	S3	S4	S5	Avg.
2015-CVPR	Zhang 2015 [14]	9.8	14.1	14.3	22.2	3.7	12.9
2016-CVPR	MCNN [15]	3.4	20.6	12.9	13.0	8.1	11.6
2017-ICCV	ConvLSTM-nt [114]	8.6	16.9	14.6	15.4	4	11.9
2017-CVPR	Switching CNN [42]	4.4	15.7	10	11	5.9	9.4
2017-ICCV	ConvLSTM [114]	7.1	15.2	15.2	13.9	3.5	10.9
2017-ICCV	CP-CNN [41]	2.9	14.7	10.5	10.4	5.8	8.86
2017-ICCV	Bidirectional ConvLSTM [114]	6.8	14.5	14.9	13.5	3.1	10.6
2018-CVPR	DecideNet [85]	2	13.14	8.9	17.4	4.75	9.23
2018-CVPR	CSRNet [44]	2.9	11.5	8.6	16.6	3.4	8.6
2018-CVPR	ACSCP [85]	2.8	14.05	9.6	8.1	2.9	7.5
2018-ECCV	SANet [47]	2.6	13.2	9	13.3	3	8.2
2018-TIP	BSAD [135]	4.1	21.7	11.9	11	3.5	10.5
2018-IJCAI	DRSAN [45]	2.6	11.8	10.3	10.4	3.7	7.76
2019-ICCV	ADCrowdNet [50] (AMG-bAttn-DME)	1.7	14.4	11.5	7.9	3	7.7
2019-CVPR	ADCrowdNet [50] (AMG-attn-DME)	1.6	13.2	8.7	10.6	2.6	7.3
2019-CVPR	TEDnet [79]	2.3	10.1	11.3	13.8	2.6	8
2019-CVPR	CAN [88]	2.9	12	10	7.9	4.3	7.4
2019-ICCV	DSSINet [156]	1.57	9.51	9.46	10.35	2.49	6.67
2020-CVPR	ASNet [54]	2.22	10.11	8.89	7.14	4.84	6.64
2020-ECCV	AMSNet [52]	1.6	8.8	10.8	10.4	2.5	6.8

be adopted, one is to first train with GCC and then fine-tune the model with the real dataset, the other is to use only GCC to train the model. Both strategies can achieve much better performance.

C. CAC

Most current counting works are dedicated to counting only one specific kind of objects. However, CAC [20] focuses on using a model to count objects of any category. The author used the self-similarity property of images to convert the counting problem into a matching problem. The so-called self-similarity property means that the image can be represented by some specific repeated blocks, and these blocks are called exemplar. To this end, they proposed a matching network that can be used for unknown classes: Generic Matching Networks, thereby converting the counting problem into a matching problem. Moreover, an adapter module is designed to meet the needs of different users, so that it only needs a small number of labeled data to train a high-performance model, which is of great significance for scenarios lacking training data.

D. GWTA-CCNN

Sam et al. [19] proposed a nearly unsupervised method based on the Grid Winner-Take-All (GWTA) autoencoder. This method first encodes the input image, then decodes and reconstructs it, and uses the similarity between the images as a loss function to train the encoder and decoder. Almost 100% of the parameters in this model are obtained through unsupervised training. Comparative experiments showed that the performance of the GWTA method using only a small amount of data is better than the supervised counterpart.

4 Application and discussion

4.1 Application

At present, crowd counting is mainly used in scenarios such as crowd security, video surveillance and traffic analysis. Monitoring the number of individuals in assembly activities is an important part of crowd security, such as sports events, public demonstrations, political gatherings, concerts and other scenarios. Information about number of individuals can be used not only to assist the security forces but also to evacuate people in a more timely and effective manner, thereby reducing the possibility of accidents. Secondly, crowd counting can also be used to monitor traffic flow information, which not only helps road construction but also makes vehicle scheduling plans more reasonable. Moreover, statistics on the number of individuals staying in front of different shelves in the shopping mall can help to assess the popularity of different products, which is conducive to a

more reasonable layout of goods by merchants. Moreover, it is also a good research direction to expand crowd counting to the field of microscopy counting.

4.2 Discussion

4.2.1 Challenges and solutions

1. Scale variation

   The scale variation is a tough challenging in crowd counting and density estimation, including changes in the scale of the crowd and the size of the head. As shown in a and b in Fig. 5, the number of individuals in different scenes varies greatly, when the crowd is dense, the number of individuals can reach thousands, and when sparse, there are only dozens of individuals. Such a huge quantity difference is a daunting challenge for the model. In addition, as shown in c in Fig. 5, due to the camera angle, the size of the human head in the image is inevitably very different.

