Received: 8 February 2023 | Revised: 12 April 2023 | Accepted: 8 May 2023

CAAI Transactions on Intelligence Technology

DOI: 10.1049/cit2.12241

The Institution of Engineering and Technology WILEY

REVIEW

Deep learning in crowd counting: A survey

Lijia Deng[1] | Qinghua Zhou[1] | Shuihua Wang[1,2] | Juan Manuel Górriz[3] | Yudong Zhang[1,2]

[1]School of Computing and Mathematical Sciences, University of Leicester, Leicester, UK

[2]Department of Information Systems, Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah, Saudi Arabia

[3]Department of Signal Theory, Networking and Communications, University of Granada, Granada, Spain

Correspondence

Yudong Zhang.
Email: yudongzhang@ieee.org

Funding information

BHF, Grant/Award Number: AA/18/3/34220; Hope Foundation for Cancer Research, Grant/Award Number: RM60G0680; GCRF, Grant/Award Number: P202PF11; Sino-UK Industrial Fund, Grant/Award Number: RP202G0289; LIAS, Grant/Award Numbers: P202ED10, P202RE969; Data Science Enhancement Fund, Grant/Award Number: P202RE237; Sino-UK Education Fund, Grant/Award Number: OP202006; Fight for Sight, Grant/Award Number: 24NN201; Royal Society International Exchanges Cost Share Award, Grant/Award Number: RP202G0230; MRC, Grant/Award Number: MC_PC_17171; BBSRC, Grant/Award Number: RM32G0178B8

Abstract

Counting high-density objects quickly and accurately is a popular area of research. Crowd counting has significant social and economic value and is a major focus in artificial intelligence. Despite many advancements in this field, many of them are not widely known, especially in terms of research data. The authors proposed a three-tier standardised dataset taxonomy (TSDT). The Taxonomy divides datasets into small-scale, large-scale and hyper-scale, according to different application scenarios. This theory can help researchers make more efficient use of datasets and improve the performance of AI algorithms in specific fields. Additionally, the authors proposed a new evaluation index for the clarity of the dataset: average pixel occupied by each object (APO). This new evaluation index is more suitable for evaluating the clarity of the dataset in the object counting task than the image resolution. Moreover, the authors classified the crowd counting methods from a data-driven perspective: multi-scale networks, single-column networks, multi-column networks, multi-task networks, attention networks and weak-supervised networks and introduced the classic crowd counting methods of each class. The authors classified the existing 36 datasets according to the theory of three-tier standardised dataset taxonomy and discussed and evaluated these datasets. The authors evaluated the performance of more than 100 methods in the past five years on different levels of popular datasets. Recently, progress in research on small-scale datasets has slowed down. There are few new datasets and algorithms on small-scale datasets. The studies focused on large or hyper-scale datasets appear to be reaching a saturation point. The combined use of multiple approaches began to be a major research direction. The authors discussed the theoretical and practical challenges of crowd counting from the perspective of data, algorithms and computing resources. The field of crowd counting is moving towards combining multiple methods and requires fresh, targeted datasets. Despite advancements, the field still faces challenges such as handling real-world scenarios and processing large crowds in real-time. Researchers are exploring transfer learning to overcome the limitations of small datasets. The development of effective algorithms for crowd counting remains a challenging and important task in computer vision and AI, with many opportunities for future research.

KEYWORDS

artificial intelligence, computer vision, image analysis, image processing

1 | INTRODUCTION

The task of obtaining the number of people from an image of a video is called crowd counting. Crowd counting is a hot research topic in the field of computer vision and intelligent video surveillance. Since 2008, researchers have built crowd-monitoring and scene-understanding cognitive systems that can benefit society and public safety [1–3]. In addition, people can use algorithmic strategies based on crowd counting to assist in completing tasks such as behavior analysis [3, 4], congestion analysis [5, 6], anomaly detection [7, 8], and event detection [9, 10]. For example, monitoring the crowd in a square or a mall during a sports competition or a festival celebration can prevent riots and trampling accidents [11, 12]. Meanwhile, investigating the number of people in a mall or playground could be used to determine their business capabilities, and it also has a certain effect on economic research. Moreover, the study of crowd counting and density estimate can also be used in other different fields, such as psychological effects of people gathering groups [13] and animal migration [14] and bacterial activity [15]. Overall, crowd counting has many potential applications in various fields where understanding crowd behaviour is important and these applications including:

1. Safety monitoring: Video surveillance cameras are widely used for security and safety purposes, but traditional surveillance algorithms may struggle with high-density crowds. Algorithms designed for crowd analysis tasks such as behaviour analysis [3, 4, 16], congestion analysis [5, 6], anomaly detection [7, 8], and event detection [9, 10] can be leveraged for these scenarios.
2. Disaster management: Crowd analysis can be used to detect overcrowding early and manage crowds effectively, thus preventing disasters in scenarios such as sports events, music concerts, public demonstrations, and political rallies [17, 18].
3. Design of public spaces: Crowd analysis can reveal design shortcomings in public spaces such as airport terminals, train stations, and shopping malls, allowing for the optimisation of safety and crowd movement [19, 20].
4. Evidence-based decision-making: Crowd counting techniques can be used to gather intelligence for further analysis and inference, such as in retail for appropriate product placement, staff optimisation, and pedestrian flow analysis [19, 21].
5. Virtual environments: Crowd analysis methods can help establish mathematical models that accurately simulate crowd phenomena, useful for computer games, film scenes, and designing evacuation plans [22, 23].
6. Technology transfer: The methods used for crowd counting can also be used in other categories of object counting work like: crop yield estimation [24], medical testing [25], and intelligent transportation system [26].

Before 2010, researchers began working on crowd counting research [27–32]. They used some traditional methods like detection-based methods and regression-based methods to do crowd counting. For the detection-based methods, like Figure 1, researchers use a sliding window to detect the people in an image and then use this information to count the number of people [33, 34]. However, in the case of extremely crowded scenes, which are difficult to detect (e.g., dense density, severe occlusion) for classical methods, the regression-based method comes in handy.

The regression-based methods will learn a mapping between the number of people and local image patch features [1, 35, 36]. They obtain this mapping from low-level features such as global features and local features using regression approaches such as linear regression, Gaussian regression, ridge regression, and neural networks [1, 36–38]. Background subtraction techniques can extract global features based on blobs, such as area, perimeter-area ratio, and perimeter from foreground segments [36, 39]. Like gradient and edge features, local feature extraction is helpful for regression modelling [40–42].

Some researchers used segmentation methods to count the number of people [1, 39]. The people in the image are segmented into several groups. Then, researchers count the total number of people by regressing the global properties of each group. Therefore, it can be said that this is also a kind of the crowd counting method based on regression [43].

The earlier regression-based methods have good results when facing dense density and occlusion crowds, but most of them get the count without the spatial location information that detection-based methods can mark out. In 2010, Lempitsky et al. pioneered the method of density map to count the number of objects [43]. As shown in Figure 2, the density map can record the number of objects and mark the position of these objects.

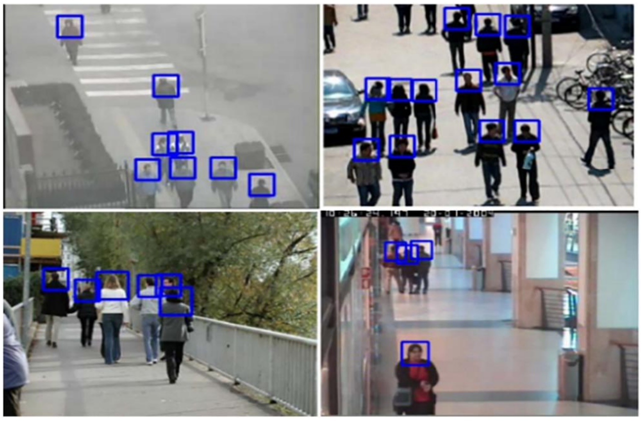


FIGURE 1 The detection based crowd counting approach [33].

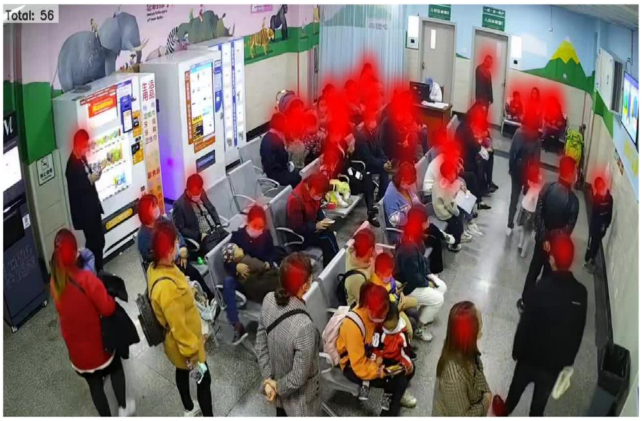


FIGURE 2 The density map estimation based crowd counting approach.

Convolutional Neural Networks (CNNs) have been successfully used in many computer-vision studies. In 2015, researchers began using CNNs to learn the non-linear regression function from the crowd images to the predicted density map [44, 45]. Then the crowd counting model based on CNNs has become the main framework used by many researchers [46–50].

However, the research on crowd counting still faces many challenges. First, the dense overlap of the crowd in the image makes the crowd counting network difficult to extract features from the entire head region. Second, the proximity of environmental features and crowd features can easily cause interference. Third, the feature size of the head shows a big difference because of severe perspective scaling and distortion. Fourth, the different image quality will affect the effect of feature extraction.

On the one hand, to overcome these challenges, researchers have improved and innovated crowd counting methods. On the other hand, since 2016, more and more researchers have noticed the lack of crowd datasets and started collecting new datasets. Nowadays, crowd counting-related datasets have started to become diverse, but some new datasets are similar to traditional datasets, and the other unique new datasets have not caught enough attention.

In past surveys [51–53], researchers mainly focused on providing an overview of recent advances in CNN-based methods for crowd counting and evaluation methods. Loy et al. [52] highlighted the importance of individual components in the processing pipeline and discuss the advantages of using regression-based techniques for handling more crowded scenes. Sindagi and Patel [53] discuss various CNN-based methods for crowd counting and density estimation, including regression-based methods, detection-based methods, and density estimation-based methods. In recent years, crowd counting algorithms and related datasets have developed rapidly. However, the above surveys do not provide a detailed description of the datasets, or many new datasets are not covered. If the validation of crowd counting algorithms is limited to only a few datasets, it will be detrimental to the development of crowd counting research.

With the enrichment of datasets, there are obvious differences between datasets. Some datasets are complex, and others are simple. For convolutional neural networks that rely significantly on data-driven, the dataset is one of the most important factors affecting the model. What are the current evaluation criteria for these datasets, and how can they be classified?

Therefore, we will describe the three-tier standard dataset taxonomy, show researchers' efforts in expanding the crowd dataset in recent years, and introduce some classic data-driven CNN methods from six different directions. Finally, the challenges in the field of crowd counting are discussed from the perspective of theory and practice. In short, our contributions are:

- We propose a three-tier standardised dataset taxonomy (TSDT).
- We propose a new evaluation index for the clarity of the dataset: average pixel occupied by each object (APO).
- Thirty-six datasets are classified and discussed.
- We distinguish the crowd counting algorithm from the data-driven perspective and discuss each category's classical algorithm.
- More than 110 algorithms have been evaluated.
- We discussed the theoretical and practical challenges of crowd counting from the perspective of data, algorithms and computing resources.

The following paper will introduce the datasets of crowd counting and describe each dataset's basic conditions and parameters first. Then, this paper will introduce some classical methods of crowd counting and compare the latest methods on estimated accuracy. Finally, this paper will describe the challenges of crowd counting and give some successful solutions.

2 | THREE-TIER STANDARDISED DATASET TAXONOMY AND CROWD COUNTING DATASETS

Data has always been one of the key points to drive artificial intelligence research. Different training data will lead to different training results. Therefore, we will first introduce the existing datasets in the crowd counting field.

2.1 | Three-tier standardised dataset taxonomy

After evaluating more than 100 crowd counting algorithms in the past, we noticed that these algorithms usually only show advantages on some datasets, and the complexity of these data sets is usually positively related to the complexity of the algorithm. The model focuses on lightweight and has obvious advantages in small-scale datasets, but it is not ideal in high-complexity datasets. In addition, the network based on small-scale datasets training is not as good as the network based on complex datasets training in the face of complex datasets [54, 55].

In addition, extremely high-density images are introduced to improve the accuracy of crowd counting algorithms. The extremely complex images usually represent the maximum challenge point of the dataset [56]. On the other hand, since the main evaluation indicator of the crowd counting model are mean absolute error (MAE) and root mean squared error (RMSE), these two indicators are greatly affected by the number of people in the image. Therefore, the maximum and mean annotation count can well reflect the difficulty of the dataset, and good for enabling the model to compare among datasets of the same level, facilitating the introduction of more datasets into the model evaluation process.

According to the average number of tags contained in each image in the dataset, we divide the existing dataset into small-scale datasets, large-scale datasets, and hyper-scale datasets. The division rules of the datasets are shown in Table 1. We use the maximum number of annotations per image (N_{max}) or the average number of annotations per image (N_{avg}) to define the scale of the dataset. In a small-scale dataset, its N_{max} is less than 100 or its N_{avg} is less than 40. And the small-scale dataset was usually obtained through surveillance cameras in daily life, so it may be called a daily-type crowd dataset. The large-scale dataset is its N_{max} is from 100 to 1000 or N_{avg} is from 40 to 200. The larger scale dataset usually includes many images taken during crowd-gathering activities, so it may be called an assembly-type crowd dataset. In the hyper-scale dataset, its N_{max} is greater than or equals to 1000 or its N_{avg} is greater than or equals to 200. In pursuit of the peak performance of neural networks, researchers have also continuously proposed a variety of challenging hyper-scale and difficult datasets to test and improve the performance of the neural networks. These datasets have image complexity far beyond small-scale datasets and large-scale datasets. Complex and challenging datasets can better test the performance of neural networks. Moreover, these datasets can help researchers improve the model's ability to handle extreme situations. Because some of the original papers lack the parameters of the datasets, we will classify the datasets according to the available parameters.

In addition, better image clarity may lead to better performance of convolutional neural network (CNN) models [57–59]. Clearer images may provide more detailed and accurate information about the objects or individuals in the crowd, making it easier for the CNN model to accurately count them. In addition, high image quality can help reduce the noise and ambiguity in the image, which can improve the accuracy of CNNs [60].

However, for a long time, the main indicator to measure image clarity is image resolution, which may not a good indicator to measure whether the image is clear for the task of object counting. The more annotations an image contains, the more blurred each annotation in the image will be. The fuzzy object will directly affect the performance of the algorithm. In object counting and detection, we mainly focus on the clarity of the target, rather than the background.

Therefore, we proposed a new parameter: "average pixel occupied by each object" (APO). APO is an indicator used to measure the resolution of one kind of object in the image. Therefore, when calculating APO, only the detected targets need to be considered. The calculation process is shown in the Formula (1). The APO usually better reflects the clarity of the crowd in the image than the resolution of the image. APO specifically captures the clarity of the object of interest, rather than the overall image clarity or resolution.

$$APO = \sqrt{\frac{H \times W}{C}} \tag{1}$$

where APO is the average pixel occupied by each object, which is the head in crowd counting. H and W represent the height and width of the image. C means the number of annotations in the image. In this survey, for the APO of the dataset, for ease of calculation, we use the mean value to calculate it for ease of calculation., as $APO = \sqrt{\frac{H_{avg} \times W_{avg}}{C_{avg}}}$, where, the average number of annotations per image C_{avg} can be calculated by the total number of annotations and number of images in the dataset: $C_{avg} = C_{total}/N_{img}$.

The clarity of people in the image greatly affects the recognition performance of a network. Based on this perspective, we evaluated the neural network trained with a dataset in different APOs. In the same dataset, the larger the APO, the clearer the head in the dataset. We compared the results of ResNet50 and VGG16 on Mall dataset with different APOs in Table 2, we noticed that the larger the image APO, the better the performance of the model.

TABLE 1 TSDT classification rules of dataset scale.

Classes	Max number of annotations per image (N_{max})	Average number of annotations per image (N_{avg})
Small-scale dataset	$N_{max} < 100$	$N_{avg} < 40$
Large-scale dataset	$100 \leq N_{max} < 1000$	$40 \leq N_{avg} < 200$
Hyper-scale dataset	$1000 \leq N_{max}$	$200 \leq N_{avg}$

2.2 | Small-scale datasets

This section will introduce the small-scale datasets. The images of small-scale datasets are usually from common scenes in daily life, such as shopping malls, schools, bus stops etc.

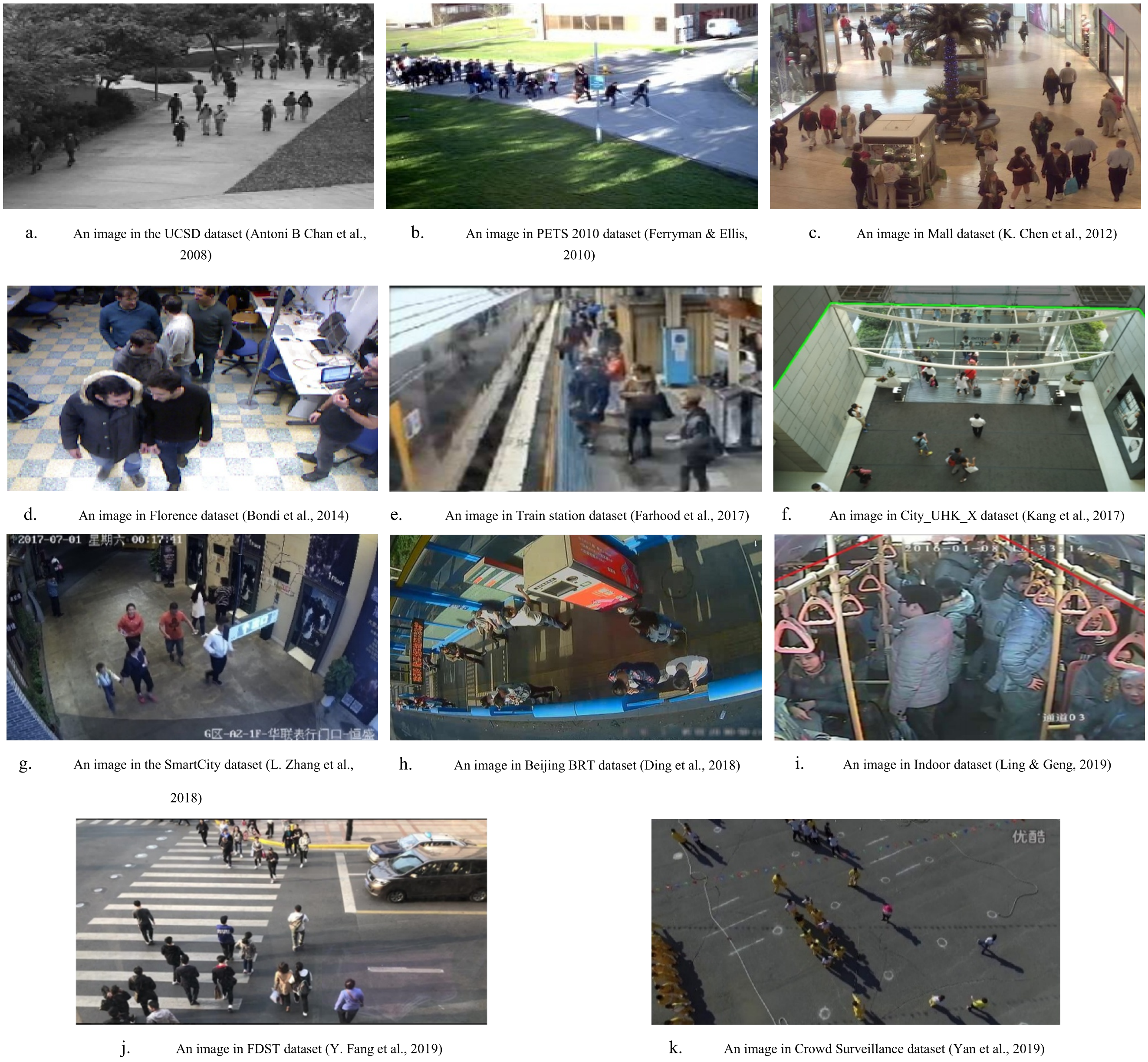
2.2.1 | UCSD dataset

The University of California San Diego (UCSD) dataset can be said as the first crowd-analysis dataset in the world. This dataset was collected by Chan et al. [1]. The UCSD dataset contains a variety of data, including crowd data related to crowd counting and other data related to object detection. As shown in Figure 3a, they used a fixed-position camera to take a one-hour video of people on a school path at UCSD. The video capture frame rate is 30 fps, and the size is 740 × 480. In order to reduce the consumption of computing resources, the original video is processed to a frame rate of 10 fps with a frame size of 238 × 158. The first 200 s of the video were

TABLE 2 Training results of models based on different APOs; Bold means the best.