   As described in Sect. 2.3.1, there are currently two main solutions: fusion of multi-scale features and fusion of multi-scale density maps. By fusing features or density maps of different levels, the scale variation can be alleviated to some extent. Some researchers have done a lot of works based on the idea of multi-scale feature fusion: Multi-column networks with different receptive fields are used for multi-scale feature extraction, such as MCNN [15] and CrowdNet [38]; fusion of feature maps generated in different stages, such as SaCNN [43] and TEDnet [79]; use inception to directly merge multi-scale features, such as ADCrowdNet [50] and SCNet [95]; multi-scale density maps fusion, such as Ref. [48] and [85].
2. Occlusion

   Almost all images contain occlusion problem, and it becomes more severe as the crowd becomes denser. As shown in d in Fig. 5, when the crowd is dense, the occlusion is becoming grievous which makes counting extremely difficult. Most current works utilize the powerful feature extraction ability and learning ability of convolutional neural network to ease this difficulty.
3. Uneven distribution

   In most cases, the individuals are not evenly distributed, so the crowd density distribution varies greatly, as shown in e in Fig. 5. Researchers have proposed two solutions: using attention mechanism and patch-based processing. The attention mechanism makes the model pay more attention to the crowded area and reduces the counting error in the corresponding part, thereby improving the performance. The MSAN [80] and Attend To Count [83] models are in this category. Other methods such as Hydra-CNN [39] and Switch-CNN [42] divide the input image into multiple patches and then process different patches, respectively, thereby alleviating the uneven distribution problem.
4. Perspectives variation

   Changes in camera position and angle of view directly lead to scale variation in the image, occlusion and uneven distribution.
5. Others

   In addition to the main difficulties mentioned above, crowd counting also faces many other difficulties, such as small datasets, high background noise and large differences in light levels.

4.2.2 Trends

1. Design of architectures

   Since the architecture of MCNN [15] was first put forward, a lot of works follows the idea of using multi-column networks. However, CSRNet [44] pointed out that the design of the multi-column network did not achieve the expected accuracy but increased extra calculation. In other words, the multi-column network design can somewhat improve the counting performance, but the calculation amount introduced is too large. Therefore, multi-column networks are not recommended for real-time applications, however, the idea of multi-scale feature fusion is worth further exploration. Multi-scale feature fusion in a single-column network can be achieved through continuous fusion of features at various levels or inception structures, such as the combination of inception and dilated convolution in ADCrowdnet [50], jump connection in SaCNN [43]. Further research is encouraged and better performance can be expected.

(a) Dense crowd (b) Sparse crowd (c) View variation (d) Occlusion (e) Uneven distribution

Fig. 5 Typical challenges in crowd counting, reproduced with permission of Ref. [15], Copyright of 2020 IEEE

Moreover, many pruning works have fully proved that a set of parameters contributes little in the model, which bloats the model and limits its further application. Therefore, more lightweight models are likely to emerge in the future.

2. Construction of datasets

For the present, lack of data or large datasets remains one of the major difficulties in crowd counting. As mentioned in Sect. 3.1, most of the existing datasets have various problems, such as single scene (UCSD [11]), single angle (Mall [12]), small capacity (UCF_CC_50 [13]) and so on. For the construction of subsequent datasets, we make a few suggestions here:

a. Sample capacity. Deep learning requires huge sample capacity, which is often the premise of training a high-performance model. For existing datasets, UCSD [11] and Mall [12] contain 2000 images, UCF_CC_50 [13] contains 50 images, WorldExpo 10 [14] contains 3980 images, ShanghaiTech [15] contains 1198 images and GCC [17] contains 15212 virtual synthetic images. Compared with the datasets in the object detection field (such as MS COCO [158] and Caltech [159]) which commonly contains tens of thousands of samples, the dataset of the crowd counting field is extremely scarce.

b. Scene diversity. A good dataset requires not only a large sample size but rich scenes as well. This is usually the prime difficulty to overcome because it needs to take multiple sets of images in multiple locations, multiple angles, multiple lighting conditions.

c. Image quality. The purpose of the dataset is to make the trained model perform the best performance under the specific image. Therefore, the dataset should match the actual application scenarios as much as possible. Considering the diversity of image resolution in reality, cluster analysis is recommended to guide the distribution of image resolution in the dataset.

d. Annotation method. The annotation steps of various datasets are relatively consistent. First, manually mark every head in the image, and then the "ground truth" density map is generated by various methods. Quite evidently, the quality of the dataset depends largely on the performance of the annotation method. As described in Sect. 3.2, current methods include conversion using Gaussian kernel [15, 36], content-aware [147], Inverse K-Nearest Neighbor [148] and density contribution probability model [149]. Further use of context information and optimization of image edge parts should be considered for continued exploration.

3. Attention mechanism

Attention mechanism is currently mainly used to highlight crowded areas in the image and then optimize its counting effect, which solves the problem of uneven population distribution to a certain extent. Researchers currently mainly use the attention mechanism by scale, channel, space, etc., and have achieved remarkable results. However, the ideal effect has not yet been achieved and there is still much room for improvement. More accurate attention mechanism can be considered in the future to be applied to the processing of complex backgrounds in images, the generation of finer-grained density maps and more accurate counting.

4. Cross-domain integration

Many fields of computer vision can complement and promote each other. As mentioned above, the combination of crowd counting and small object detection can effectively enhance the detection effect [123, 125]. However, from the perspective of actual results (as shown in Fig. 4), there is still much room for improvement. At present, there exist little related works in the research direction, and more researches are encouraged to explore. Moreover, subsequent studies may consider combining crowd counting with super-resolution to improve the counting effect of dense crowd areas in the image.

5. Crowd location

A lot of works are devoted to drawing fine-grained density maps, and the effect is remarkable. As the density map becomes more and more refined, it is gradually possible to accurately locate the crowd in the image based on the density map. But for the time being there is no similar work yet, we encourage further research in this direction.

6. Few-shot Learning

In computer vision, we occasionally encounter problems for other kinds of object counting, not only human individuals. However, these categories of target usually lack a large number of datasets that are necessary for training an acceptable model, i.e., the problem of few-shot learning in encountered. Section 3.5 summarizes some existing works of general category object detection. Further research can be considered in the future research work.

7. Image processing: patches or whole

As described in Sect. 2.3.3, the input image is divided into multiple small patches and processed separately, thereby alleviating the problem of uneven distribution of crowd to some extent. In contrast, the attention mechanism tries to solve this problem in an alternative mode. We need to devote more efforts to figure out a loss function suitable for specific situations.

8. Loss function

Designing an appropriate loss function is critical for training a well-performed model. The most commonly used loss functions in crowd counting and density estimation are MAE, MSE, RMSE, etc. In addition, there are some customized loss functions, such as the

combination loss designed in TEDnet [79], a combination of adversarial loss and pixel-level Euclidean loss in CP-CNN [41], the combination of the L2 loss and scale-aware losses in Ref. [80], etc. Other specific loss functions can be examined for more accurate performance.

5 Conclusion

This article briefly summarizes the traditional crowd counting methods and the procedure of their development. We focused on the CNN-based crowd counting methods and divided them into seven categories according to their guiding ideology and then expounded separately. Secondly, we described some datasets and pointed out various existing annotation methods in detail. In order to compare the performance of the model more specifically, we tabulated the results of some popular methods on the mainstream datasets. In addition, we also summarized the work of video-based crowd counting and few-shot learning. Finally, we discussed the difficulties faced in crowd counting and their corresponding solutions and conclusions, future trends. We hope that this review will give an overall understanding of crowd counting and density estimation. The field still needs further research, and it is encouraged that more researchers focus on this field and make it more applicable in the future.