Methods	APO	24.8	49.61	99.2
ResNet50	MAE	2.35	2.14	**2.07**
	RMSE	2.93	2.75	**2.62**
VGG16De	MAE	7.44	7.4	**7**
	RMSE	8.6	8.56	**8.1**

Note: The bold values mean the best results.



a. An image in the UCSD dataset (Antoni B Chan et al., 2008)

b. An image in PETS 2010 dataset (Ferryman & Ellis, 2010)

c. An image in Mall dataset (K. Chen et al., 2012)

d. An image in Florence dataset (Bondi et al., 2014)

e. An image in Train station dataset (Farhood et al., 2017)

f. An image in City_UHK_X dataset (Kang et al., 2017)

g. An image in the SmartCity dataset (L. Zhang et al., 2018)

h. An image in Beijing BRT dataset (Ding et al., 2018)

i. An image in Indoor dataset (Ling & Geng, 2019)

j. An image in FDST dataset (Y. Fang et al., 2019)

k. An image in Crowd Surveillance dataset (Yan et al., 2019)

FIGURE 3 Some sample images from small-scale datasets.

clipped to be a crowd dataset. There are a total of 2000 images in the UCSD crowd dataset, where 800 images are used as the training set, and other 1200 images are used as the test set. A larger proportion of test sets can better test the ability of the network to resist overfitting. The UCSD crowd dataset annotated 49885 people in total, including 29410 annotations in the test set and 20475 annotations in the training set. The maximum number of annotations per image is 46, and the minimum number of annotations per image is 11.

2.2.2 | PETS 2010 dataset

PETS dataset was collected by J. Ferryman et al. [61]. The dataset was collected by multiple cameras in Whiteknights Campus, University of Reading, UK, which included crowd counting, crowd density estimation, trajectory tracking, flow analysis, and event recognition. The data relating to crowd counting is in Dataset S1. This dataset has 1076 frames and 18289 annotations. The size of the image is 384 × 288. Each picture contains 0 to 40 people. The image of PETS data is shown in Figure 3b. The people in this dataset move in the same direction as one or more large clusters.

2.2.3 | Mall dataset

Chen et al. collected the mall dataset in a shopping mall through a publicly accessible surveillance camera [36]. As shown in Figure 3c, the image background of the MALL dataset is more complicated than it is in the UCSD dataset. The new features, such as the glass's and ground's reflection, greater perspective zoom, and models in the shop window, make the dataset more challenging. There are 2000 images in this dataset, and the image size is 640 × 480. The dataset has 62,325 people in total. Each image includes 13 to 53 people. The crowd presents two situations: some people are moving, and some are standing still or sitting down. In terms of space, the distribution of crowd areas is very uneven.

2.2.4 | Florence dataset

Florence dataset is a crowd dataset based on indoor surveillance images collected from the University of Florence by Enrico Bondi et al. [62]. As shown in Figure 3d, the background of the image in the Florence dataset is not very complex. This dataset contains 3358 images with a total of 17,630 people, ranging from 0 to 28 people in one image. This RGB-D imagery dataset has three video sequences: FLOW, QUEUE and GROUPS. In the FLOW part, people all walk from one side to another side. For the QUEUE sequence, the images show people queuing. This scene is similar to the path of the ticket gate at a station and a park. The slow-moving or stationary crowd may be recognised as a background. It raises the requirement for the robustness of the model. The GROUP sequence has two groups of people in each image. The image shows the interaction between the two groups of people, simulating the security issues in open and closed spaces. This dataset is more focused on crowd behaviour than other datasets.

2.2.5 | Train station dataset

The train station dataset was collected in a train station platform through a publicly accessible surveillance camera by Helia Farhood et al. [63]. The dataset contains 2000 images with 62581 people in total, ranging from 1 to 53 people in one image. However, as shown in Figure 3e, because of the camera's limitations and the high compression ratio of images, the image size is 256 × 256, which means the image resolution and quality are very low. On average, each head is only 45 pixels wide, and some heads are even hard to recognise by the human eye, which presents a huge challenge for crowd counting models. Besides, over time, the density of passengers on the platform will also change greatly.

2.2.6 | City_UHK_X dataset

The City_UHK_X dataset contains 3191 images with a size of 512 × 384, which was collected from the City University of Hong Kong (City_UHK) by Di Kang et al. [64]. The dataset has 106783 annotations in total, with an average of 33 annotations per image. The dataset contains 55 scenes, with an average of 58 images per scene. The scenes of the training set and the test set are completely different, which requires higher robustness of the model. As shown in Figure 3f, because the camera's tilt angle is greater than it is in the WorldExpo dataset, people in the image are more distorted.

2.2.7 | SmartCity dataset

SmartCity dataset was collected by Lu Zhang et al. [65]. This dataset mainly focuses on urban scenes such as office buildings, pedestrian streets, shopping malls etc. There are ten scenes and 50 images in total. This dataset contains 369 annotations, ranging from 1 to 14 per image, and the size of the images is 1920 × 1080. As shown in Figure 3g, this dataset contains some rare indoor scenes. This small-scale dataset requires a high generalisation ability of the model.

2.2.8 | Beijing BRT dataset

The Beijing BRT dataset was collected from a Beijing Bus Rapid Transit (BRT) platform by Xinghao Ding et al. [66]. This dataset contains 1280 images with a size of 360 × 640. There are 16,795 annotated people in total, ranging from 1 to 64 annotations in each image. The dataset collected images of the day from morning to night at the BRT station. As a result of this, this dataset contains obvious natural light changes, such as strong light, glare, shadows, reflections etc. As shown in Figure 3h, due to the use of a Fisheye Wide-angle Lens camera, objects at the edge of the image are distorted. In addition, the scenes in this dataset are very close to the application scenarios in real life. Combining the above factors, this dataset has a high requirement for the robustness of the model. It is worth mentioning that this dataset provides the perspective map, which reduces the difficulty of density map generation.

2.2.9 | Indoor dataset

The Indoor dataset is a dataset that focuses on indoor scenes collected by Miaogen Ling et al. [67]. This dataset contains three different scenes: classroom, canteen, and bus. The images of classroom scenes are from four Closed-Circuit Television (CCTV) videos of three different classrooms, and images of canteen scenes are from two different canteens. Three classroom videos were recorded in the daytime, and the other one was recorded at night. The indoor dataset contains a huge number of samples, including 148243 images, with a size of 740 × 576. The dataset contains 1834770 annotations in total, with the number of annotations per image ranging from 0 to 49. As Figure 3i shows, there are too many hindrances in the images of the bus scene. The crowd is highly dense, and the occlusion overlap is serious, which is very challenging for crowd counting.

2.2.10 | FDST dataset

Fudan-ShanghaiTech (FDST) dataset was collected by Yanyan Fang et al. [68]. This dataset contains 15000 images from 100 videos. This dataset includes 13 scenes. The image size is 1920 × 1080. A total of 394081 people have been annotated in total, with an average of 27 annotations per image. The training set and the test set are independent. The training set has 9000 images, and the test set has 6000 images.

2.2.11 | Crowd Surveillance

Crowd Surveillance was collected by Zhaoyi Yan et al. [50]. This dataset contains 13945 images with a size of 1342 × 840. A total of 386513 people have been annotated in the dataset, with an average number of 28 annotations per image. The dataset provides regions of interest (ROI), which can avoid the influence of some invalid regions with complex backgrounds. As shown in Figure 3k, apart from the conventional images, the dataset also contains aerial images.

2.3 | Large-scale datasets

This section will introduce the large-scale datasets. The large-scale datasets usually record scenes of various events, festivals, and parties.

2.3.1 | Zhengzhou Airport dataset

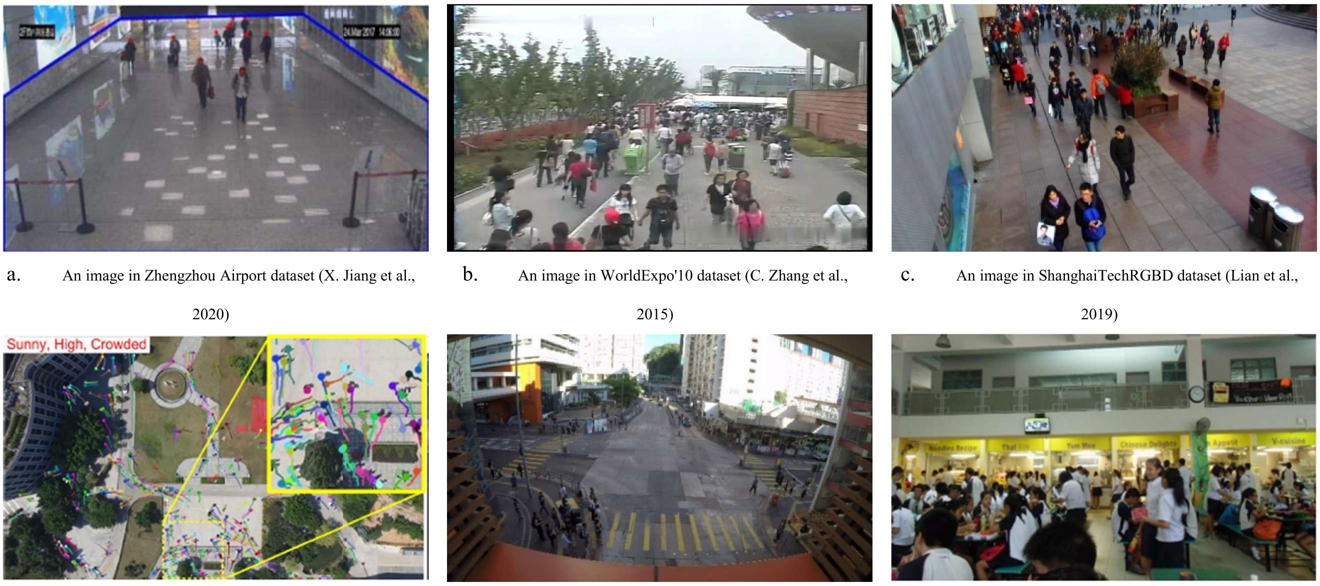
Zhengzhou Airport dataset was collected by Xiaoheng Jiang et al. [69]. The dataset is a cross-scenes dataset that covers six scenes from six CCTVs in Zhengzhou Airport. It contains 49061 annotations in 1111 images in total, with an average of 44 annotations per image. The number of annotations in each image is from 7 to 128. The huge change in crowd density puts higher demands on the robustness of the crowd counting model.

2.3.2 | WorldExpo'10 dataset

WorldExpo'10 dataset is a classical large-scale dataset that was collected by Cong Zhang et al. [70]. This dataset contains 3980 images with a size of 720 × 576 from 108 surveillance cameras around the Shanghai 2010 World Expo Park, as shown in Figure 4b. It has annotated 199,923 people with a range from 1 to 253 annotations per image. The training set and test set are independent of each other. The training set has 103 scenes, and the test set has the other five scenes. A perspective map has been provided for each scene.

2.3.3 | ShanghaiTechRGBD dataset

ShanghaiTechRGBD dataset is a large-scale RGB-D image dataset collected by Dongze Lian et al. [71]. As Figure 4c shows, the dataset shows people in public areas. This dataset contains 2193 images with a size of 1920 × 1080 and 144,512 annotations in total. Each image contains 6 to 234



a. An image in Zhengzhou Airport dataset (X. Jiang et al., 2020)

b. An image in WorldExpo'10 dataset (C. Zhang et al., 2015)

c. An image in ShanghaiTechRGBD dataset (Lian et al., 2019)

d. An image in Drone Crowd dataset (Wen et al., 2021)

e. An image in City Street dataset (Q. Zhang & Chan, 2019)

f. An image in Fine-Grained crowd dataset (Jia Wan et al., 2021)

FIGURE 4 Some sample images from large-scale datasets.

annotations. The dataset contains outdoor scenes, including streets and parks. The illumination of these scenes changes significantly. The images were taken by a stereo camera with a depth of field measurement of 20 m. An RGB-D image contains an ordinary RGB three-channel image and a Depth image. The Depth image is similar to the grayscale image, but each pixel value is the actual distance between the sensor and the object. Usually, RGB images and Depth images are registered [72].

2.3.4 | Drone Crowd dataset

The Drone Crowd dataset was collected by Longyin Wen et al. [73]. This dataset focuses on the use of drones in computer vision. They used a drone-mounted camera to take videos with 25 FPS and a resolution of 1920 × 1080. There are 33600 images from 112 videos in the dataset, including 4864280 annotations, and each image contains 25 to 455 annotations. The average number of annotations per image is 144.5. As Figure 4d shows, the image taken by the drone is very different from the image taken by the surveillance camera. This dataset is more similar to that of cell counting. Both focus on counting the number of objects on a plane.

2.3.5 | City Street dataset

The City Street dataset was collected by Qi Zhang et al. [74]. There are 500 images cut from videos of which 300 images are used for training and 200 images for testing. Each image has annotations from 70 to 150 with a resolution of 2704 × 1520. As shown in Figure 4e, The view angle is very high, and the camera is far away from pedestrians, which makes the pedestrians in the picture appear very small. This is a multi-view video dataset. Five surveillance cameras have been used to take videos from a busy street. The camera view and ground-plane ROIs of each view have been provided.

Moreover, this dataset provides tools to link the images of five independent surveillance cameras to build a 3D model of the street. The dataset contains two kinds of annotation based on 2D images and 3D street models. Compared to other datasets, this dataset provides three-dimensional spatial environment information.

2.3.6 | Fine-grained crowd dataset

This dataset was collected by Jia et al. [75]. The dataset contains more than 3700 images. The average number of annotations per image is 57, and its maximum number of annotations is 344. Jia et al. indicated that the current crowd counting algorithm only pays attention to the number of people in the image and lacks analysis of other crowd information [75]. In real life, the comprehensive information of the crowd is usually the most valuable. The behavior classification of people can guide social life more effectively, such as distinguishing people waiting in line and passers-by to determine the popularity of stores and distinguishing violent people and non-violent people to ensure social security. The dataset contains four categories of behaviour: the direction of driving on the sidewalk, standing or sitting, whether waiting in a queue, and whether showing violent behaviour. Because the features of different groups categories are similar, the challenge of fine-grained crowd counting is how to effectively use contextual information to distinguish categories. The task of this dataset is novel and practical. And this dataset is a good challenge and supplement to the existing crowd counting work.

2.4 | Hyper-scale datasets

This section will introduce the hyper-scale datasets. The hyper-scale dataset has some extremely dense crowd images, or the crowd density span is very large, which is suitable to be used as a challenging dataset to evaluate the network performance.

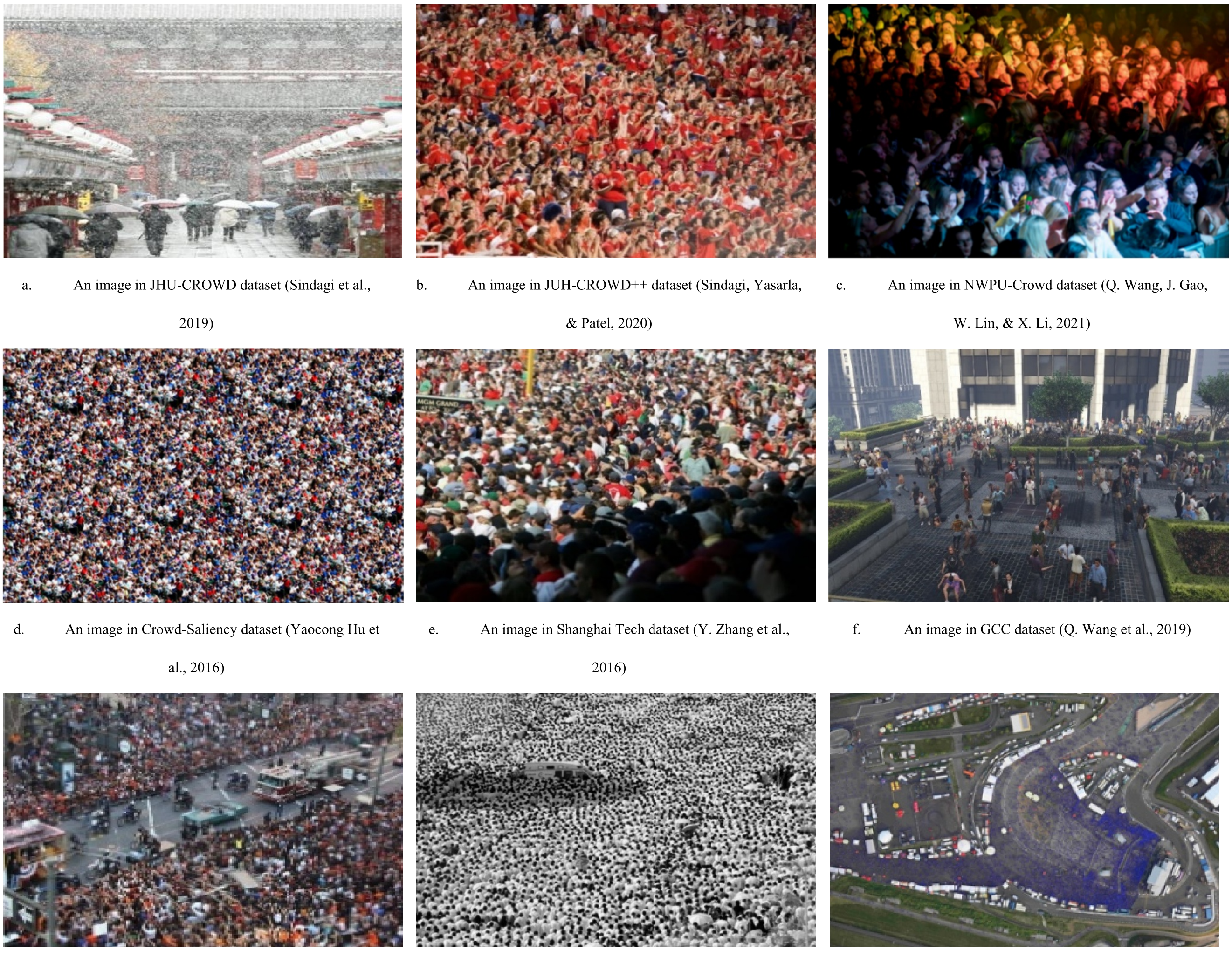
2.4.1 | JHU-CROWD and JHU-CROWD++ dataset

Johns Hopkins University crowd (JHU-CROWD) dataset collected 4250 images by Vishwanath A. Sindagi et al. [76]. The average number of annotations per image is 262, and the maximum number of annotations per image is 7286. The images come from the internet, so they have different scenes. Images from different scenes increase the diversity of data. Besides, as shown in Figure 5a, this dataset covered images in different weathers such as rain, snow, haze etc. For every image, information such as scenes and weather has been annotated. The dataset includes 100 images without a crowd to reduce the learning bias of the crowd counting model. Moreover, every annotation in the image contains more information, including the location of the head, the size of the head, corresponding occlusion level, and the blur level. Abundant information helps improve the learning efficiency of models.