References

1. Ryan D, Denman S, Fookes C et al (2009) Crowd counting using multiple local features. In: 2009 Digital image computing: techniques and applications. IEEE, pp 81–88
2. Subburaman VB, Descamps A, Carincotte C (2012) Counting people in the crowd using a generic head detector. In: 2012 IEEE ninth international conference on advanced video and signal-based surveillance
3. Hou Y, Pang G (2011) People counting and human detection in a challenging situation. IEEE Trans Syst Man Cybern Part A 41(1):24–33
4. Handte M, Iqbal MU, Wagner S et al (2014) Crowd density estimation for public transport vehicles. EDBT/ICDT Workshops
5. Hussain N, Yatim HSM, Hussain NL et al (2011) Cdes: a pixel-based crowd density estimation system for masjid al-haram. Saf Sci 49(6):824–833
6. Yuan Y, Qiu C, Xi W et al (2011) Crowd density estimation using wireless sensor networks. In: 2011 Seventh international conference on mobile ad-hoc and sensor networks, 138–145
7. Zhe W, Hong L, Qian Y et al (2012) Crowd density estimation based on local binary pattern co-occurrence matrix. 2012 IEEE International Conference on Multimedia and Expo Workshops, 372–377.
8. Cho SY, Chow TWS, Leung CT (1999) A neural-based crowd estimation by hybrid global learning algorithm. IEEE Trans Syst Man Cybern B Cybern 29(4):535–541
9. Marana AN, Velastin SA, Costa LF et al (1998) Automatic estimation of crowd density using texture. Saf Sci 28(3):165–175
10. Ma R, Li L, Huang W et al (2004) On pixel count based crowd density estimation for visual surveillance. IEEE Conf Cybern Intell Syst 1:170–173
11. Chan AB, Liang Z-SJ, Vasconcelos N (2008) Privacy preserving crowd monitoring: counting people without people models or tracking. IEEE Conf Comput Vis Pattern Recognit 2008:1–7
12. Ke C, Chen CL, Gong S et al (2012) Feature mining for localised crowd counting. In: British machine vision conference
13. Idrees H, Saleemi I, Seibert C et al (2013) Multi-source multi-scale counting in extremely dense crowd images. In: IEEE conference on computer vision and pattern recognition
14. Cong Z, Li H, Wang X et al (2015) Cross-scene crowd counting via deep convolutional neural networks. In: 2015 IEEE conference on computer vision and pattern recognition (CVPR)
15. Zhang Y-Y, Zhou D, Chen S et al (2016) Single-image crowd counting via multi-column convolutional neural network. IEEE Conf Comput Vis Pattern Recognit 2016:589–597
16. Idrees H, Tayyab M, Athrey K et al (2018) Composition loss for counting, density map estimation and localization in dense crowds. In: ECCV
17. Wang Q, Gao J, Lin W et al (2019) Learning from synthetic data for crowd counting in the wild. IEEE CVF Conf Comput Vis Pattern Recognit 2019:8190–8199
18. Liu X, Weijer JVD, Bagdanov AD (2018) Leveraging unlabeled data for crowd counting by learning to rank. In: 2018 IEEE/CVF conference on computer vision and pattern recognition (CVPR)
19. Sam DB, Sajjan NN, Maurya H et al (2019) Almost unsupervised learning for dense crowd counting. In: AAAI
20. Lu E, Xie W, Zisserman A (2018) Class-agnostic counting. In: ACCV
21. Liu X, van de Weijer J, Bagdanov AD (2019) Exploiting unlabeled data in cnns by self-supervised learning to rank. IEEE Trans Pattern Anal Mach Intell 41:1862–1878
22. Zhan B, Monekosso D, Remagnino P et al (2008) Crowd analysis: a survey. Mach Vis Appl 19:345–357
23. Loy CC, Chen K, Gong S et al. (2013) Crowd counting and profiling: methodology and evaluation. In: Modeling, simulation and visual analysis of crowds
24. Ryan D, Denman S, Sridharan S et al (2015) An evaluation of crowd counting methods, features and regression models. Comput Vis Image Underst 130:1–17
25. Zitouni MS, Bhaskar H, Dias J et al (2016) Advances and trends in visual crowd analysis: a systematic survey and evaluation of crowd modelling techniques. Neurocomputing 186:139–159
26. Saleh SMA, Suandi SA, Ibrahim H (2015) Recent survey on crowd density estimation and counting for visual surveillance. Eng Appl Artif Intell 41:103–114
27. Sindagi V, Patel V (2018) A survey of recent advances in cnn-based single image crowd counting and density estimation. **abs/1707.01202**
28. Ilyas N, Shahzad A, Kim K (2020) Convolutional-neural network-based image crowd counting: review, categorization, analysis, and performance evaluation. Sensors (Basel, Switzerland), 20(1)
29. Nguyen VTT, Ngo TD (2019) Single-image crowd counting: a comparative survey on deep learning-based approaches. Int J Multimed Inf Retr 9:63–80
30. Cenggoro TW (2019) Deep learning for crowd counting: a survey. EMACS, 1(1)
31. Abdou M, Erradi A (2020) Crowd counting: a survey of machine learning approaches. In: 2020 IEEE international conference on informatics, IoT, and enabling technologies (ICIoT). pp 48–54