On April 7th, 2020, Vishwanath et al. updated the JHU-CROWD dataset. They collected 4372 images in the new dataset named JHU-CROWD++ [77]. These new images are also collected from the internet. It contains pictures of different weather, such as 145, 201, and 168 images of rain, snow, and haze. This new dataset has 1515005 annotations in total, which is 31% more than the JHU-CROWD dataset. The largest number of annotations per image is 25791.

2.4.2 | NWPU-crowd dataset

The Northwestern Polytechnical University Crowd (NWPU-Crowd) dataset was collected by Qi Wang et al. [56]. The annotated 5109 images have an average resolution of 3383 × 2311 based on 2000 images and 200 videos collected by their team and from the internet. There are 2133238



a. An image in JHU-CROWD dataset (Sindagi et al., 2019)

b. An image in JUH-CROWD++ dataset (Sindagi, Yasarla, & Patel, 2020)

c. An image in NWPU-Crowd dataset (Q. Wang, J. Gao, W. Lin, & X. Li, 2021)

d. An image in Crowd-Saliency dataset (Yaocong Hu et al., 2016)

e. An image in Shanghai Tech dataset (Y. Zhang et al., 2016)

f. An image in GCC dataset (Q. Wang et al., 2019)

g. An image in the UCF-QNRF dataset (Idrees et al., 2018)

h. An image in UCF_CC_50 dataset (Idrees et al., 2013)

i. An image in the DLR-ACD dataset (Reza et al., 2019)

FIGURE 5 Some sample images from hyper-scale datasets.

annotations with a range from 0 to 20033 per image. Although most images in this dataset are small-scale crowd images, the dataset has approximately 1000 images with more than 500 people annotated and some images with more than 20000 annotations. The latter subset is very challenging for crowd counting models. As Figure 5c shows, some images show a huge change in illumination: some are extremely bright, and some are very dark. Some images have a very large resolution that is 4028 × 19044. The NWPU-Crowd dataset does not publish the labels of the test set because of fair evaluation.

2.4.3 | Crowd-Saliency dataset

The Crowd-Saliency dataset includes 107 images with a resolution of 720 × 576 collected by Yaocong Hu et al. [79]. The dataset annotated 45000 people with a range from 58 to 2201 per image. The average number of annotations per image is 422. The dataset covered different scenes, which can help to improve the robustness of the crowd counting model. Because of the low image resolution, such as Figure 5d, this dataset has the least average pixel width per head, which is 31.4.

2.4.4 | ShanghaiTech dataset

The ShanghaiTech dataset is a classic dataset collected by Yinyin Zhang et al. [80]. This dataset contains 1198 images and 330165 annotations in total. The dataset has been divided into two parts: Shanghai Tech A and Shanghai Tech B. Shanghai Tech A was collected from the internet, and Shanghai Tech B was collected from a busy street in Shanghai. Shanghai Tech A has 482 images and 241677 annotations, ranging from 33 to 3139 annotations per image. Moreover, Shanghai Tech B has 716 images with a range from 9 to 578 annotations per image. The average number of annotations is 501.3 for Shanghai Tech A and 501.3 and 123.6 for Shanghai Tech B. As shown in Figure 5e, Shanghai Tech A contains images with illumination changes. Researchers widely use this dataset because of its early publish time and high image quality. The different scenes and densities of the image can test the robustness of the crowd counting model. Besides, the images contain various scale changes and perspective distortion, which poses a higher challenge to crowd counting models.

2.4.5 | GCC dataset

The GTA5 Crowd Counting (GCC) dataset was collected from Grand Theft Auto V (GTA5) by Qi Wang et al. [81]. GTA5 is a computer game that contains a virtual Los Angeles-based city model. Qi Wang et al. use this game to simulate many crowd scenes. Through simulation, they obtained a very large amount of data with 400 different scenes. There are 15211 images with a resolution of 1920 × 1080, with a range from 0 to 3995 annotations per image. The images of the crowd with different sizes are evenly distributed in the dataset. The GCC dataset covered seven different weather scenes: clear, clouds, rain, foggy, thunder, overcast, and extra sunny. However, because of the limit of GTA5, the GCC dataset only has 256 different person models. As Figure 5f shows, there are still differences between simulated person models and real people.

2.4.6 | CrowdX dataset

The CrowdX dataset is a realistic simulation dataset generated by Hou et al. based on Unity3D [82]. It has 24,000 annotated images of crowds and can control the factors that influence the simulation. Compared to real-world datasets, it is easier to gather and annotate, and compared to other simulation datasets, it has a more realistic outcome and can adjust the factors affecting it.

2.4.7 | UCF-QNRF dataset

The UCF-QNRF dataset collected 1535 images with a resolution of 2902 × 2013 by Haroon Idrees et al. [83]. The train set has 1201 images, and the test set has 334. The images have been collected from three sources: Flickr, Web Search, and Hajj footage accounting for 90%, 7%, and 3%, respectively. There are 1251642 annotations with a range from 49 to 12865 annotations per image. The average number of annotations per image is 815.4, and the medium number of annotations per image is 425. As Figure 5g shows, the dataset's images are complex, and the crowd distribution is dense.

2.4.8 | UCF_CC_50 dataset

The UCF_CC_50 dataset was collected by the University of Central Florida (UCF) from the web including Flickr (CC is crowd counting in short) [84]. As Figure 5h shows, this is an extremely dense crowds' dataset. The dataset contains 50 images with 63974 annotations. The number of people annotated on each image varies from 94 to 4543, with an average of over 1000 annotations per image. Although the average resolution of the image is large, the pixel value of each head is still very small. These hyper-scale crowd images bring great challenges to model training. Since the dataset was published very early, this dataset is widely used to test the performance of the crowd counting model.

2.4.9 | DLR-ACD dataset

The DLR's Aerial Crowd Dataset (DLR-ACD) contains 33 aerial images of the extremely dense crowds collected by German Aerospace Center (DLR). [85]. The dataset contains 226291 annotations with a range from 285 to 24368 and an average of 6857.3 annotations per image. As Figure 5i shows, the image is obtained by standard Digital Single Lens Reflex (DSLR) cameras on a helicopter. Like the Drone Crowd dataset, it is also a dataset from an overlooking perspective where the crowd has little variation in perspective distortion. The image contains ground sampling distance information (GSD) which can be used to calculate the real area of the image. This dataset is extremely challenging for the crowd counting model.

2.5 | Summary of the datasets

We performed statistics on the parameters of these datasets in Table 3. These parameters include the total number of images, the average image pixel size, content information, publication time, and the scale of datasets. The content information of the picture includes the total number of annotations in the dataset, the average number of annotations, and the minimum number of annotations for a single image.

3 | DEEP LEARNING THEORIES IN CROWD COUNTING AND THEIR QUANTIFIED EVALUATION

Since "Crowd density estimation using texture analysis and learning" (2006) [87] and "Face recognition using kernel ridge regression" (2007) [88], during this decade, the methods of crowd counting can be divided into detection-based methods and regression-based methods. For the detection-based approach, a framework will be used to detect the people in images. It collects information from the input images by sliding boxes to count the number of people [33, 34]. The regression-based method aims to find out a mapping of features to the number of people [35, 36]. Regression-based methods usually use foreground feature and edge feature extraction to construct the mapping between features and the number of people, such as standard background subtraction techniques to get foreground features [36, 39].

With the development of artificial intelligence technology, crowd counting is no longer satisfied with counting the number of people in simple scenarios. More and more researchers are focusing on large-scale high-density crowd counting in complex situations. For this research direction, the main difficulty lies in the mutual occlusion of dense crowds and the

TABLE 3 A list of crowd datasets.

		Avg. Resolution		Count statistics						Scale of dataset
Dataset	**Number**	**H**	**W**	**Total**	**Min**	**Average**	**Max**	**APO**	**Years**	
Florence [62]	3358	-	-	17630	0	5.3	28	-	2014	Small
SmartCity [65]	50	1080	1920	369	1	7.4	14	529.4	2018	Small
Indoor2 [67]	148243	576	740	1834770	0	12.4	40	185.4	2019	Small
Beijing-BRT [66]	1280	640	360	16795	1	13.1	64	133.1	2018	Small
PETS [61]	1076	288	384	18289	0	16.9	40	80.9	2010	Small
UCSD [1]	2000	158	238	49885	11	24.9	46	38.9	2008	Small
FDST [68]	15000	1080	1920	394081	-	26.3	-	277.1	2019	Small
Crowd surveillance [50]	13945	840	1342	386513	-	27.7	-	200.6	2019	Small
MALL [36]	2000	480	640	62325	13	31.2	53	99.2	2012	Small
Train station [63]	2000	256	256	62581	1	31.3	53	45.8	2017	Small
CityUHK-X [64]	3191	384	512	106783	-	33.5	-	77.2	2017	Small
VSCrowd [86]	62938	1080	1920	2344276	-	37	-	236.7	2022	Small
ZhengzhouAirport [69]	1111	-	-	49061	7	44.2	128	-	2019	Large
City street [74]	500	1520	2074	-	70	-	150	-	2019	Large
WorldExpo'10 [70]	3980	576	720	199923	1	50.2	253	90.9	2015	Large
Fine-grained crowd [75]	3728	158	238	112344	2	57	344	-	2021	Large
ShanghaiTechRGBD [71]	2193	1080	1920	144512	6	65.9	234	177.4	2019	Large
ShanghaiTech part B [80]	716	768	1024	88488	9	123.6	578	79.8	2016	Large
DroneCrowd [73]	33600	1080	1920	4864280	25	144.8	455	119.7	2019	Large
JHU-CROWD [76]	4250	900	1450	1114785	-	262.3	7286	70.6	2019	Hyper
GTA head	5098	1080	1920	1732505	-	340	-	78.1	2022	Hyper
JHU-CROWD++ [77]	4372	910	1430	1515005	-	346.5	25791	61.3	2020	Hyper
NWPU-crowd [56]	5109	2311	3383	2133238	0	418.5	20033	136.8	2020	Hyper
Crowd-saliency [79]	107	576	720	45000	58	421.6	2201	31.4	2016	Hyper
CrowdX [82]	24000	768	1024	2400	-	500	-	39.7	2022	Hyper
ShanghaiTech part A [80]	482	589	868	241677	33	501.3	3139	31.9	2016	Hyper
GCC [81]	15211	1080	1920	7625843	0	501.4	3995	64.3	2019	Hyper
UCF-QNRF [83]	1535	2013	2902	1251642	49	815.4	12865	84.7	2018	Hyper
UCF_CC_50 [84]	50	2888	2101	63974	94	1280.5	4543	68.9	2013	Hyper
DLR-ACD [85]	33	-	-	226291	285	6857.3	24368	-	2019	Hyper

Note: In the table, H, W, Total, Min, Average, Max, and APO represent the average height and width of the picture, the total number, the minimum number, average number, maximum number, and average width occupied by each object of instances in the datasets, respectively. "-" means that the data is not mentioned in the original paper, resulting in the lack of data and unable to calculate. The table is sorted from least to most according to the average number of annotations per image.

complex background environment, which increases the counting error of detection-based crowd counting methods. On the other hand, the traditional regression-based crowd counting method does not show the location of each target as clearly as the detection method. This drawback has led to certain doubts about the credibility of the regression model. To solve this problem, researchers proposed a method based on density estimation. It will transfer the image to a density map and then use a density map to estimate the number of people [43, 89, 90]. This method can reflect the distribution position of the crowd in the image and can be better used in real life.

After the extensive application of the convolutional neural network structure in recent years, researchers also began to use the convolutional neural networks in the field of crowd counting. In 2015, CNNs were first used in crowd counting. Wang et al. and Fu et al. used basic CNNs to estimate the density maps [44, 45]. Their research has inspired other scholars to start upgrading CNNs for crowd counting. From the beginning, the CNN network showed excellent learning ability. The use of CNN is efficient and fast, but researchers still need to transform the network for actual use scenarios. To improve the capabilities of the network, researchers have made various improvements. For example, the Cross Scene Crowd Counting [70] gave a new method to get density maps and let CNNs perform perspective normalisation. Also, the 'End to End Count Estimate Method' allowed the input to use the whole image rather than a patch of the image. It reduced the complication of programming based on CNNs [91].

Some researchers also made achievements by upgrading the network structure. For example, a deep network is used in combination with a shallow network to analyse images [46]. Zhang [80] proposed the multi-column convolutional neural network (MCNN) to combine different size filters [80]. This model combines several independent convolutional neural networks in parallel to accommodate the different sizes of people in the image. Sam et al., in 2017, proposed an upgraded network Switch-CNN based on MCNN [92]. It used a switching layer to select each patch of image and choose a suitable network to analyse it. Then a Mixture of CNNs (MoCNN), which can better adapt to changes in different background environments, was proposed in 2017 [48]. Additionally, to improve the CNNs performance in high-density crowd image research, a cascaded Multi-task CNN model has been proposed by Sindagi [93].

Data have largely driven improvements in these methods. Therefore, from the perspective of data-driven, we divide crowd counting algorithms into the following six categories:

(1) Multi-scale networks: These networks use multiple scales of input images to capture different levels of detail in the crowd. This can improve accuracy when the crowd is dense or when individuals are close together.
(2) Single-column networks: These networks process the input image with a single column of convolutional layers. They are often used when the data is relatively simple and straightforward, and when the goal is to achieve fast and efficient processing.
(3) Multi-column networks: These networks use multiple columns of convolutional layers to process different regions of the input image simultaneously. This can improve accuracy when the crowd is highly varied in density or when there are significant variations in illumination or other environmental factors.
(4) Multi-task networks: These networks use multiple loss functions to simultaneously optimise for different tasks, such as counting people and estimating their positions. This can improve accuracy and reduce overfitting, especially when there are limited amounts of training data, while this may need more complex data preprocessing.
(5) Attention networks: These networks use attention mechanisms to selectively focus on important parts of the input image. Some methods will need a specific attention area for training, like a perspective map. This can improve accuracy when the crowd is highly varied, and the key features for counting are not easily distinguished.
(6) Weak-supervised networks: These networks use weakly-supervised learning techniques, such as using only image-level labels or partial annotations, to learn from limited amounts of labelled data. This can improve accuracy when labelled data is scarce or difficult to obtain.

3.1 | Multi-scale networks

Multi-scale networks refer to neural networks that are designed to process information at multiple scales. This means that the network can process and analyse the same input data at different levels of granularity, such as at different levels of resolution or at different scales of size. This capability allows multi-scale networks to capture more detailed and diverse features from the input data and thus to improve their performance in tasks such as image classification, object detection, and crowd counting.

The usual crowd counting model can only deal with the learned crowd scenes, and it is often difficult to deal with new scenes. Zhang et al. proposed a cross-scene crowd counting method in 2015 to count people in new crowd scenarios [70]. This method eliminates the need to label and relearn new crowd scenes, which can greatly save network deployment costs. The new model proposed by this method is mainly trained on the two tasks of crowd density estimation and crowd counting. Through the simultaneous learning of these two tasks, the model can obtain a better local optimum to adapt to both crowd counting and crowd density estimation. Since it is very difficult to segment the foreground of the crowd completely, causing some immobile people to be removed by mistake, this model does not use foreground segmentation as the basis of training. Therefore, the model is data-driven, and Zhang's team also created a new dataset including 108 scenes and 200000 head annotations for this purpose.

This method is shown in Figure 6. The whole model is divided into two parts. The first part is to train a crowd density estimation network, and the second part is to adjust the parameters of the pre-trained crowd density estimation network based on the distribution characteristics of the crowd in the dataset. In the first part, the input image is image blocks of different sizes cropped from the training image according to the perspective value of the block centre pixel. After scaling these image blocks to the same size, the person in the image can maintain the same size. These images are then used for the training of the CNN network. Participating in training along with these images is the density map as the global truth. Each label on the density map contains two parts, the head and the body. As formula (2) shows, the density map D is generated by two parts: head and body. The pedestrian head P_h is generated using a standardised 2-dimensional Gaussian kernel N_h, and the body P_b is generated using a bivariate normal distribution N_b. The second part is the switchable training process of this method, which can alternately optimise the effects of crowd density map estimation and crowd counting. This method will collect the loss of the density map and the loss of the global number. The network mainly learns to estimate the density of input patches, supplemented by the estimation of the global number. At the beginning of training, the loss convergence of the density map is prioritised, and when the loss of the density map converges, it is switched to minimise the loss of global number.

$$D_i(p) = \sum_{P\in P_i}\frac{1}{\|Z\|}\left(N_h(p;P_h,\sigma_h) + N_b(p;P_b,\sigma_b)\right), \qquad (2)$$

where $\sigma_h = 0.2M(p)$ for the term N_h, and $\sigma_x = 0.2M(p)$, $\sigma_y = 0.5M(p)$ for the term N_b.

In view of the difference between the test scene and the training scene, this method proposes non-parametric fine-tuning of the target scene. This method uses candidate scene retrieval and partial patch retrieval to add images similar to the salient features in the test scene into the training set so that the neural network extraction adapts to the possible situations of the test scene.

3.2 | Single-column networks

Single-column networks, also known as single-stream networks, are neural networks that use a single column of neurons to process input data. Unlike multi-column networks, which use multiple columns to process the data, single-column networks process the input data using only one set of neurons. Single-column networks are simpler in architecture and require fewer computations, making them faster and more efficient to train and run.

Li et al. proposed the CSRnet for congested scene recognition (Li, Zhang, and Chen). This method constructs a single column network to estimate the crowd density. This model has achieved amazing results on the ShanghaiTech Part B dataset, and the team also applied this model to tasks such as vehicle detection.

This method used the density map generation method based on the geometrically adapted Gaussian kernel [80], which is shown in Equation (4).

The network structure is shown in Figure 7. This network is front-end by a fine-tuning network based on VGG-16, which has good feature extraction capabilities. This method also uses a data-driven approach, but it pioneered the use of dilated CNN as a back-end network to increase feature acquisition while avoiding the loss of spatial information caused by the use of pooling layers. The formula of dilated CNN is shown in Equation (3).

$$y(m,n) = \sum_{i=1}^{M}\sum_{j=1}^{N} x(m+r\times i, n+r\times j)w(i,j) \qquad (3)$$

In which $x(m,n)$ is the input with length and width of M and N, respectively, and the output $y(m,n)$ is obtained through the convolution kernel $w(i,j)$, where the parameter r represents the dilation rate. If $r = 1$, the dilated convolution is an ordinary convolution layer. In the study, the dilated rate is set as $r = 2$.

The use of the deconvolution layer will increase the complexity of the network, but the use of dilated convolution can well control the number of parameters and the amount of calculation. In general, this method is based on the idea of encoding-decoding and innovatively improves the VGG16 network. The ten convolutional layers in the front-end network come from the already trained VGG-16, so only fine-tuning training is required. The dilated convolution is used at the end of the VGG-16 network to expand the receiving domain without reducing the resolution.

In general, CSRNet is a single-column convolutional neural network that can use end-to-end methods to generate density maps. This method has become one of the mainstream methods currently studied.

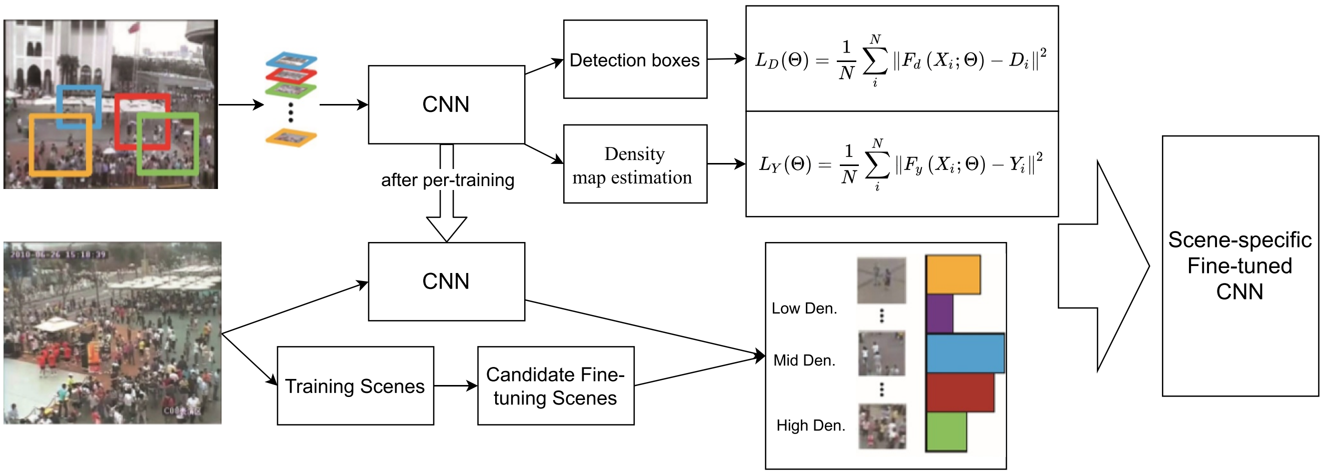
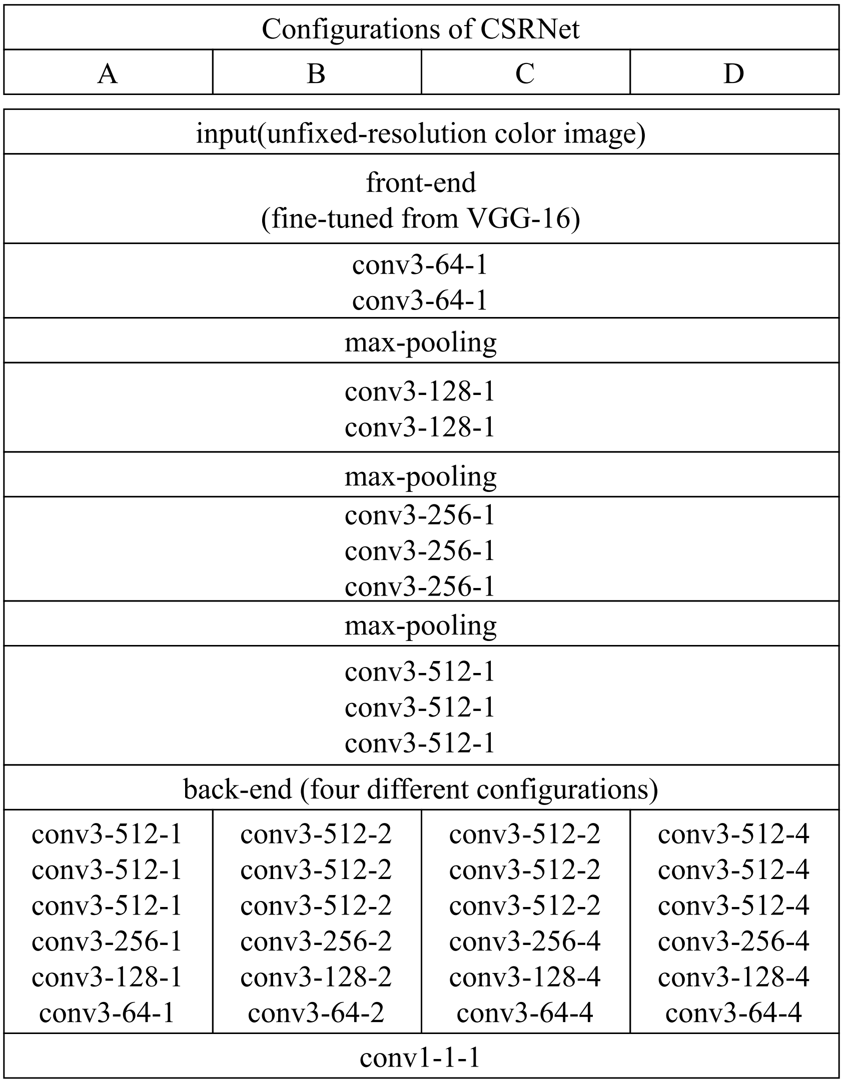

CNN
Detection boxes
Density map estimation
$L_D(\Theta) = \frac{1}{N}\sum_i^N \|F_d(X_i;\Theta) - D_i\|^2$
$L_Y(\Theta) = \frac{1}{N}\sum_i^N \|F_y(X_i;\Theta) - Y_i\|^2$
after per-training
CNN
Training Scenes
Candidate Fine-tuning Scenes
Low Den.
Mid Den.
High Den.
Scene-specific Fine-tuned CNN

FIGURE 6 Illustration of the cross-scene crowd counting method.



Configurations of CSRNet			
A	B	C	D
input(unfixed-resolution color image)			
front-end (fine-tuned from VGG-16)			
conv3-64-1 conv3-64-1			
max-pooling			
conv3-128-1 conv3-128-1			
max-pooling			
conv3-256-1 conv3-256-1 conv3-256-1			
max-pooling			
conv3-512-1 conv3-512-1 conv3-512-1			
back-end (four different configurations)			
conv3-512-1 conv3-512-1 conv3-512-1 conv3-256-1 conv3-128-1 conv3-64-1	conv3-512-2 conv3-512-2 conv3-512-2 conv3-256-2 conv3-128-2 conv3-64-2	conv3-512-2 conv3-512-2 conv3-512-2 conv3-256-4 conv3-128-4 conv3-64-4	conv3-512-4 conv3-512-4 conv3-512-4 conv3-256-4 conv3-128-4 conv3-64-4
conv1-1-1			

FIGURE 7 Configuration of CSRNet (Y. Li et al.).

3.3 | Multi-column networks

Multi-column networks, also known as multi-stream networks, are neural networks that use multiple columns of neurons to process input data. Each column is designed to process information at a different scale or level of abstraction, allowing the network to capture more diverse and detailed features from the input data. This makes multi-column networks more effective in tasks that require processing high-dimensional and multi-scale information. Multi-column networks are generally more complex in architecture and require more computations than single-column networks, but they are also more powerful in terms of performance.

Faced with the large perspective zoom phenomenon of crowd images, Zhang et al. proposed a multi-column convolutional neural network (MCNN) [80]. This network can tolerate input images of any size, instead of using fixed-size input images like VGG16. At the same time, this research puts forward an adaptive Gaussian convolution kernel to obtain an accurate crowd density map for the problem that the perspective view of the crowd image is difficult to obtain. At the same time, the team collected and sorted out a new crowd dataset, which is the ShanghaiTech dataset that is now widely used.

For the input image, MCNN randomly crops each original image nine times to obtain nine training images, and each training image is 1/4 the size of the original image. In this way, the input image size becomes smaller, speeding up network training. Moreover, the training of the complete image is completed through the training of the partial image block, which can also achieve good results.

The network structure is shown in Figure 8. Due to high-density crowd images usually having a serious perspective, in early research, multi-scale image processing methods are usually used to deal with people of different sizes[44, 45, 70]. The multi-scale method maintains the same size of the person and crops the image into patches of different sizes. The multi-column network uses different receptive fields of the different size convolution kernel in different columns so that the sub-networks in different columns can learn head features of different sizes. The three-column sub-network contains three different sizes of convolution kernels: large, medium, and

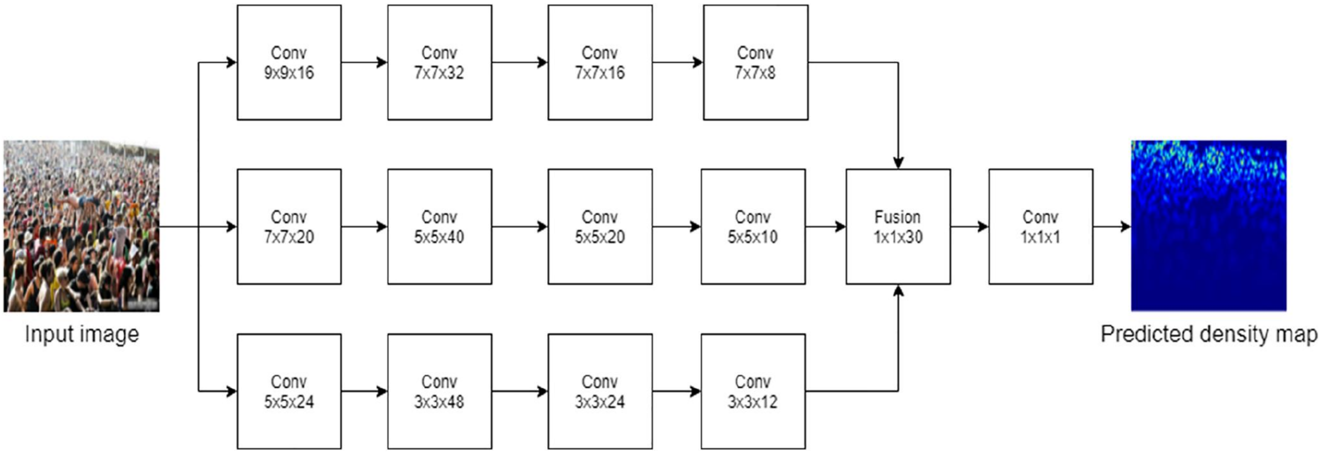


FIGURE 8 The structure of the proposed multi-column convolutional neural network for crowd density map estimation.

small, corresponding to people at three distances: near, middle, and far. Finally, the 1 × 1 convolution layer is used to merge the density maps generated by the three subnets together. The size of each convolution kernel is carefully designed to ensure that the network can accept input images of any size.

In training, at first, MCNN needs to pre-train the three single-column neural networks and combine the three pre-trained networks as a multi-column network, then train this network.

At the same time, Zhang et al also contributed to the ShanghaiTech dataset. To facilitate the use of the dataset, Zhang et al. proposed an adaptive Gaussian kernel to obtain an accurate crowd density map. The adaptive Gaussian kernel automatically obtains the Gaussian kernel size according to the distance of the people around a person, which is shown in Equation (4). The crowd density map generated in this way can automatically adapt to the distribution of the crowd in the image, thereby obtaining good accuracy.

$$F(x) = \sum_{i=1}^{N} \delta(x - x_i) \times G_{\delta_i}(x), with\ \delta_i = \beta\overline{d_i} \qquad (4)$$

In which δ is the ground truth of the input x_i, $\overline{d_i}$ is the average distance of k nearest neighbours. $\delta(x - x_i)$ will be convolved by a Gaussian kernel with the parameter δ_i, and x is the position of pixel in the image. In the study, $\beta = 0.3$ shows the best performance in the result.

MCNN is the first time that a multi-column convolutional neural network has been proposed. At the same time, the ShanghaiTech dataset released by Zhang et al. is also one of the most commonly used crowd datasets.

3.4 | Multi-task networks

Multi-task networks are neural networks that are trained to perform multiple tasks simultaneously. Unlike traditional single-task networks, which are trained to perform only one task, multi-task networks can process input data and produce outputs for multiple tasks. This allows them to share information and learn common features that are relevant to multiple tasks, making them more efficient in terms of training time and resource utilisation. Multiple tasks are related and can benefit from shared knowledge.

In reality, the crowd is not always in a state of high density. Considering the coexistence of high-density and low-density people, Jiang et al. noted that the crowd counting method based on detection could better complete the counting work of low-density crowds, but this method has great disadvantages for the high-density crowd. While the crowd counting method based on the density map can complete the estimation of the high-density crowd well, it has big errors in the face of the low-density crowd. They proposed DecideNet [94] to combine the advantages of these two methods. DecideNet can simultaneously complete the two tasks of crowd target detection and crowd density counting. Moreover, it will adaptively select the counting method of target detection or the counting method of density estimation according to the distribution of people on the image. When the crowd density is high, choose more counting methods for density estimation, and when the crowd density is low, choose more counting methods for target detection.

Figure 9 shows the architecture of DecideNet. In this model, the input image will be counted based on target detection (DetNet) and count based on density estimation (RegNet).

RegNet is a fully convolutional network that uses regression to obtain density maps. The formula is shown in Equation (5) in which the F^{reg} is the crowd density obtained by a fully convolutional network, I_i is the input image D_i^{reg} is the ground truth, p is the pixel of I_i and Ω is the weight of RegNet. This network uses a larger convolution kernel (7 × 7, 5 × 5) to obtain more environment information.

$$F^{reg}\left(I_i \middle| \Omega_{reg}\right) = D_i^{reg}\left(p \middle| \Omega_{reg}, I_i\right) \qquad (5)$$

DetNet is a Fast-RCNN head detection network based on ResNet-101 with a Gaussian convolutional layer for output at the end. The formula is shown in (6) in which D_i^{det} is the

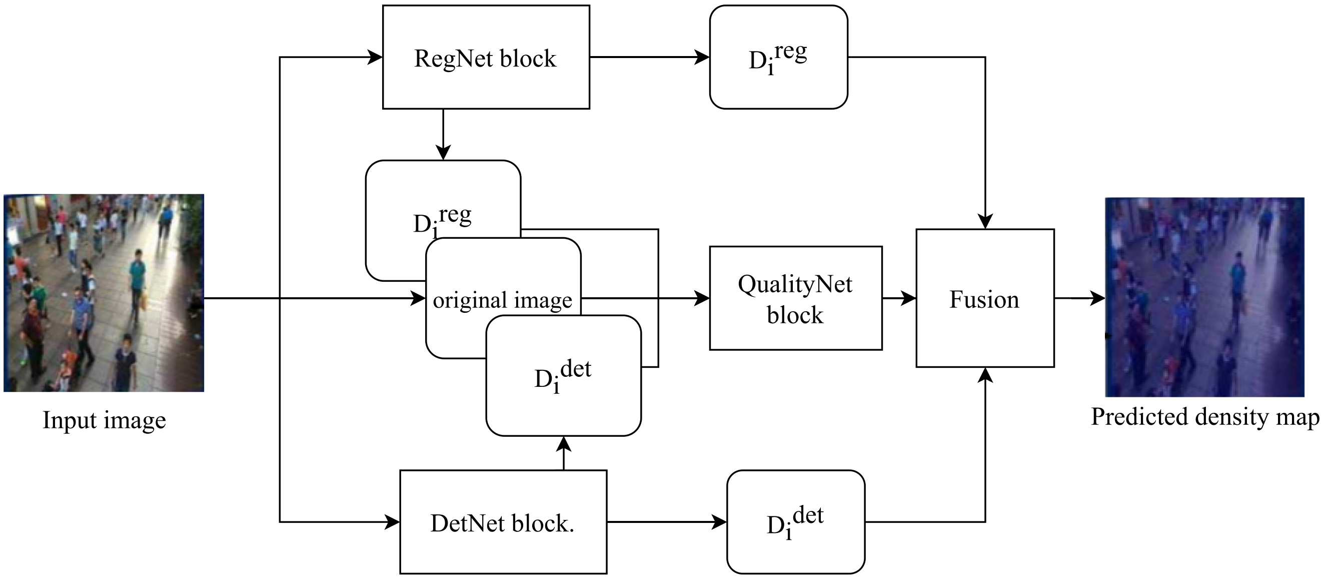


FIGURE 9 The architecture of DecideNet.

detection-based density map, P_i^{det} is the detection output, and $N^{det}(p|\mu = P, \sigma^2)$ is the constant Gaussian function.

$$D_i^{det}(p|\Omega_{det}, I_i) = \sum_{P \in P_i^{det}} N^{det}(p|\mu = P, \sigma^2) \tag{6}$$

At the same time, the model uses a QualityNet to determine whether the crowd density of the input image is of high density or low density to determine the respective weighting parameters of the regression output and the detection output. This network dynamically evaluates the weights of the two density maps through the quality of each pixel. The formula is shown in (7) in which D_i^{final} is the weighted sum of the density map from DetNet and RegNet, and K_i.is the attention map; ⊙ is the Hadamard product for two matrices, and the J is an all-one-matrix with the same size of K_i.

In training, to increase the number of images and improve the robustness of the model, the training images are cut into 43 patches, and a uniform noise of [−5,5] is randomly added to each patch with a probability of 50%.

DecideNet is the first framework to estimate the number of people adaptively by using detection or regression-based methods through the attention mechanism.

$$\begin{aligned} D_i^{final}(p|I_i) = {} & K_i\left(p\middle|\Omega_{qua}, I_i\right) \odot D_i^{det}(p|\Omega_{det}, I_i) \\ & + \left(\mathrm{J} - K_i\left(p\middle|\Omega_{qua}, I_i\right)\right) \odot D_i^{reg}\left(p\middle|\Omega_{reg}, I_i\right) \end{aligned} \tag{7}$$

3.5 | Attention networks

Attention networks are neural networks that use an attention mechanism to selectively focus on certain parts of the input data and weigh their importance in making predictions. The attention mechanism allows the network to dynamically attend to the most relevant features and information in the input and to automatically adjust the importance of different features based on their relevance to the task. Attention networks have proven to be effective in improving performance and reducing computational complexity compared to traditional feedforward neural networks.

To solve the challenge of large changes in the density of crowd images, Hossain et al. proposed an attention-based method: SAAN [95]. The commonly used crowd counting method is to use the neural network to generate density and then estimate the number of people through the density map. This model can learn the area of interest in the image from both the global and local directions. The attention mechanism is to simulate the human observation of objects. Generally, when people observe an object, they do not scan an object but focus on observing various parts of the object. The attention mechanism is not to extract features from the entire image but to focus the model on the most needed features.

The structure of SAAN is shown in Figure 10. The network has three main parts: Multi-scale Feature Extractor (MFE), Global Scale Attention (GSA), and Local Scale Attention (LSA). MFE is a module for multi-scale feature extraction of images based on multi-column convolutional networks. It acquires image features from three scales: large, medium, and small. GSA is to obtain the information of the entire image and then score the crowd density of the image.