32. Luo Y, Lu J, Zhang B (2020) Crowd counting for static images: a survey of methodology. In: 2020 39th Chinese control conference (CCC). pp 6602–6607
33. Jeevitha S, Rajeswari R(2019) A Review of Crowd Counting Techniques. Internat J Res Anal Rev 5(3)
34. Gao G, Gao J, Liu Q et al (2020) Cnn-based density estimation and crowd counting: a survey. **abs/2003.12783**)
35. Rabichith SPK, Nithya S, Borra S (2018) Crowd density estimation using image processing: a survey. Int J Appl Eng Res 13(9): 6855–6864
36. Lempitsky VS, Zisserman A (2010) Learning to count objects in images. In: NIPS
37. Pham VQ, Kozakaya T, Yamaguchi O et al (2015) Count forest: co-voting uncertain number of targets using random forest for crowd density estimation. In: International conference on computer vision (ICCV 2015)
38. Boominathan L, Kruthiventi SSS, Babu RV (2016) Crowdnet: a deep convolutional network for dense crowd counting. In: Proceedings of the 24th ACM international conference on multimedia
39. Ooro-Rubio D, Lpez-Sastre RJ (2016) Towards perspective-free object counting with deep learning. ECCV
40. Zeng L, Xu X, Cai B et al (2017) Multi-scale convolutional neural networks for crowd counting. IEEE International Conference on Image Processing (ICIP), 465-469
41. Sindagi VA, Patel VM (2017) Generating high-quality crowd density maps using contextual pyramid cnns. In: IEEE international conference on computer vision
42. Sam DB, Surya S, Babu RV (2017) Switching convolutional neural network for crowd counting. In: 2017 IEEE conference on computer vision and pattern recognition (CVPR)
43. Zhang L, Shi M, Chen Q (2017) Crowd counting via scale-adaptive convolutional neural network. IEEE Winter Conf Appl Comput Vis (WACV) 2018:1113–1121
44. Li Y, Zhang X, Chen D (2018) Csrnet: dilated convolutional neural networks for understanding the highly congested scenes. IEEE CVF Conf Comput Vis Pattern Recognit 2018:1091–1100
45. Liu L, Wang H, Li G et al (2018) Crowd counting using deep recurrent spatial-aware network. In: IJCAI
46. Ranjan V, Le H, Hoai M (2018) Iterative crowd counting. In: ECCV
47. Cao X, Wang Z, Zhao Y et al (2018) Scale aggregation network for accurate and efficient crowd counting. In: Proceedings of the European conference on computer vision (ECCV), pp 734–750
48. Wu X, Zheng Y, Ye H et al (2019) Adaptive scenario discovery for crowd counting. In: ICASSP 2019–2019 IEEE international conference on acoustics, speech and signal processing (ICASSP). pp 2382–2386
49. Liu J, Gao C, Meng D et al (2017) Decidenet: counting varying density crowds through attention guided detection and density estimation. IEEE CVF Conf Comput Vis Pattern Recognit 2018:5197–5206
50. Liu N, Long Y, Zou C et al (2018) Adcrowdnet: an attention-injective deformable convolutional network for crowd understanding. IEEE CVF Conf Comput Vis Pattern Recognit 2019:3220–3229
51. Dong L, Zhang H, Ji Y et al (2020) Crowd counting by using multi-level density-based spatial information: a multi-scale cnn framework. Inf Sci 528:79–91
52. Hu Y, Jiang X, Liu X et al (2020) Nas-count: counting-by-density with neural architecture search. arXiv:2003.00217
53. Liu L, Lu H, Zou H et al (2020) Weighing counts: sequential crowd counting by reinforcement learning. arXiv:2007.08260
54. Jiang X, Zhang L, Xu M et al (2020) Attention scaling for crowd counting. IEEE CVF Conf Comput Vis Pattern Recognit 2020:4705–4714
55. Bai S, He Z, Qiao Y et al (2020) Adaptive dilated network with self-correction supervision for counting. IEEE CVF Conf Comput Vis Pattern Recognit 2020:4593–4602
56. Liu X, Yang J, Ding W (2020) Adaptive mixture regression network with local counting map for crowd counting. arXiv :2005.05776
57. Wu B, Nevatia R (2005) Detection of multiple, partially occluded humans in a single image by Bayesian combination of edgelet part detectors. In: Tenth IEEE international conference on computer vision, 2005. ICCV 2005
58. Sabzmeydani P, Mori G (2007) Detecting pedestrians by learning shapelet features. In: 2007 IEEE computer society conference on computer vision and pattern recognition (CVPR 2007), 18–23 June 2007. Minneapolis, Minnesota, USA
59. Dalal N, Triggs B (2005) Histograms of oriented gradients for human detection. In: 2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05), vol 1. pp 886–893
60. Viola P, Jones MJ (2004) Robust real-time face detection. Int J Comput Vis 57(2):137–154
61. Gao C, Liu J, Feng Q et al (2016) People-flow counting in complex environments by combining depth and color information. Multimed Tools Appl 75(15):9315–9331
62. Viola PA, Jones MJ, Snow D (2003) Detecting pedestrians using patterns of motion and appearance. In: Proceedings ninth IEEE international conference on computer vision, vol 2. pp 734–741
63. Gall J, Member IEEE et al (2011) Hough forests for object detection, tracking, and action recognition. IEEE Trans Pattern Anal Mach Intell 33(11):2188–2202
64. Min L, Zhang Z, Huang K et al (2008) Estimating the number of people in crowded scenes by mid based foreground segmentation and head-shoulder detection, in 19th International Conference on Pattern Recognition (ICPR 2008), December 8–11, 2008. Tampa, Florida
65. Wu B, Nevatia R (2007) Detection and tracking of multiple, partially occluded humans by Bayesian combination of edgelet based part detectors. Int J Comput Vis 75(2):247–266
66. Lin S, Chen J, Chao H (2001) Estimation of number of people in crowded scenes using perspective transformation. IEEE Trans Syst Man Cybern Part A Syst Hum 31(6):645–654
67. Felzenszwalb PF, Girshick RB, Mcallester D et al (2010) Object detection with discriminatively trained part-based models. IEEE Trans Pattern Anal Mach Intell 32(9):1627–1645
68. Laradji IH, Rostamzadeh N, Pinheiro PHO et al (2018) Where are the blobs: counting by localization with point supervision. In: ECCV
69. Liu Y, Shi M, Zhao Q et al (2019) Point in, box out: beyond counting persons in crowds. IEEE CVF Conf Comput Vis Pattern Recognit 2019:6462–6471
70. Kong D, Gray D, Tao H (2006) A viewpoint invariant approach for crowd counting. In: 18th international conference on pattern recognition (ICPR 2006), 20–24 August 2006. China, Hong Kong
71. Chan AB, Vasconcelos N (2009) Bayesian poisson regression for crowd counting. In: 2009 IEEE 12th international conference on computer vision
72. Chen K, Gong S, Xiang T et al (2013) Cumulative attribute space for age and crowd density estimation. In: IEEE conference on computer vision and pattern recognition
73. Rodriguez M, Laptev I, Sivic J et al (2011) Density-aware person detection and tracking in crowds. In: IEEE international conference on computer vision, ICCV 2011, Barcelona, Spain, November 6–13, 2011

74. KONG D (2005) Counting pedestrians in crowds using viewpoint invariant training. In: British machine vision conference
75. Li J, Huang L, Liu C (2011) Robust people counting in video surveillance: dataset and system. In: 2011 8th IEEE international conference on advanced video and signal based surveillance (AVSS). pp 54–59
76. Phu D, Julien C, Brmond BF et al (2013) Author manuscript, published in "ieee international conference on advanced video and signal-based surveillance (2013)" online tracking parameter adaptation based on evaluation
77. Lin T-Y, Lin Y-Y, Weng M-F et al (2011) Cross camera people counting with perspective estimation and occlusion handling. IEEE Int Workshop Inf Forensics Secur 2011:1–6
78. Min F, Pei X, Li X et al (2015) Fast crowd density estimation with convolutional neural networks. Eng Appl Artif Intell 43:81–88
79. Jiang X, Xiao Z, Zhang B et al (2019) Crowd counting and density estimation by trellis encoder–decoder networks. IEEE CVF Conf Comput Vis Pattern Recognit 2019:6126–6135
80. Varior RR, Shuai B, Tighe J et al (2019) Multi-scale attention network for crowd counting. CVPR
81. Gao J, Wang Q, Yuan Y (2019) Scar: spatial-/channel-wise attention regression networks for crowd counting. Neurocomputing 363:1–8
82. Zhu L, Zhao Z, Lu C et al (2019) Dual path multi-scale fusion networks with attention for crowd counting. arXiv:1902.01115
83. Zou Z, Cheng Y, Qu X et al (2019) Attend to count: crowd counting with adaptive capacity multi-scale cnns. Neurocomputing 367:75–83
84. Sam DB, Sajjan NN, Babu RV (2018) Divide and grow: capturing huge diversity in crowd images with incrementally growing cnn. IEEE CVF Conf Comput Vis Pattern Recognit 2018:3618–3626
85. Shen Z, Xu Y, Ni B et al (2018) Crowd counting via adversarial cross-scale consistency pursuit. IEEE CVF Conf Comput Vis Pattern Recognit 2018:5245–5254
86. Yang J, Zhou Y, Kung S-Y (2018) Multi-scale generative adversarial networks for crowd counting. In: 2018 24th international conference on pattern recognition (ICPR)
87. Wang L, Li Y, Xue X (2019) Coda: counting objects via scale-aware adversarial density adaption. IEEE Int Conf Multimed Expo (ICME) 2019:193–198
88. Liu W, Salzmann M, Fua P (2018) Context-aware crowd counting. IEEE CVF Conf Comput Vis Pattern Recognit CVPR 2019:5094–5103
89. Sang J, Wu W, Luo H et al (2019) Improved crowd counting method based on scale-adaptive convolutional neural network. IEEE Access 7:24411–24419
90. Zou Z, Su X, Qu X et al (2018) Da-net: learning the fine-grained density distribution with deformation aggregation network. IEEE Access 6:60745–60756
91. Szegedy C, Liu W, Jia Y et al (2015) Going deeper with convolutions. IEEE Conf Comput Vis Pattern Recognit CVPR 2015:1–9
92. Chen Z, Cheng J, Yuan Y et al (2019) Deep density-aware count regressor. arXiv:1908.03314
93. Deb D, Ventura J (2018) An aggregated multicolumn dilated convolution network for perspective-free counting. In: 2018 IEEE/cvf conference on computer vision and pattern recognition workshops (CVPRW)
94. Liu M, Jiang J, Guo Z et al (2018) Crowd counting with fully convolutional neural network. In: 2018 25th IEEE international conference on image processing (ICIP). IEEE, pp 953–957
95. Wang Z, Xiao Z, Xie K et al (2018) In defense of single-column networks for crowd counting. In: BMVC
96. Dai F, Liu H, Ma Y et al (2019) Dense scale network for crowd counting. arXiv:1906.09707
97. Kang D, Chan AB (2018) Crowd counting by adaptively fusing predictions from an image pyramid. In: BMVC