Moreover, it uses VGG16 as the basic network. GSA will also judge whether this crowd is high-density, medium-density, or low-density. Since the density distribution of each picture is not uniform, it is necessary to use LSA to determine the local density level of different areas in the picture. LSA will generate three pixel-level attention maps to

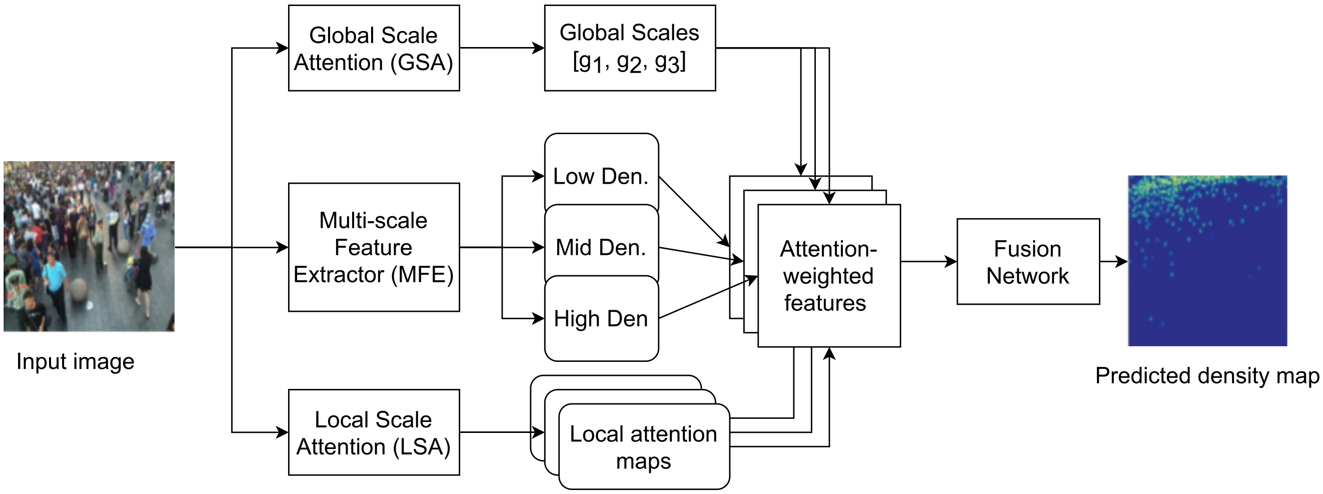

Input image
Global Scale Attention (GSA)
Global Scales [g1, g2, g3]
Multi-scale Feature Extractor (MFE)
Low Den.
Mid Den.
High Den
Attention-weighted features
Fusion Network
Predicted density map
Local Scale Attention (LSA)
Local attention maps

FIGURE 10 The architecture of SAAN.

label the attention levels of different regions in the image. Finally, the fusion network multiplies the density map and density score obtained by the three modules to obtain a final crowd density map. The Fusion network has two deconvolutional layers to extend the feature map to the same size as the input image. This feature map will be flattened by the 1 × 1 Conv layer to generate the final 2D density map (DM), and the pixel value sum of the DM is the predicted number of people, which is shown as follow:

$$Count = \sum_{j=1}^{H} \sum_{i=1}^{W} DM(j,k) \tag{8}$$

To better control the network's learning, the loss of the model uses a compound loss as shown in Equation (9).

$$L_{final} = L_{DM} + \lambda_g L_{GSA} + \lambda_l L_{LSA} \tag{9}$$

L_{DM} is the square of the Frobenius norm between the predicted density map C and the ground truth C_{gt}:

$$L_{DM} = \frac{1}{2} \left\| vec(C) - vec\left(C_{gt}\right) \right\| \tag{10}$$

L_{GSA} is the standard cross-entropy loss:

$$L_{GSA} = CE\left(g, g_{gt}\right), \tag{11}$$

where $CE()$ means the multi-class cross-entropy loss function.

L_{LSA} is the sum of cross-entropy losses overall spatial locations:

$$L_{LSA} = \sum_{h=1}^{H} \sum_{w=1}^{W} CE\left(l[h,w,:], l_{gt}[h,w]\right) \tag{12}$$

Both SAAN and DecideNet predict the crowd density of the image to improve the accuracy of crowd density estimation. However, SAAN pays more attention to the use of the attention mechanism, which makes the structure of SAAN simpler.

3.6 | Weak-supervised networks

Weakly supervised networks are neural networks that are trained using weak supervision signals, such as image-level labels or tags, instead of densely annotated data. Weak supervision signals provide limited information about the input data and require the network to learn from indirect or partial signals. This makes weakly-supervised networks more efficient and cost-effective to train compared to fully-supervised networks, which require dense annotations for every instance in the training data. Weakly supervised networks have been widely used in various applications, such as object detection, segmentation, and classification, where obtaining dense annotations is challenging or expensive. Despite the limitations of weak supervision signals, weakly supervised networks have proven to be effective in learning high-level representations from the input data and achieving competitive performance compared to fully supervised networks.

The task of crowd counting is still limited by crowd data. The existing labelled data is less and single type. Moreover, collecting data specifically for a specific task and labelling such data requires much effort. Therefore, from the perspective of big data theory, the study of crowd counting requires a model that can use unlabelled data to increase the total amount of data. Sam et al. proposed a semi-supervised method for crowd counting [96]. This model mainly relies on unlabelled data for training. The network structure of the model is shown in Figure 11. The GWTA-CCNN is a single-column 6-layer convolutional neural network. The first four layers use the GWTA method for training, and the last two layers use the supervised method for training. GWTA is essentially a training method based on an auto-encoder-decoder. The auto-encoder

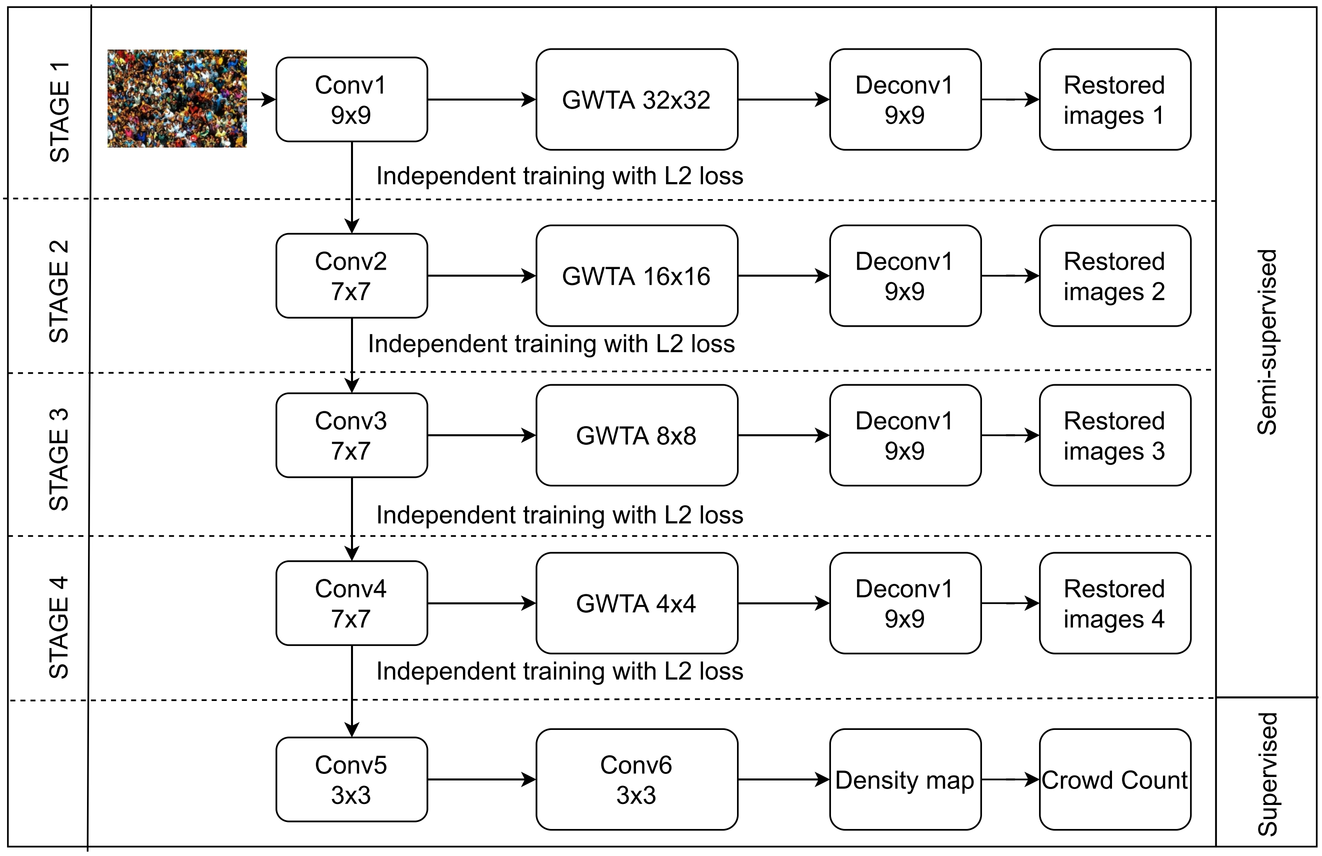


FIGURE 11 The architecture of GWTA-CCNN.

extracts features from the input image and then the decoder restores the input image based on the extracted features, then train the network model by comparing the feature gap before and after restoration. WTA is Winner Takes All, where only the maximum value is kept, and the remaining values are zero. The original WTA is performed on the entire feature map, while GWTA is a grid of the feature map, and WTA operations are performed on each grid.

Each encoder will be trained separately. When training Conv1, the image is first input into Conv1, and the output features are processed by GWTA, and then the restored image is output through a deconvolution module. Then, calculate the loss between the restored image and the original input image of Conv1, which is shown in Equation (13). Then train Conv2 individually, and the features obtained by Conv1 are pooled as the input of the second layer Conv2.

$$L_{l_2}^{D} = \frac{1}{2N}\sum_{i=1}^{N}\left\| D_{x_i}(x;\Theta_s) - D_{x_i}^{GT}(x)\right\|_2^2 \tag{13}$$

When training the last two layers of networks, the parameters of the first four layers of networks are fixed, just like transfer learning. In this way, the network parameters that need to be trained with annotated data are greatly reduced. Conversely, the need for annotated data for network training is not as large as training the entire network.

3.7 | Quantified evaluation

In the past 5 years, the field of crowd counting has developed particularly rapidly. We have collected most of the methods (114 methods in total) in crowd counting fields and their estimated accuracies on seven popular datasets from 2019 to early 2023 in Table 4. Mean absolute error (MAE) and root mean squared error (RMSE), which represent error counts, are used to evaluate the accuracy of crowd estimation. In order to better reflect the impact of the dataset on each method, we have divided the table according to the scale of the dataset. From left to right, the three main columns are small-scale datasets, large-scale datasets, and hyper-scale datasets, respectively. The methods are arranged according to the year of publication. We bolded the top three results with the lowest error count on each dataset, and the top one result on each dataset is underlined.

$$MAE = \frac{1}{N}\sum_{i=1}^{N}\left|C_{gt_i} - C_{est_i}\right| \tag{14}$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left|C_{gt_i} - C_{est_i}\right|^2} \tag{15}$$

TABLE 4 Compare the performance of different methods on the popular crowd counting dataset.

	UCSD		Mall		WorldExpo'10		SHT B		SHT A		UCF-QNRF		UCF_CC_50		
Method	**MAE**	**RMSE**	**MAE**	**RMSE**	**MAE**	**RMSE**	**MAE**	**RMSE**	**MAE**	**RMSE**	**MAE**	**RMSE**	**MAE**	**RMSE**	**Year**
Cross-scene [70]	1.6	3.31	-	-	12.9	-	32	49.8	181.8	277.7	-	-	467	498.5	2015
MCNN [80]	1.07	1.35	2.24	8.5	11.6	-	26.4	41.3	110.2	173.2	-	-	377.6	509.1	2016
DecideNe t[94]	-	-	1.52	1.9	9.23	-	21.53	31.98	-	-	-	-	-	-	2018
CSRNet [97]	1.16	1.47	-	-	8.6	-	10.6	16	68.2	115	120.3	208.5	266.1	397.5	2018
Stacked-Pool [98]	-	-	-	-	12.92	-	18.73	31.86	93.98	150.59	-	-	-	-	2018
SAAN (M. [95])	-	-	1.28	1.68	-	-	16.86	28.41	-	-	-	-	271.6	391	2019
SWTA-CCNN [96]	-	-	-	-	-	-	-	-	154.7	229.4	-	-	433.7	583.3	2019
IkNN [99]	-	-	-	-	-	-	13.4	21.4	68	117.7	104	172	237.76	305.7	2019
TAN [100]	1.08	1.41	2.03	2.6	8.3	-	15.1	23.3	93.3	157	-	-	262	358.6	2019
MobileCount [101]	-	-	-	-	12.2	-	9.1	15.1	98.6	162.9	137.8	238.2	321.7	437.1	2019
FADA [102]	2	2.43	2.47	3.25	21.6	-	16	24.7	-	-	-	-	-	-	2019
DUBNet [103]	-	-	-	-	-	-	7.7	12.5	64.6	106.8	105.6	180.5	243.8	329.3	2019
DENet [104]	1.05	1.31	-	-	8.2	-	9.6	15.4	65.5	101.2	-	-	241.9	345.4	2019
SD-CNN [105]	1.01	1.28	-	-	7.04	-	-	-	-	-	-	-	235.74	345.6	2019
LSC-CNN [106]	-	-	-	-	-	-	8.1	12.7	66.4	117	-	-	225.6	302.7	2019
EPF [107]	0.81	1.07	-	-	6.6	-	-	-	-	-	-	-	-	-	2019
DeepCount [108]	-	-	1.55	2	-	-	7.2	11.3	65.2	112.5	95.7	167.1	-	-	2020
RRP [109]	-	-	-	-	-	-	9.4	13.9	63.2	105.7	93	156	216.3	316.6	2020
CAT-CNN [110]	-	-	-	-	7.2	-	11.2	20	66.7	101.7	-	-	235.5	324.8	2020
ASDF [111]	1.15	1.44	1.5	1.91	7.1	-	8.5	13.7	65.6	98	-	-	196.2	270.9	2020
SRN + PS [112]	1.07	1.35	-	-	-	-	13.8	18.8	75	115.2	-	-	289.7	384.2	2020
MLSTN [51]	1.02	1.32	1.8	2.42	-	-	-	-	-	-	-	-	-	-	2020
FMLF [113]	-	-	1.85	2.34	8.6	-	10.2	14.9	69.8	114.7	-	-	271.3	376.3	2020
DensityCNN [114]	-	-	-	-	6.88	-	9.12	16.34	63.06	106.34	101.52	186.9	244.57	341.76	2020
Deem [115]	-	-	2.1	2.66	8.34	-	8.09	12.98	-	-	-	-	253.4	364.4	2020
ZoomCount [116]	-	-	-	-	8.7	-	-	-	66.6	94.5	130	204	-	-	2020
DCL [117]	-	-	-	-	11	-	-	-	64.97	107.96	108	182	-	-	2020
MS-GAN [118]	1.78	3.03	1.27	2.55	-	-	18.7	30.5	-	-	-	-	345.7	418.3	2020
BNFDD [119]	-	-	-	-	-	-	6.64	10.93	58.99	106.99	97.58	198.79	174.28	240.86	2020
CRNet [120]	-	-	-	-	7.1	-	7.4	11.9	56.4	90.4	101	162	203.3	263.4	2020
PaDNet [121]	0.85	1.06	-	-	-	-	8.1	12.2	59.2	98.1	96.5	170.2	185.8	278.3	2020
HA-CCN [122]	-	-	-	-	-	-	8.1	13.4	62.9	94.9 8	118.1	180.4	256.2	348.4	2020
CTN [123]	-	-	-	-	-	-	7.5	11.9	61.5	103.4	86	146	210	305.4	2020
CC-2P [124]	-	-	-	-	-	-	-	-	67.8	86.2	94.5	141.9	-	-	2020
FSSA [125]	3.08	4.16	2.44	3.12	7.12	9.88	-	-	-	-	-	-	-	-	2020
SDANet [126]	-	-	-	-	8.1	12.9	7.8	10.2	63.6	101.8	-	-	227.6	316.4	2020
HyGnn [127]	-	-	-	-	-	-	7.5	12.7	60.2	94.5	100.8	185.3	184.4	270.1	2020
C-CNN [128]	-	-	-	-	9.9	-	14.9	22.1	88.1	141.7	-	-	-	-	2020
FSC [129]	-	-	-	-	-	-	16.9	24.7	129.3	187.6	221.2	390.2	-	-	2020

(Continues)

TABLE 4 (Continued)

	UCSD		Mall		WorldExpo'10		SHT B		SHT A		UCF-QNRF		UCF_CC_50		
Method	**MAE**	**RMSE**	**MAE**	**RMSE**	**MAE**	**RMSE**	**MAE**	**RMSE**	**MAE**	**RMSE**	**MAE**	**RMSE**	**MAE**	**RMSE**	**Year**
MSPNET [130]	-	-	-	-	-	-	7.5	14.1	59.8	98.2	206.7	299.3	-	-	2020
SMANet [131]	-	-	-	-	-	-	7.3	12.9	59.7	102.1	92.5	176.7	178.4	256.3	2020
BBA-NET [132]	-	-	-	-	-	-	7.8	12	63.8	93.8	-	-	230.5	316.9	2020
AGRD [133]	-	-	-	-	-	-	7.2	11.8	61.4	97.5	-	-	194.7	246.8	2020
CWAN [134]	-	-	2.06	2.9	7.17	-	-	-	-	-	-	-	-	-	2020
SOFA-Net [135]	-	-	-	-	-	-	6.8	10.38	57.5	92.12	96.2	158.7	185	281	2020
SRF-Net [136]	-	-	-	-	-	-	7.1	11.5	60.4	97.2	98	170	197.3	271.8	2020
HSRNet [118]	1.03	1.32	1.8	2.28	7.44	-	7.2	11.8	62.3	100.3	-	-	-	-	2020
ASNet [114]	-	-	-	-	6.64	-	-	-	57.78	90.13	91.59	159.71	174.84	251.63	2020
RPNet [137]	1.32	1.23	-	-	8.2	-	8.1	11.6	61.2	96.9	120.5	218.2	-	-	2020
ADSCNet [138]	-	-	-	-	-	-	6.4	11.3	55.4	97.7	71.3	132.5	198.4	267.3	2020
FOCNN [139]	-	-	1.22	2.54	-	-	-	-	-	-	-	-	-	-	2020
CCLS [140]	1.8	2.8	-	-	9.6	-	12.3	21.2	104.6	145.2	-	-	-	-	2020
PSSW [141]	-	-	3.8	5.4	-	-	14.4	17.9	84.4	93.8	-	-	318.7	421.6	2020
IRAST [142]	-	-	-	-	11.1	-	14.7	22.9	86.9	148.9	135.6	233.4	-	-	2020
GP [143]	2	2.4	-	-	12.8	-	15.7	27.9	102	172	160	275	-	-	2020
LibraNet [144]	-	-	-	-	-	-	7.3	11.3	55.9	97.1	88.1	143.7	181.2	262.2	2020
AMRNet [145]	-	-	-	-	-	-	7.02	11	61.59	98.36	86.6	152.2	184	265.8	2020
NAS-count [146]	-	-	-	-	6.8	-	6.7	10.2	56.7	93.4	101.8	163.2	208.4	297.3	2020
RDBT [147]	-	-	-	-	-	-	13.38	29.25	112.24	218.18	175.02	294.76	368.01	518.92	2020
PWCU [148]	-	-	-	-	9.4	-	7.6	13	64.8	107.5	102	171.4	214.2	318.2	2020
SKT [94]	-	-	-	-	7.34	-	7.48	11.68	71.55	114.4	96.24	156.82	-	-	2020
KDMG [149]	-	-	-	-	-	-	7.8	12.7	63.8	99.2	99.5	173	-	-	2020
MNA [150]	1	1.35	-	-	-	-	7.4	11.3	61.9	99.6	85.8	150.6	-	-	2020
DM-count [151]	-	-	-	-	-	-	7.4	11.8	59.7	95.7	85.6	148.3	211	291.5	2020
JHU-CROWD++ [77]	-	-	-	-	-	-	7.5	12.1	60.2	94	95.5	164.3	-	-	2020
M-SFANet [152]	-	-	-	-	7.32	-	6.32	10.06	57.55	94.48	87.64	147.78	167.51	256.26	2020
MH-MetroNet [153]	-	-	-	-	-	-	7.93	13	67.52	113.47	-	-	170	221.95	2020
AdaCrowd [154]	-	-	4	5	14.56	22.75	-	-	-	-	-	-	-	-	2020
MATT [155]	-	-	-	-	9.7	-	11.7	17.5	80.1	129.4	-	-	355	550.2	2020
NLT [117]	1.42	1.76	1.8	2.42	12.5	-	10.8	18.3	90.1	151.6	165.8	279.7	-	-	2020
BSCC [156]	-	-	-	-	7.9	-	6.7	10.7	58.3	100.1	86.3	153.1	-	-	2020
SDIHD [157]	0.97	1.12	-	-	-	-	-	-	-	-	112	173	-	-	2020
CFANet [158]	-	-	1.2	1.56	-	-	6.5	10.2	56.1	89.6	89	152.3	203.6	287.3	2020
STDNet [159]	0.76	1.01	1.47	1.88	7.05	-	-	-	-	-	-	-	-	-	2021
CRANet [160]	-	-	-	-	-	-	6.6	11	54.6	87.5	95	174	216.4	299	2021
IDK [161]	-	-	-	-	-	-	7.8	12.2	-	-	132	191	212.2	243.7	2021
Crowd-SDNet [162]	-	-	-	-	-	-	7.8	12.6	65.1	104.4	-	-	-	-	2021
Gloss [163]	-	-	-	-	-	-	7.3	11.7	61.3	95.4	84.3	147.5	-	-	2021