98. Gao J, Wang Q, Li X (2019) Pcc net: perspective crowd counting via spatial convolutional network. arXiv:1905.10085
99. Hossain M, Hosseinzadeh M, Chanda O et al (2019) Crowd counting using scale-aware attention networks. IEEE Winter Conf Appl Comput Vis WACV 2019:1280–1288
100. Sindagi V, Patel VM (2019) Ha-ccn: hierarchical attention-based crowd counting network. IEEE Trans Image Process 29:323–335
101. Ranjan V, Shah M, Nguyen MH (2019) Crowd transformer network. arXiv:1904.02774
102. Sindagi V, Patel VM (2019) Inverse attention guided deep crowd counting network. In: 2019 16th IEEE international conference on advanced video and signal based surveillance (AVSS). pp 1–8
103. Kasmani SA, He X, Jia W et al (2018) A-ccnn: adaptive ccnn for density estimation and crowd counting. In: 2018 25th IEEE international conference on image processing (ICIP). pp 948–952
104. Tian Y, Lei Y, Zhang J et al (2018) Padnet: pan-density crowd counting. IEEE Trans Image Process 29:2714–2727
105. Zhang Y, Chang F, Wang M et al (2018) Auxiliary learning for crowd counting via count-net. Neurocomputing 273:190–198
106. Han K, Wan W, Yao H et al (2017) Image crowd counting using convolutional neural network and markov random field
107. Shi M, Yang Z, Xu C et al (2018) Revisiting perspective information for efficient crowd counting. IEEE CVF Conf Comput Vis Pattern Recognit CVPR 2019:7271–7280
108. Kumagai S, Hotta K, Kurita T (2017) Mixture of counting cnns: adaptive integration of cnns specialized to specific appearance for crowd counting. arXiv:1703.09393
109. Olmschenk G, Hao T, Zhu Z (2018) Crowd counting with minimal data using generative adversarial networks for multiple target regression. In: 2018 IEEE winter conference on applications of computer vision (WACV)
110. Chong S, Ai H, Bo B (2016) End-to-end crowd counting via joint learning local and global count
111. Xu C, Qiu K, Fu J et al (2019) Learn to scale: generating multipolar normalized density maps for crowd counting. IEEE CVF Int Conf Comput Vis (ICCV) 2019:8381–8389
112. Xiong H, Lu H, Liu C et al (2019) From open set to closed set: counting objects by spatial divide-and-conquer. IEEE CVF Int Conf Comput Vis (ICCV) 2019:8361–8370
113. Liu C, Weng X, Mu Y (2019) Recurrent attentive zooming for joint crowd counting and precise localization. In: 2019 IEEE/CVF conference on computer vision and pattern recognition (CVPR). IEEE, pp 1217–1226
114. Xiong F, Shi X, Yeung D-Y (2017) Spatiotemporal modeling for crowd counting in videos. IEEE Int Conf Comput Vis ICCV 2017:5161–5169
115. Shi X, Chen Z, Wang H et al (2015) Convolutional lstm network: a machine learning approach for precipitation nowcasting, pp 802–810
116. Fang Y, Zhan B, Cai W et al (2019) Locality-constrained spatial transformer network for video crowd counting. IEEE Int Conf Multimed Expo ICME 2019:814–819
117. Wu X, Xu B, Zheng Y et al (2019) Video crowd counting via dynamic temporal modeling. arXiv:1907.02198
118. Zou Z, Shao H, Qu X et al (2019) Enhanced 3d convolutional networks for crowd counting. arXiv:1908.04121
119. Zheng H, Lin Z, Cen J et al (2019) Cross-line pedestrian counting based on spatially-consistent two-stage local crowd density estimation and accumulation. IEEE Trans Circuits Syst Video Technology 29(3):787–799
120. He G, Chen Q, Jiang D et al (2017) A double-region learning algorithm for counting the number of pedestrians in subway surveillance videos. Eng Appl Artif Intell 64:302–314