TABLE 4 (Continued)

	UCSD		Mall		WorldExpo'10		SHT B		SHT A		UCF-QNRF		UCF_CC_50		
Method	**MAE**	**RMSE**	**MAE**	**RMSE**	**MAE**	**RMSE**	**MAE**	**RMSE**	**MAE**	**RMSE**	**MAE**	**RMSE**	**MAE**	**RMSE**	**Year**
URC [164]	-	-	-	-	-	-	12.03	18.7	72.79	111.61	119.2	211.4	293.99	443.09	2021
FDC18 [165]	-	-	-	-	-	-	11.4	19.1	65.4	109.2	93	157.3	-	-	2021
MFDC18 [165]	-	-	-	-	-	-	6.9	10.3	55.4	91.3	76.2	121.5	-	-	2021
SDNET [166]	-	-	-	-	-	-	-	-	55	92.7	80.7	146.3	197.5	264.1	2021
P2PNat [167]	-	-	-	-	-	-	6.25	9.9	52.74	85.06	85.32	154.5	172.72	256.18	2021
EDIREC [168]	1.79	2.47	2.36	3.12	-	-	-	-	-	-	-	-	-	-	2021
SASNet [169]	-	-	-	-	-	-	6.35	9.9	53.59	88.38	85.2	147.3	161.4	234.46	2021
UOT [170]	-	-	-	-	-	-	6.5	10.2	58.1	95.9	83.3	142.3	-	-	2021
TopoCount [171]	-	-	-	-	-	-	7.8	13.7	61.2	104.6	89	159	184.1	258.3	2021
DSNet [172]	0.82	1.06	-	-	-	-	6.7	10.5	61.7	102.6	91.4	160.4	183.3	240.6	2021
SS-CNN [173]	-	-	-	-	6.83	-	-	-	-	-	115.2	175.7	229.4	325.6	2021
DFNet [174]	-	-	-	-	-	-	14.1	21.1	77.58	129.7	218.2	357.4	402.3	434.1	2021
SSR-HEF [175]	-	-	-	-	6.5	-	6.1	9.5	55	88.3	70.2	128.6	173.3	260.4	2022
STNet [176]	-	-	-	-	-	-	6.25	10.3	52.85	83.64	87.88	166.44	161.96	230.39	2022
SGANet [177]	-	-	-	-	-	-	6.6	10.2	57.6	101.1	87.6	152.5	221.9	289.8	2022
TransCrowd [178]	-	-	-	-	-	-	9.3	16.1	66.1	105.1	97.2	168.5	-	-	2022
HANet [179]	-	-	-	-	-	-	6.8	11.5	54.9	91.2	98	179	195.2	268.6	2022
RAN [175]	-	-	-	-	-	-	7.15	11.86	57.92	99.23	83.38	141.79	155.01	219.45	2022
NDConv [180]	-	-	-	-	-	-	7.8	13.8	61.4	104.18	91.2	165.6	167.2	240.6	2022
FIDTM [181]	-	-	-	-	-	-	6.9	11.8	57	103.4	89	153.5	171.4	233.1	2022
S-GNANet [86]	-	-	-	-	-	-	11.2	22.6	70.4	124.3	-	-	-	-	2022
AGCCM [182]	-	-	-	-	-	-	5.98	9.72	52.75	85.5	-	-	-	-	2022
CLRNet [183]	-	-	1.45	1.84	-	-	-	-	-	-	-	-	-	-	2022
HDNet [184]	-	-	-	-	-	-	-	-	53.4	89.9	83.2	148.3	-	-	2022
MPS [185]	-	-	-	-	-	-	9.6	15	71.4	110.7	-	-	-	-	2022
ESA-NET [82]	-	-	-	-	-	-	8.3	12.9	-	-	-	-	-	-	2022
CrowdFormer [186]	-	-	-	-	-	-	5.7	9.6	56.9	97.4	78.8	136.1	-	-	2022
BLA [187]	-	-	-	-	17.9	-	11.9	18.9	99.3	145	198.9	316.1	346.8	480	2022
MAN [188]	-	-	-	-	-	-	-	-	56.8	90.3	77.3	131.5	-	-	2022
CDCC [189]	-	-	-	-	16.6	-	11.4	17.1	76.3	144.2	134.3	240.3	336.5	486.1	2022
RSICNN [190]	-	-	-	-	-	-	-	-	54.8	89.1	81.6	153.7	186.9	256.5	2022
ChfL [191]	-	-	-	-	-	-	6.9	11	57.5	94.3	80.3	137.6	-	-	2022
DACount [188]	-	-	-	-	-	-	9.6	14.6	67.5	110.7	91.1	153.4	-	-	2022
S-DCNet [192]	-	-	-	-	-	-	6.8	11.5	59.8	100	84.8	142.3	-	-	2022
CLTR [193]	-	-	-	-	-	-	6.5	10.6	56.9	95.2	85.8	141.3	-	-	2022
PAP [189]	-	-	-	-	-	-	17.5	27.5	-	-	-	-	382	594.9	2022
DGCC [54]	-	-	-	-	-	-	12.6	24.6	121.8	203.1	119.4	216.6	-	-	2023
MFCN [19]	-	-	1.6	2.1	-	-	-	-	-	-	-	-	-	-	2023
DMCNet [194]	-	-	-	-	-	-	8.64	13.67	58.46	84.55	96.52	163.99	-	-	2023

Note: Red means the best result, Yellow means the second best result and Green means the third best result; "-" denotes that results are not available in the original paper. The table is sorted according to the year of publication of these methods. From left to right, the three columns are small-scale datasets, large-scale datasets, and hyper-scale datasets. Please note that the results in the WorldExpo'10 dataset are the average result of five intersecting scenes. All results are the lower the value, the better the performance.

where C_{gt_i} and C_{est_i} is the ground truth and prediction results, respectively. N is the number of test images.

In the past five years, the performance of over 100 methods was evaluated on various levels of popular datasets. However, advancements in research on small datasets have lately slowed down with limited new datasets and algorithms being developed for small-scale datasets. The conventional studies on large or hyper-scale datasets seem to have reached a plateau. Consequently, the integration of multiple methods has become a primary area of research.

From Table 4, advancements in research on small datasets have lately slowed down with limited new datasets and algorithms being developed for small-scale datasets. The conventional studies on large or extremely large datasets seem to have reached a plateau. Consequently, the integration of multiple methods has become a primary area of research.

As a result, the combination of multiple approaches and techniques has become a crucial area of research in the field. This is because, for complex real-world problems, a single approach, or technique may not be sufficient to provide optimal solutions. The integration of multiple methods and techniques can help to overcome the limitations of individual methods and provide more robust and accurate results.

3.8 | Summary of crowd counting methods

In the previous sections, we classified different model structures from the perspective of data-driven methods. As the performance of crowd counting models based on convolutional neural networks has improved rapidly on classical datasets in recent years, although models constructed using a single method have significant advantages in lightweight and fast modelling, they have gradually been unable to form breakthrough progress. Models that combine multiple methods have become mainstream, like 'Adaptive Dilated Network with Self-Correction Supervision' (ADSCNet) combines multi-scale network and multi-task network [138]. Due to space limitations, we will briefly summarise several models that have achieved the best validation results on popular datasets in the past 5 years.

SAAN [95]: The method uses a multi-branch network with shared parameters and different receptive fields to capture multi-scale features. Attention mechanisms are introduced to focus on the regions that contribute to the count, and a scale-aware module is designed to adaptively learn the scale variation in crowd scenes.

EPF [195]: The method involves first computing the optical flow of the scene and then using a density map and a velocity map to estimate the people flow. The method also incorporates a temporal component to account for the movement of people in the scene over time.

DeepCount [108]: DeepCount uses a deep convolutional neural network to estimate the crowd density map and then regresses the crowd count from the density map. The network is trained end-to-end to minimise a combination of mean square error and a density-aware loss function. The density-aware loss function takes into account the varying density of the crowd and weighs the loss for each pixel based on its density. The method also includes a post-processing step that uses a Gaussian kernel to refine the density map and improve the accuracy of the crowd count estimation.

DCL [196]: The method involves progressively training a model on increasingly difficult images based on their density levels. The curriculum is designed such that the model learns to count smaller groups of people before moving on to more crowded scenes. The approach also uses a multi-task loss function that combines counting and density estimation.

MS-GAN [197]: The method consists of two networks: a generator and a discriminator. The generator network generates a density map from the input image, while the discriminator network estimates the count of the people in the input image by comparing the generated density map with the ground truth. The generator network is trained to minimise the count estimation error of the discriminator network, while the discriminator network is trained to maximise the count estimation error of the generator network. The proposed method uses a multiscale architecture to capture people of different scales in the crowd.

PaDNet [121]: The method uses a deep neural network that consists of three modules: a backbone network, a multi-scale density map estimator, and a density-aware refinement module. The backbone network extracts feature from the input image, which are then used by the multi-scale density map estimator to generate a set of density maps with different resolutions. The density-aware refinement module then combines the information from these density maps to produce the final crowd count. The method also introduces a novel pan-density loss function that improves the model's ability to count crowds with varying densities.

FSSA [125]: The method consists of a meta-learner that learns to quickly adapt to new scenes and a task-specific network that is trained on the adapted data to perform crowd counting. The meta-learner uses a few labelled samples from the new scene and a large number of labelled samples from a source scene to learn how to adapt the task-specific network to the new scene. The task-specific network is a convolutional neural network with dilated convolutional layers that can capture both local and global features.

SDANet [126]: The method consists of two main parts: a feature extractor and a dense attention module. The feature extractor is a shallow convolutional neural network that extracts features from the input image. The dense attention module is used to weigh the importance of each feature map to improve the accuracy of the density map.

ASNet [69]: The method consists of two modules: a feature extraction module and an attention scaling module. The feature extraction module uses a convolutional neural network (CNN) to extract features from the input image. The attention scaling module then uses the attention mechanism to selectively weigh the features based on their importance for crowd counting. The weighed features are then aggregated using global average pooling and fed into a regression layer to obtain the final crowd count.

FOCNN [139]: In order to minimise the computational cost of training the network, a completely optimised method based on the neural structure search was used to reduce network complexity while achieving better counting performance. Due to the extreme lightweight of network parameters, this network is only suitable for working on small-scale datasets.

ADSCNet [138]: The method is composed of a dilated convolutional neural network (DCNN) with adaptive dilation rates to capture features at different scales and an additional self-correction module to improve counting accuracy. The self-correction module uses a regression approach to correct the predicted count by learning the residuals between the predicted count and the ground truth count. The adaptive dilation rates and self-correction module are jointly optimised through a multi-task learning framework.

MH-MetroNet [153]: The network has three heads, with each head performing a specific task: (1) estimating the total number of passengers, (2) estimating the number of passengers in the ticketing area, and (3) estimating the number of passengers in the platform area. The network uses dilated convolutions and multi-level feature fusion to handle scale variations in the crowd.

AdaCrowd [154]: The method uses the unlabelled target data during training to adapt to the target domain. It uses a teacher-student learning framework, where the teacher network learns to generate crowd density maps from the source data, and the student network learns to estimate the count from the target data. AdaCrowd introduces an adaptation module that adapts the teacher network to the target domain using adversarial learning.

CFANet [158]: The method is composed of two modules: a coarse-grained attention module (CGAM) and a fine-grained attention module (FGAM). The CGAM attends to the global information by downsampling the input and extracting features with larger receptive fields, while the FGAM attends to the local information by upsampling the features and capturing fine details. In addition, a background-aware loss is introduced to better handle the unbalanced foreground and background regions in crowd scenes.

STDNet [159]: The method uses spatiotemporal dilated convolution and uncertain matching for video-based crowd estimation. The method first extracts spatial and temporal features using a convolutional neural network (CNN) and then applies spatiotemporal dilated convolution to capture long-range dependencies in the crowd dynamics. To deal with uncertain matching of people across frames, the method uses a probabilistic matching scheme that estimates the probability of a person appearing in a specific frame.

MFDC18 [165]: The method consists of multiple branches, each processing the input image at different scales to capture multi-scale features. Additionally, a novel correlation-based weighing scheme is introduced to combine the predictions of the different branches. The correlation between samples is learned using a novel Siamese network architecture, which compares the feature maps of different samples to learn their correlation.

P2PNet [167]: The method, called Point-CNN, works by first detecting points of interest in an image by using a density map. Then, a convolutional neural network is trained to estimate the count and density at each point in the image. The authors also introduce a novel loss function called Point Loss, which incorporates both count and density errors. The method can also be extended to handle multiple object types and can be used for other point-based tasks such as localisation and tracking.

SASNet [169]: This method selects the optimal scale for crowd counting by comparing multiple single-scale neural networks and fusing their outputs. This method includes training multiple neural networks with different receptive field to capture different levels of detail and scale and using a gating network to determine the appropriate scale of each input image. The gating network selects the most relevant scale based on the image features and outputs a weight map for each scale, which is used to fuse the outputs of the individual CNNs.

DSNet [172]: The method involves training a deep neural network using densely connected convolutional layers and residual connections and integrating a scale aggregation module into the network architecture to combine information from different scales. The scale aggregation module is designed to explicitly model the scale variation in the input images and to learn a scale map that assigns weights to different scales based on their relevance for counting.

SSR-HEF [175]: The method consists of two stages which are multi-scale semantic refining and hard example focusing. In the first stage, a multi-scale convolutional neural network is used to extract features from the input image, and a semantic refining module is applied to the features to refine the predictions. In the second stage, a hard example focusing module is applied to the refined features to better handle hard examples in the data.

STNet [179]: The method uses a multi-scale feature extraction module to capture scale variation and a multi-level auxiliary branch to enhance the feature representation. The proposed hard example focusing (HEF) method is employed to handle occlusion by emphasising the training samples with high loss. Additionally, STNet adopts a scale-adaptive fusion strategy to improve the accuracy of the final counting results.

RAN [198]: The model consists of two main components: a Region-Aware Module (RAM) and a Scale-Aware Module (SAM). The RAM is a multi-scale and multi-context module that captures different regions of the image and extracts their features. The SAM is designed to handle different scales of objects within the image and learn their scale-wise representations. The RAN model also uses a regional-aware loss function that takes into account the regional density variation of the crowd. This loss function helps the model to focus more on the high-density regions, where it is more important to accurately count the number of people.

AGCCM [182]: The method uses two types of attention mechanisms to selectively focus on the most informative regions of the input image. The first type of attention is spatial attention, which assigns different weights to different regions of the image based on their importance for counting.

The second type is channel attention, which adaptively scales the feature maps to emphasise the most informative channels. Additionally, the method also employs a collaborative counting strategy that combines the outputs of multiple individual models to improve the counting accuracy. Specifically, each model is trained to focus on a different aspect of the input image, such as low- or high-density regions, and the outputs of all models are averaged to produce the final count.

CLRNet [183]: This is a novel deep learning architecture for crowd counting in videos. The model leverages both spatial and temporal information by introducing a cross locality relation module to learn the complex spatiotemporal correlation among different frames in a video sequence. The network consists of two main modules: a backbone network for feature extraction and a cross locality relation module for temporal feature fusion. Additionally, a density-aware loss function is proposed to better handle the density variation problem in crowd counting.

CrowdFormer [186]: CrowdFormer is a crowd counting method based on a vision transformer architecture that utilises the overlap-patching technique. It uses an input image patching strategy to address the spatial variability of crowd density. CrowdFormer also incorporates the global and local information of the image patches with a multi-level feature aggregation approach. It utilises both positional encoding and spatial attention mechanism to process the input image patches. The overlap-patching technique reduces the noise in the density map and improves the counting accuracy.

MAN [199]: The method generates a set of attention maps for each scale based on the multi-scale features of the input image. These attention maps are then combined using a multifaceted attention mechanism to highlight the most relevant regions for crowd counting. The proposed method also introduces a feature fusion module that merges multi-scale features in a weighted manner. Additionally, the paper suggests the use of a multi-objective loss function to optimize the model.

DMCNet [194]: The network has a two-stage architecture. The first stage is a feature extraction network that uses a pretrained VGG16 network. The second stage is a dynamic mixture of counter network that contains a set of counting modules, each having a different receptive field size. The receptive field sizes of the counting modules are adaptively adjusted based on the input image size, enabling location-agnostic crowd counting. The network uses a dynamic weighting scheme to combine the predictions of the counting modules.

4 | THEORETICAL AND PRACTICAL CHALLENGES IN CROWD COUNTING

In the previous section, we discussed the datasets and various experimental methods used in the field of crowd counting. However, while these researches have yielded valuable insights, they also highlight the challenges faced in the crowd counting. The first is that limited dataset diversity and representation can lead to biased and inaccurate results. In addition, current algorithms may not be robust enough to handle complex and dynamic real-world scenarios, such as those involving hyperscale data and multiple modalities. Moreover, the computing resources required for these experiments can be prohibitively expensive and time-consuming.

In this section, we will analyse these challenges from three perspectives that impact AI development and provide suggestions based on existing research and practical experience.

The challenges of crowd counting research mainly come from three aspects:

(1) the challenge in the data;
(2) the challenges in the crowd counting algorithm;
(3) the challenge in computing resources.

4.1 | Challenges in the data

In terms of data, crowd counting faces these difficulties: occlusion, scale changes, uneven crowd distribution, background confusion, diverse illumination and weather, perspective distortion, and image resolution. Among them, occlusion, scale changes, and uneven crowd distribution are the basic challenges in almost all datasets. At the same time, the other challenges do not exist alone, and each dataset contains several or even all these challenges. For example, the Beijing BRT dataset contains both perspective distortion and illumination changes, and the NWPU-crowd dataset covers all these changes.

4.1.1 | Occlusion

Images of the high-density crowd, like Figure 12a, show that people usually overlap and occlude with each other. This makes it difficult for the original detection-based crowd-counting method to complete the people identification. To solve this problem, the researchers changed the method of crowd counting from object detection to density map estimation.