121. He G, Ma Z, Huang B et al (2019) Dynamic region division for adaptive learning pedestrian counting. IEEE Int Conf Multimed Expo ICME 2019:1120–1125
122. Redmon J, Farhadi A (2018) Yolov3: an incremental improvement. arXiv:1804.02767
123. Lian D, Li J, Zheng J et al (2019) 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. pp 1821–1830
124. Lin TY, Goyal P, Girshick R et al (2017) Focal loss for dense object detection. IEEE Trans Pattern Anal Mach Intell 99:2999–3007
125. Sam DB, Peri SV, Mukuntha NS et al (2020) Locate, size and count: accurately resolving people in dense crowds via detection. In: IEEE transactions on pattern analysis and machine (intelligence)
126. Hu P, Ramanan D (2017) Finding tiny faces. IEEE Conf Comput Vis Pattern Recognit CVPR 2017:1522–1530
127. Jiang S, Lu X, Lei Y et al (2019) Mask-aware networks for crowd counting. arXiv:1901.00039
128. Kang D, Dhar D, Chan AB (2016) Crowd counting by adapting convolutional neural networks with side information. arXiv :1611.06748
129. Marsden M, McGuinness K, Little S et al (2017) Fully convolutional crowd counting on highly congested scenes. arXiv :1612.00220
130. Valloli VK, Mehta K (2019) W-net: reinforced u-net for density map estimation. arXiv:1903.11249
131. Ronneberger O, Fischer P, Brox T (2015) U-net: convolutional networks for biomedical image segmentation. arXiv:1505.04597
132. Ding X, Lin Z, He F et al (2018) A deeply-recursive convolutional network for crowd counting. In: 2018 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, pp 1942–1946
133. Zhao M, Zhang J, Zhang C et al (2019) Leveraging heterogeneous auxiliary tasks to assist crowd counting. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 12736–12745
134. Wan J, Luo W, Wu B et al (2019) Residual regression with semantic prior for crowd counting. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp 4036–4045
135. Huang S, Li X, Zhang Z et al (2018) Body structure aware deep crowd counting. IEEE Trans Image Process 27:1049–1059
136. Yang Y, Li G, Wu Z et al (2020) Reverse perspective network for perspective-aware object counting. IEEE CVF Conf Compu Vis Pattern Recognit CVPR 2020:4373–4382
137. hwan Oh M, Olsen PA, Ramamurthy KN (2020) Crowd counting with decomposed uncertainty. arXiv:1903.07427
138. Shi Z, Zhang L, Sun Y et al (2018) Multiscale multitask deep netvlad for crowd counting. IEEE Trans Industr Inf 14(11):4953–4962
139. Wei X, Du J, Liang M et al (2017) Boosting deep attribute learning via support vector regression for fast moving crowd counting. Pattern Recognit Lett 119:12–23
140. Shi Z, Le Z, Yun L et al (2018) Crowd counting with deep negative correlation learning. In: 2018 IEEE/CVF conference on computer vision and pattern recognition (CVPR)
141. Oncel T (2008) Pedestrian detection via classification on riemannian manifolds. IEEE Trans Pattern Anal Mach Intell 30:1713–1727
142. Arteta C, Lempitsky V, Zisserman A (2016) Counting in the wild
143. Marsden M, McGuinness K, Little S et al (2017) Resnetcrowd: a residual deep learning architecture for crowd counting, violent behaviour detection and crowd density level classification. In: 2017 14th IEEE international conference on advanced video and signal based surveillance (AVSS), pp 1–7
144. Liu Y, Liu L, Wang P et al. (2020) Semi-supervised crowd counting via self-training on surrogate tasks. arXiv:2007.03207
145. Yang Y, Wu Z, Su L et al (2020) Weakly-supervised crowd counting learns from sorting rather than locations
146. Wang Q, Gao J, Lin W et al. (2020) Nwpu-crowd: a large-scale benchmark for crowd counting. In: IEEE transactions on pattern analysis and machine intelligence
147. Oghaz MM, Khadka AR, Argyriou V et al (2019) Content-aware density map for crowd counting and density estimation. arXiv :1906.07258
148. Olmschenk G, Tang H, Zhu Z (2019) Improving dense crowd counting convolutional neural networks using inverse k-nearest neighbor maps and multiscale upsampling.arXiv:1902.05379
149. Ma Z, Wei X, Hong X et al (2019) Bayesian loss for crowd count estimation with point supervision. IEEE CVF Int Conf Comput Vis ICCV 2019:6141–6150
150. Wang Z, Bovik AC, Sheikh HR et al (2004) Image quality assessment: from error visibility to structural similarity. IEEE Trans Image Process 13:600–612
151. Sindagi VA, Patel VM (2017) Cnn-based cascaded multi-task learning of high-level prior and density estimation for crowd counting. In: 2017 14th IEEE international conference on advanced video and signal based surveillance (AVSS). pp 1–6
152. Sam DB, Babu RV (2018) Top-down feedback for crowd counting convolutional neural network. In: AAAI
153. Zhang L, Shi Z, Cheng M-M et al (2019) Nonlinear regression via deep negative correlation learning. In: IEEE transactions on pattern analysis and machine intelligence
154. Shi Z, Mettes P, Snoek CGM (2019) Counting with focus for free. IEEE CVF Int Conf Comput Vis ICCV 2019:4199–4208
155. Chen X, Bin Y, Sang N et al (2019) Scale pyramid network for crowd counting. IEEE Winter Conf Appl Comput Vis WACV 2019:1941–1950
156. Liu L, Qiu Z, Li G et al (2019) Crowd counting with deep structured scale integration network. IEEE CVF Int Conf Comput Vis ICCV 2019:1774–1783
157. Yan Z, Yuan Y, Zuo W et al (2019) Perspective-guided convolution networks for crowd counting. IEEE CVF Int Conf Comput Vis ICCV 2019:952–961
158. Lin T-Y, Maire M, Belongie SJ et al (2014) Microsoft coco: common objects in context. arXiv:1405.0312
159. Dollár P, Wojek C, Schiele B et al (2012) Pedestrian detection: an evaluation of the state of the art. IEEE Trans Pattern Anal Mach Intell 34:743–761