4.1.2 | Scale changes

The change in the size of the person in the image is caused by the distance between the person in the image and the camera, as Figure 12b shows. The person far away from the camera will look smaller in the picture than the person close to the camera. Therefore, it is not easy for computers to identify all of them. The problem of vertical scale change exists in almost all datasets, so almost all crowd counting methods need to consider how to solve this problem. This problem can be translated into the question of how to recognise objects of different sizes in the images. Using detection boxes with different sizes like the SSD [200] or

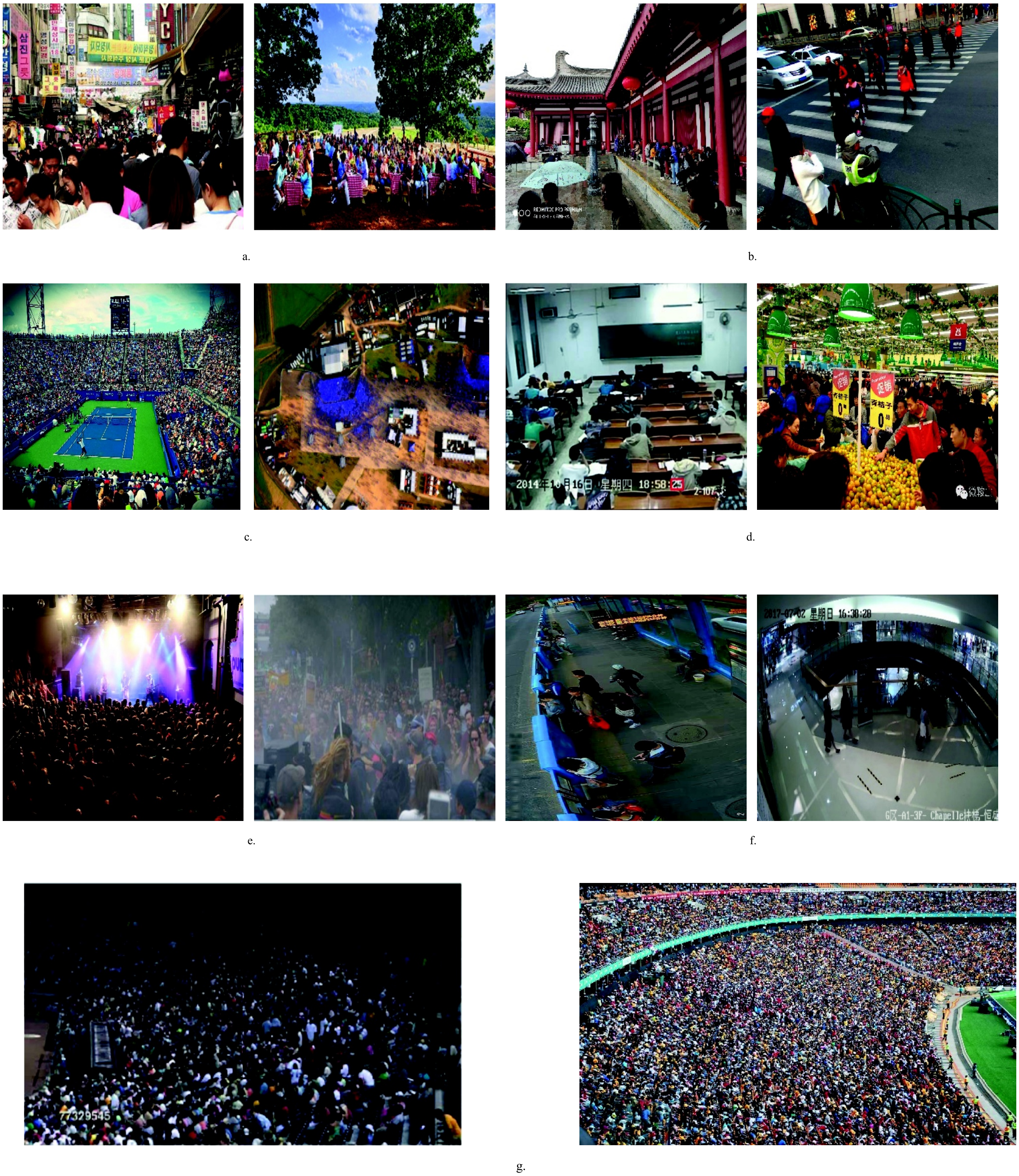


FIGURE 12 Challenges in the data. (a) Occlusion in the image [77, 83]. (b) Scale changes in the image [71, 196]. (c) Uneven crowd distribution in the image [77, 85]. (d) Background confusion in images [67, 148]. (e) Diverse illumination [56] and weather [76]. (f) Perspective distortion in the image [65, 66]. (g) Different image resolutions: left is an image sample in low resolution and right is an image sample in high resolution [79].

YOLO [201] based on the detection method, or using the multi-column [80, 92, 94, 202] and multi-scale networks [65, 84, 136] based on the density map estimation method could solve this problem.

4.1.3 | Uneven crowd distribution

The distribution of crowd in a scene is usually different, and even the distribution of crowd in the same image could show large inhomogeneities, such as competition venues and parks with activities in Figure 12c. For the crowd with varying levels of density, a good approach would be to identify them separately. For example, the Quality Net in DecideNet ranks images [94]; and the multi-level convolutional neural network generates a density map that contains multi-level features [69]; and the Multi-layer Regression Network uses multiple regression to predict people at multiple densities [203].

4.1.4 | Background confusion

Background chaos means that the background of the picture and the foreground have similar colours or textures. For example, in Figure 12d, the complex background can lead to the misidentification of the model. This problem can usually be suppressed and circumvented by using regions of interest, semantic segmentation, and attention methods [33, 136, 204–206].

4.1.5 | Diverse illumination and weather

The difference in illumination includes the difference between natural light and artificial light. Natural light sources mainly affect the brightness of the image during the day and night. In addition, artificial light also affects the colour of the picture. Changes in the weather, such as water reflections on rainy days and blurred images on foggy days, will seriously affect the confusion of the background and the clarity of the image. As Figure 12e shows, under different illumination, the colour of the features will also differ and many interferences in the background of the image on snowy days. These challenges will seriously test the robustness of the model.

4.1.6 | Perspective distortion

The error caused by perspective distortion is a long-standing problem in the field of image recognition. This is usually caused by the lens used to take the pictures, such as a wide-angle lens and a fisheye lens shown in Figure 12f. This phenomenon will drastically cause person scale variation. On the one hand, calibrating the camera and obtaining the perspective map can improve this problem. On the other hand, it can correct the perspective distortion map from the programme method or enable the network to adapt to the distortion of the image [207].

4.1.7 | Image resolution

The resolution level of the picture will affect the richness of the features that the model can extract, especially in hyper-scale crowd data. For example, in Figure 12g, the resolution of the left image is only 562 × 368, and the resolution of the right image is 1020 × 614. When the two images are scaled to the same size, the features retained in the right image will be more than that in the left image. Researchers often enhanced images [208, 209] and expanded the images reasonably [210] to meet this challenge. In addition, whether the image's resolution is conducive to object recognition depends on the number of objects in the image. In order to facilitate the horizontal comparison of datasets, we propose "average pixel occupied by each object" (APO) to measure whether the image size is conducive to target recognition.

4.2 | Challenges in the crowd counting algorithm

There are several aspects that affect the performance of a model: the input of the model, the design of the model, and training and feedback. Researchers will be concerned about the quality of the input. High-quality input is more likely to train a good network. On the other hand, the design of the model directly determines the performance of the trained network. So, the design of the model has been the focus of research in this field.

Meanwhile, the training process is also important. The training system mainly depends on the feedback system during training. Usually, we pay more attention to the performance of the loss function. In addition to the loss used in training, the evaluation criteria used in determining the result of network training are also very important.

4.2.1 | The input of the model

The input of the network includes a training image and a ground truth image. With the development of camera technology, the images of most datasets have begun to enter the era of high resolution. However, in real life, there are still many low-resolution cameras and old-fashioned surveillance cameras. Therefore, researchers may need to perform various preprocessing on the input image to adapt to the actual data situation. In the preprocessing stage, researchers need to process input, such as scaling or segmenting the images to meet the needs of the network. Foreground segmentation and image enhancement are used to reduce the influence of the background. It is used to eliminate irrelevant information in the image, enhance the detectability of related information and simplify the data to the greatest extent, thereby improving the reliability of feature extraction, matching, and recognition.

On the other hand, the generation of ground truth density maps is also crucial. Usually, the dataset will not contain ready-made ground truth density maps; they are all generated by researchers based on the annotation coordinates. The earliest density map was used to solve the problem of cell density on the plane [43]. After introducing it into the field of crowd counting [70], the density maps need to be zoomed, to follow the scale changes and perspective distortion of the images. When the perspective map is available, such as the Beijing BRT dataset [66], the most accurate mapping relationship between the head size and coordinate points can be obtained, resulting in an accurate ground truth density map. In the face of no perspective map, the commonly used method now is to use the adaptive density map generation method proposed by Zhang et al. [80]. This method is based on a high-density crowd, and the distance between each head is related to the density of the crowd. However, in low-density crowd images, the density map generated by this method will be biased due to the number of selected points. The adaptive density map generation method has inspired other researchers. Researchers have made many improvements based on this method, such as the A depth-adaptive kernel-based density map generation method developed by the authors in Ref. [71] and alternative inverse k-nearest neighbour (ikNN) maps developed by Olmschenky et al [99].

4.2.2 | The design of the model

The construction of network models has always been the focus of crowd counting research. From a practical point of view, the size of the network model, the training speed, and the running speed of the model are all worth studying. At present, the development of the network model is basically to solve the problem of high-density crowd counting, so the new model is getting bigger and more complicated. However, scenes of extremely high-density crowds only account for a part of crowd counting application scenarios in real life. Moreover, some studies indicate that the model's size should match the size of the data; that is, the complex networks may not generalise well in daily life [139]. From Table 3, it is easily found that networks, which perform well in large-scale crowd data, rarely achieve the same optimal performance in small-scale crowd data. In addition, restricted by hardware conditions, even though a complex model can obtain good results in experiments, it is not easy to apply in life.

On the other hand, the small network model consumes fewer computing resources and computing time than the large-scale network model [94, 128], which can be an advantage, especially in the days of distributed computing development. The miniaturisation of network models may become a new trend.

4.2.3 | Training and feedback

Network learning requires constant iterations to complete. The loss function has the greatest impact on the iterative learning of the network. Now the mainstream crowd counting method is a regression prediction based on the density map. Euclidean distance is the basic loss function in this field. But only using the Euclidean loss function may ignore some spatial information. The innovative design of the loss function may be a good way to improve network performance. For example, Adversarial Loss [211], SmoothL1 Loss [212], Tukey Loss [213], the spatial correlation loss [204], and the Maximum Excess over Pixels (MEP) loss [106]. On the other hand, the multi-column network can also perform corresponding loss calculations on the output of the sub-networks to obtain a comprehensive loss function. For example, in SFAnet, compared with the use of a single loss function, the network performance is improved by using more appropriate losses for the density graph generation network and attention graph generation network, respectively [214]. Therefore, an innovative loss function or the use of a composite loss function may be a good idea for crowd counting method improvement.

In addition, the evaluation method used in training is also related to the final performance of the network. Usually, researchers are more concerned about the accuracy of the number of people prediction, but in real life, the location information of the crowd may be more important than the number of people. We have noticed that the results of network predictions are usually evaluated using MAE and RMSE (Some papers will use MSE to refer to RMSE). However, these two evaluation methods ignore location information, so in the study, researchers usually add corresponding density maps to prove that the actual predicted target of the network is human. However, these density maps cannot be directly compared quantitatively. Sindagi et al. [215] proposed the use of PSNR and SSIM [216] to evaluate the quality of density maps. To a certain extent, a high-quality density map means high-quality prediction results.

4.3 | Challenges in the computing resources

The limitation of computing power mainly comes from the development of hardware. In the field of deep learning, the main tool of computing has been transformed from CPU to GPU because contemporaneous GPU has more treating Multiprocessors and better parallel computing capabilities. Although dedicated computing cards, such as Tesla and other ASCI, have more cores and faster computing speeds, their high prices are still unaffordable for individual users. Nowadays, the use of Distributed Processing Units (DPUs) to deploy neural network models and 5G-based central computation is becoming the new trend. In general, there are several types of deep learning hardware currently available, including CPUs, GPUs, TPUs, FPGAs, and DPUs.

CPUs (central processing units) are the traditional processors used in most computers, but they are not specifically designed for deep learning and can be slow for complex computations [217].

Advantages: CPUs are widely available and can handle a wide range of tasks, including deep learning. They are relatively inexpensive and can be easily integrated into existing computer systems.

Disadvantages: CPUs are not optimised for deep learning and can be slow when processing large amounts of data. They also consume a lot of power, making them less energy-efficient than other hardware options.

GPUs (graphics processing units) were originally designed for graphics rendering, but their parallel processing capabilities demonstrate tremendous value in deep learning.

Advantages: GPUs offer superior energy efficiency when compared to CPUs, because of their robust parallel processing capabilities which enable them to process voluminous datasets with remarkable efficiency. And GPUs have gained widespread acceptance in the field of deep learning owing to their great proficiency in handling diverse neural network architectures.

Disadvantages: GPUs are a pricier investment and necessitate complementary software to unleash their multithreaded processing potential, thereby elevating their utilisation thresholds and constraints.

TPUs (tensor processing units) are specialised hardware developed by Google specifically for deep learning. They are designed to perform matrix multiplication and other tensor operations more efficiently than GPUs [218, 219].

Advantages: TPUs are designed specifically for deep learning and can process large amounts of data quickly and efficiently. They are highly energy-efficient and can handle a wide range of neural network architectures.

Disadvantages: TPUs are expensive and can be difficult to integrate into existing systems. They also require specialised software and training methods, which can be a barrier to entry for some users [81].

FPGAs (field-programmable gate arrays) are highly customisable hardware that can be programmed to perform specific computations. They are more flexible than other deep learning hardware, but they can be more difficult to programme [220].

Advantages: FPGAs are highly customisable and can be configured to handle a wide range of tasks, including deep learning. They are also more energy-efficient than CPUs and can be programmed to optimise the performance for specific neural network architectures.

Disadvantages: FPGAs are expensive and require specialised expertise to programme and integrate into existing systems. They can also be slower than other hardware options when processing large amounts of data [220].

DPUs (distributed processing units) are a type of deep learning hardware designed to accelerate neural network computation through parallel processing. Unlike traditional CPUs and GPUs, DPUs are optimised specifically for deep learning workloads and are often integrated into larger systems such as edge devices or data centers. DPU combines three key elements [221]: a multi-core CPU that is high-performance, software-programmable, and follows the industry-standard Arm architecture, closely linked to other SoC parts; a packet processing engine that is programmable and has high performance, which can boost the processing of network traffic and reduce the workload on the host CPU; and a hardware accelerator that is programmable and specialised for deep learning inference [222].

Advantages: DPUs in deep learning offer advantages such as high throughput, low latency, low power consumption, and high performance.

Disadvantages: DPUs are expensive and difficult to integrate into existing systems. Additionally, DPUs may require specialised programming expertise in order to take full advantage of their capabilities, which can be a barrier to adoption for some users.

Because of the hardware limitations of distributed devices, smaller and faster network models may be the best user-oriented choice for convolution networks. During the research, we noticed that many convolutional neural network models have great potential. These models often contain a lot of redundancy to handle hyper-density crowd counting that is rarely encountered in daily life. By simplifying the underlying network, we can get a model with the same high recognition accuracy but smaller size and faster speed. For example, our team achieved the best performance on a daily-type crowd dataset by streamlining the design of VGG16 [139].

5 | CONCLUSION

In the past few years, crowd counting has improved significantly. Recently, the field of crowd counting has seen a shift towards using multiple methods in combination, as opposed to relying on a single approach. The need for fresh, targeted datasets that are specifically designed for crowd counting requirements has become increasingly apparent.

Despite the advancements that have been made in the field, crowd counting still presents many challenges. These include handling real-world scenarios such as camera distortions, cluttered backgrounds, and occlusions. There is also a need for algorithms that can perform well under different lighting conditions and weather conditions and for methods that can process large crowds in real time.

In an effort to overcome the limitations of small datasets, researchers are also exploring the use of transfer learning. By using pre-trained models on large datasets, transfer learning can make predictions on smaller datasets, providing a solution to the issue of limited data availability in crowd counting.

This paper provides an overview of the existing crowd technology work from the aspects of datasets and network architecture etc. We summarised almost all existing datasets related to crowd counting and made brief explanations and applicability recommendations for these datasets. At the same time, we have observed that the performance of varied crowd counting networks on different scales of datasets is different, where complex networks are often more capable of estimating high-density crowds, while simple networks are good at estimating low-density crowds. Therefore, we have divided the scale of these datasets. We proposed a three-tier standardised dataset taxonomy to divide the datasets into small-scale datasets belonging to daily-type crowd datasets, large-scale datasets belonging to assembly-type crowd datasets and hyper-scale datasets that have often been used as challenging datasets. We hope that the taxonomy of datasets could help researchers conduct targeted research. In addition, we distinguished the crowd counting algorithm from the perspective of data-driven into six categories and discussed the classical algorithm of each category.

Moreover, we conducted a comprehensive performance benchmark evaluation of the latest 100 models since 2019 and their performance on popular datasets. Although it is not possible to cover all the work, we have highlighted the top three best performers. Through the above summary, we have summarised and discussed several factors that affect the performance of crowd counting and the challenges faced by crowd counting research and put forward some suggestions and thoughts. We hope that this work can help new researchers understand the recent development and cutting-edge works of crowd counting. More importantly, we hope that different data can bring new inspiration to researchers and help them find ways to combine crowd counting with research in other fields.

ACKNOWLEDGEMENT

Additionally, the insightful comments from anonymous reviewers greatly improved the quality of our paper.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

DATA AVAILABILITY STATEMENT

The datasets used and/or analyzed in the current study are public and can be obtained from public sources.

ORCID

Yudong Zhang https://orcid.org/0000-0002-4870-1493

REFERENCES

1. Chan, A.B., Liang, Z.-S.J., Vasconcelos, N.: Privacy preserving crowd monitoring: counting people without people models or tracking. In: Paper Presented at the 2008 IEEE Conference on Computer Vision and Pattern Recognition (2008)
2. Shao, J., et al.: Deeply learned attributes for crowded scene understanding. In: Paper Presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2015)
3. Zhou, B., Wang, X., Tang, X.: Understanding collective crowd behaviors: learning a mixture model of dynamic pedestrian-agents. In: Paper Presented at the 2012 IEEE Conference on Computer Vision and Pattern Recognition (2012)
4. Saxena, S., et al.: Crowd behavior recognition for video. In: Paper Presented at the Advanced Concepts for Intelligent Vision Systems, International Conference (2008)
5. Yi, S., Li, H., Wang, X.: Pedestrian behavior modeling from stationary crowds with applications to intelligent surveillance. IEEE Transactions on Image Processing 25(9), 1–4368 (2016). https://doi.org/10.1109/TIP.2016.2590322
6. Zhou, B., Tang, X., Wang, X.: Learning collective crowd behaviors with dynamic pedestrian-agents. International Journal of Computer Vision 111(1), 50–68 (2015). https://doi.org/10.1007/s11263-014-0735-3
7. Chaker, R., Aghbari, Z.A., Junejo, I.N.: Social network model for crowd anomaly detection and localization. Pattern Recognition 61(Complete), 266–281 (2017). https://doi.org/10.1016/j.patcog.2016.06.016
8. Li, W., Mahadevan, V., Vasconcelos, N.: Anomaly detection and localization in crowded scenes. IEEE Transactions on Pattern Analysis and Machine Intelligence 36(1), 18–32 (2014). https://doi.org/10.1109/tpami.2013.111
9. Benabbas, Y., Ihaddadene, N., Djeraba, C.: Motion pattern extraction and event detection for automatic visual surveillance. EURASIP Journal on Image and Video Processing 2011(1), 1–15 (2011). https://doi.org/10.1155/2011/163682
10. Lee, S., Kim, H.G., Yong, M.R.: STAN: spatio-temporal adversarial networks for abnormal event detection. In: Paper Presented at the ICASSP 2018 - 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (2018)
11. Alnabulsi, H., Drury, J.: Social identification moderates the effect of crowd density on safety at the Hajj. Proceedings of the National Academy of Sciences 111(25), 9091–9096 (2014). https://doi.org/10.1073/pnas.1404953111
12. Moussaïd, M., et al.: The walking behaviour of pedestrian social groups and its impact on crowd dynamics. PLOS ONE 5(4), e10047 (2010). https://doi.org/10.1371/journal.pone.0010047
13. Aveni, A.F.: The not-so-lonely crowd: friendship groups in collective behavior. Sociometry 40(1), 96–99 (1977). https://doi.org/10.2307/3033551
14. Parrish, J.K., Edelstein-Keshet, L.: Complexity, pattern, and evolutionary trade-offs in animal aggregation. Science 284(5411), 99–101 (1999). https://doi.org/10.1126/science.284.5411.99
15. Zhang, H.P., et al.: Collective motion and density fluctuations in bacterial colonies. Proceedings of the National Academy of Sciences 107(31), 13626–13630 (2010). https://doi.org/10.1073/pnas.1001651107
16. Chen, Y., et al.: Large group activity security risk assessment and risk early warning based on random forest algorithm. Pattern Recognition Letters 144, 1–5 (2021). https://doi.org/10.1016/j.patrec.2021.01.008
17. Koswatte, S., McDougall, K., Liu, X.: Crowd-assisted flood disaster management. In: Singh, V.P., et al. (eds.) Application of Remote Sensing and GIS in Natural Resources and Built Infrastructure Management, pp. 39–55. Springer International Publishing (2022)
18. Liu, J., Chen, Y., Chen, Y.: Emergency and disaster management-crowd evacuation research. Journal of Industrial Information Integration 21, 100191 (2021). https://doi.org/10.1016/j.jii.2020.100191
19. Deng, L., et al.: Hospital crowdedness evaluation and in-hospital resource allocation based on image recognition technology. Scientific Reports 13(1), 299 (2023). https://doi.org/10.1038/s41598-022-24221-6
20. Lu, L., et al.: A study of pedestrian group behaviors in crowd evacuation based on an extended floor field cellular automaton model. Transportation Research Part C: Emerging Technologies 81, 317–329 (2017). https://doi.org/10.1016/j.trc.2016.08.018
21. Liu, Z., et al.: Decision-making framework for GI layout considering site suitability and weighted multi-function effectiveness: a case study in beijing sub-center. Water 14(11), 1765 (2022). https://doi.org/10.3390/w14111765
22. Hu, R., et al.: RDC-SAL: refine distance compensating with quantum scale-aware learning for crowd counting and localization. Applied Intelligence 52(12), 14336–14348 (2022). https://doi.org/10.1007/s10489-022-03238-4
23. Perez, H., et al.: Task-based crowd simulation for heterogeneous architectures. In: Hassan, Q.F. (ed.) Innovative Research and Applications in Next-Generation High Performance Computing, pp. 194–219. IGI Global (2016)
24. Häni, N., Roy, P., Isler, V.: MinneApple: a benchmark dataset for apple detection and segmentation. IEEE Robotics and Automation Letters 5(2), 852–858 (2020). https://doi.org/10.1109/LRA.2020.2965061
25. Deng, L., Wang, S.-H., Zhang, Y.-D.: ELMGAN: a GAN-based efficient lightweight multi-scale-feature-fusion multi-task model. Knowledge-Based Systems 252, 109434 (2022). https://doi.org/10.1016/j.knosys.2022.109434
26. Xu, H., et al.: Efficient CityCam-to-edge cooperative learning for vehicle counting in ITS. IEEE Transactions on Intelligent Transportation Systems 23(9), 16600–16611 (2022). https://doi.org/10.1109/TITS.2022.3149657
27. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: Paper Presented at the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) (2005)
28. Leibe, B., Seemann, E., Schiele, B.: Pedestrian detection in crowded scenes, CVPR 2005. IEEE Computer Society Conference on. In: Paper Presented at the Computer Vision and Pattern Recognition, 2005 (2005)
29. Lin, S.F., Chen, J.Y., Chao, H.X.: Estimation of number of people in crowded scenes using perspective transformation. IEEE Transactions on Systems Man and Cybernetics Part A 31(6), 0–654 (2001)

30. Viola, P., Jones, M.J.: Robust real-time face detection. International Journal of Computer Vision 57(2), 137–154 (2004). https://doi.org/10.1023/b:visi.0000013087.49260.fb
31. Viola, P., Jones, M.J., Snow, D.: Detecting pedestrians using patterns of motion and appearance. In: Paper Presented at the Proceedings Ninth IEEE International Conference on Computer Vision (2003). https://ieeexplore.ieee.org/document/1238422/
32. Wu, B., Nevatia, R.: Detection of multiple, partially occluded humans in a single image by Bayesian combination of edgelet part detectors. In: Paper Presented at the Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1 (2005)
33. Li, M., et al.: Estimating the number of people in crowded scenes by mid based foreground segmentation and head-shoulder detection. In: Paper Presented at the 2008 19th International Conference on Pattern Recognition (2008)
34. Dollar, P., et al.: Pedestrian detection: an evaluation of the state of the art. IEEE transactions on pattern analysis 34(4), 743–761 (2011). https://doi.org/10.1109/tpami.2011.155
35. Chan, A.B., Vasconcelos, N.: Counting people with low-level features and bayesian regression. IEEE Transactions on Image Processing 21(4), 2160–2177 (2012). https://doi.org/10.1109/TIP.2011.2172800
36. Chen, K., et al.: Feature mining for localised crowd counting. In: Paper Presented at the BMVC (2012)
37. Marana, A.N., et al.: On the efficacy of texture analysis for crowd monitoring. In: Paper Presented at the Proceedings SIBGRAPI'98. International Symposium on Computer Graphics, Image Processing, and Vision (Cat. No.98EX237) (1998)
38. Paragios, N., Ramesh, V.: A MRF-based approach for real-time subway monitoring. In: Paper Presented at the Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (2001). CVPR 2001
39. Ryan, D., et al.: Crowd counting using multiple local features. In: Paper Presented at the 2009 Digital Image Computing: Techniques and Applications (2009)
40. Han, M., Xu, W., Gong, Y.: Video foreground segmentation based on sequential feature clustering. In: Paper Presented at the International Conference on Pattern Recognition (2006)
41. Ma, Z., Chan, A.B.: Crossing the line: crowd counting by integer programming with local features. In: Paper Presented at the 2013 IEEE Conference on Computer Vision and Pattern Recognition (2013)
42. Wang, Y., et al.: Counting people with support vector regression. In: Paper Presented at the International Conference on Natural Computation (2014)
43. Lempitsky, V., Zisserman, A.: Learning to count objects in images. In: Paper Presented at the Proceedings of the 23rd International Conference on Neural Information Processing Systems (NIPS'10) (2010). https://dl.acm.org/doi/10.5555/2997189.2997337
44. Fu, M., et al.: Fast crowd density estimation with convolutional neural networks. Engineering Applications of Artificial Intelligence 43, 81–88 (2015). https://doi.org/10.1016/j.engappai.2015.04.006
45. Wang, C., et al.: Deep people counting in extremely dense crowds. In: Paper Presented at the Proceedings of the 23rd ACM International Conference on Multimedia (2015)
46. Boominathan, L., Kruthiventi, S.S., Babu, R.V.: Crowdnet: a deep convolutional network for dense crowd counting. In: Paper Presented at the Proceedings of the 24th ACM International Conference on Multimedia (2016)
47. Junyu, G., Qi, W., Xuelong, L.: PCC Net: perspective crowd counting via spatial convolutional network. IEEE Transactions on Circuits and Systems for Video Technology 30(10), 3486–3498 (2020). https://doi.org/10.1109/TCSVT.2019.2919139
48. Kumagai, S., Hotta, K., Kurita, T.: Mixture of counting CNNs. Machine Vision and Applications 29(7), 1119–1126 (2018). https://doi.org/10.1007/s00138-018-0955-6
49. Shen, Z., et al.: Crowd counting via adversarial cross-scale consistency pursuit. In: Paper Presented at the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (2018)
50. Yan, Z., et al.: Perspective-guided convolution networks for crowd counting. In: Paper Presented at the 2019 IEEE/CVF International Conference on Computer Vision (ICCV) (2019). https://ieeexplore.ieee.org/document/9010874
51. Fang, Y., et al.: Multi-level feature fusion based Locality-Constrained Spatial Transformer network for video crowd counting. Neurocomputing 392, 98–107 (2020). https://doi.org/10.1016/j.neucom.2020.01.087
52. Loy, C.C., et al.: Crowd counting and profiling: methodology and evaluation. In: Ali, S., et al. (eds.) Modeling, Simulation and Visual Analysis of Crowds: A Multidisciplinary Perspective, pp. 347–382 (2013)
53. Sindagi, V.A., Patel, V.M.: A survey of recent advances in CNN-based single image crowd counting and density estimation. Pattern Recognition Letters 107, 3–16 (2018). https://doi.org/10.1016/j.patrec.2017.07.007
54. Du, Z., Deng, J., & Shi, M. (2022). Domain-general Crowd Counting in Unseen Scenarios. arXiv e-prints, arXiv:2212.02573. https://doi.org/10.48550/arXiv.2212.02573
55. Liu, X., et al.: Exploiting sample correlation for crowd counting with multi-expert network. In: Paper Presented at the 2021 IEEE/CVF International Conference on Computer Vision (ICCV) (2021)
56. Wang, Q., et al.: NWPU-crowd: a large-scale benchmark for crowd counting. IEEE Transactions on Pattern Analysis and Machine Intelligence 43(6), 2141–2149 (2021). https://doi.org/10.1109/TPAMI.2020.3013269
57. Oktay, O., et al.: Multi-input cardiac image super-resolution using convolutional neural networks. In: Paper Presented at the Medical Image Computing and Computer-Assisted Intervention - MICCAI 2016 (2016)
58. Sang, J., et al.: Improved crowd counting method based on scale-adaptive convolutional neural network. IEEE Access 7, 24411–24419 (2019). https://doi.org/10.1109/ACCESS.2019.2899939
59. Yu, H., et al.: Convolutional neural networks for medical image analysis: state-of-the-art, comparisons, improvement and perspectives. Neurocomputing 444, 92–110 (2021). https://doi.org/10.1016/j.neucom.2020.04.157
60. Kadimesetty, V.S., et al.: Convolutional neural network-based robust denoising of low-dose computed tomography perfusion maps. IEEE Transactions on Radiation and Plasma Medical Sciences 3(2), 137–152 (2019). https://doi.org/10.1109/TRPMS.2018.2860788
61. Ferryman, J., Ellis, A.: PETS2010: dataset and challenge. In: Paper Presented at the 2010 7th IEEE International Conference on Advanced Video and Signal Based Surveillance (2010)
62. Bondi, E., et al.: Real-time people counting from depth imagery of crowded environments. In: Paper Presented at the 2014 11th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS) (2014)
63. Farhood, H., et al.: Counting people based on linear, weighted, and local random forests. In: Paper Presented at the 2017 International Conference on Digital Image Computing: Techniques and Applications (DICTA) (2017)
64. Kang, D., Dhar, D., Chan, A.: Incorporating side information by adaptive convolution. In: Paper Presented at the Conference and Workshop on Neural Information Processing Systems 2017 (NIPS 2017) (2017). http://papers.nips.cc/paper/6976-incorporating-side-information-by-adaptive-convolution.pdf
65. Zhang, L., Shi, M., Chen, Q.: Crowd counting via scale-adaptive convolutional neural network. In: Paper Presented at the 2018 IEEE Winter Conference on Applications of Computer Vision (WACV) (2018). http://doi.ieeecomputersociety.org/10.1109/WACV.2018.00127
66. Ding, X., et al.: A deeply-recursive convolutional network for crowd counting. In: Paper Presented at the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (2018)
67. Ling, M., Geng, X.: Indoor crowd counting by mixture of Gaussians label distribution learning. IEEE Transactions on Image Processing 28(11), 5691–5701 (2019). https://doi.org/10.1109/TIP.2019.2922818

68. Fang, Y., et al.: Locality-constrained spatial transformer network for video crowd counting. In: Paper Presented at the 2019 IEEE International Conference on Multimedia and Expo (ICME) (2019)
69. Jiang, X., et al.: Learning multi-level density maps for crowd counting. IEEE Transactions on Neural Networks and Learning Systems 31(8), 2705–2715 (2020). https://doi.org/10.1109/TNNLS.2019.2933920
70. Zhang, C., et al.: Cross-scene crowd counting via deep convolutional neural networks. In: Paper Presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2015)
71. Lian, D., et al.: Density Map Regression Guided Detection Network for RGB-D Crowd Counting and Localization (2019). http://openaccess.thecvf.com/content_CVPR_2019/html/Lian_Density_Map_Regression_Guided_Detection_Network_for_RGB-D_Crowd_Counting_CVPR_2019_paper.html
72. Fox, D.: A large-scale hierarchical multi-view RGB-D object dataset. In: Paper Presented at the IEEE International Conference on Robotics and Automation, ICRA 2011 (2011)
73. Wen, L., et al.: Detection, tracking, and counting meets drones in crowds: a benchmark. In: Paper Presented at the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2021)
74. Zhang, Q., Chan, A.B.: Wide-area crowd counting via ground-plane density maps and multi-view fusion CNNs. In: Paper Presented at the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
75. Wan, J., Kumar, N.S., Chan, A.B.: Fine-grained crowd counting. IEEE Transactions on Image Processing 30, 2114–2126 (2021). https://doi.org/10.1109/TIP.2021.3049938
76. Sindagi, V.A., Yasarla, R., Patel, V.M.: Pushing the Frontiers of Unconstrained Crowd Counting: New Dataset and Benchmark Method (2019). arXiv e-print arXiv:1910.12384
77. Sindagi, V.A., Yasarla, R., Patel, V.M.: JHU-CROWD++: Large-Scale Crowd Counting Dataset and A Benchmark Method (2020). arXiv e-print arXiv: 2004.03597
78. Jiang, X., et al.: Attention scaling for crowd counting. In: Paper Presented at the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
79. Hu, Y., et al.: Dense crowd counting from still images with convolutional neural networks C %J J. Vis. Comun. Image Represent. Journal of Visual Communication and Image Representation 38, 530–539 (2016). https://doi.org/10.1016/j.jvcir.2016.03.021
80. Zhang, Y., et al.: Single-image crowd counting via multi-column convolutional neural network. In: Paper Presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2016)
81. Wang, Q., et al.: Learning from synthetic data for crowd counting in the wild. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2019). https://doi.org/10.1109/cvpr.2019.00839
82. Hou, Y., et al.: Enhancing and dissecting crowd counting by synthetic data. In: Paper Presented at the ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Singapore (2022)
83. Idrees, H., et al.: Composition Loss for Counting, Density Map Estimation and Localization in Dense Crowds (2018). arXiv:1808.01050v1
84. Idrees, H., et al.: Multi-source multi-scale counting in extremely dense crowd images. In: Paper Presented at the IEEE Conference on Computer Vision and Pattern Recognition (2013)
85. Reza, B., Eleonora, V., Peter, R.: MRCNet: crowd counting and density map estimation in aerial and ground imagery. In: Paper Presented at the BMVC's Workshop on Object Detection and Recognition for Security Screenin (BMVC-ODRSS) (2019)
86. Li, H., et al.: Video crowd localization with multifocus Gaussian neighborhood attention and a large-scale benchmark. IEEE Transactions on Image Processing 31, 6032–6047 (2022). https://doi.org/10.1109/TIP.2022.3205210
87. Wu, X., et al.: Crowd density estimation using texture analysis and learning. In: Paper Presented at the 2006 IEEE International Conference on Robotics and Biomimetics (2006). https://ieeexplore.ieee.org/document/4141867
88. An, S., Liu, W., Venkatesh, S.: Face recognition using kernel ridge regression. In: Paper Presented at the 2007 IEEE Conference on Computer Vision and Pattern Recognition (2007). https://ieeexplore.ieee.org/document/4270130
89. Chen, J.-C., et al.: A cascaded convolutional neural network for age estimation of unconstrained faces. In: Paper Presented at the 2016 IEEE 8th International Conference on Biometrics Theory, Applications and Systems (BTAS) (2016)
90. Wang, Y., Zou, Y.: Fast visual object counting via example-based density estimation. In: Paper Presented at the 2016 IEEE International Conference on Image Processing (ICIP) (2016)
91. Shang, C., Ai, H., Bai, B.: End-to-end crowd counting via joint learning local and global count. In: Paper Presented at the 2016 IEEE International Conference on Image Processing (ICIP) (2016)
92. Sam, D.B., Surya, S., Babu, R.V.: Switching convolutional neural network for crowd counting. In: Paper Presented at the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
93. Sindagi, V.A., Patel, V.M.: Cnn-based cascaded multi-task learning of high-level prior and density estimation for crowd counting. In: Paper Presented at the 2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS) (2017)
94. Liu, J., et al.: DecideNet: counting varying density crowds through attention guided detection and density estimation. In: Paper Presented at the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (2018). https://ieeexplore.ieee.org/document/8578643
95. Hossain, M., et al.: Crowd counting using scale-aware attention networks. In: Paper Presented at the 2019 IEEE Winter Conference on Applications of Computer Vision (WACV) (2019). https://ieeexplore.ieee.org/document/8659316
96. Sam, D.B., et al.: Almost unsupervised learning for dense crowd counting. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp. 8868–8875 (2019). https://doi.org/10.1609/aaai.v33i01.33018868
97. Li, Y., Zhang, X., Chen, D.: CSRNet: dilated convolutional neural networks for understanding the highly congested scenes. In: Paper Presented at the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (2018)
98. Huang, S., et al.: Stacked pooling: improving crowd counting by boosting scale invariance. In: Paper Presented at the ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (2020). https://ieeexplore.ieee.org/document/9053070
99. Olmschenk, G., Tang, H., Zhu, Z.: Improving dense crowd counting convolutional neural networks using inverse k-nearest neighbor maps and multiscale upsampling. In: Paper Presented at the Proceedings of the 15th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (2020). https://www.scitepress.org/Link.aspx?doi=10.5220/0009156201850195
100. Wu, X., et al.: Fast video crowd counting with a temporal aware network. Neurocomputing 403, 13–20 (2020). https://doi.org/10.1016/j.neucom.2020.04.071
101. Gao, C., Wang, P., Gao, Y.: MobileCount: an efficient encoder-decoder framework for real-time crowd counting. In: Paper Presented at the Pattern Recognition and Computer Vision (2019). https://doi.org/10.1007/978-3-030-31723-2_50
102. Gao, J., Yuan, Y., Wang, Q.: Feature-aware adaptation and density alignment for crowd counting in video surveillance. IEEE Transactions on Cybernetics 51(10), 4822–4833 (2021). https://doi.org/10.1109/TCYB.2020.3034316
103. Oh, M.-h., Olsen, P.A., Natesan Ramamurthy, K.: Crowd counting with decomposed uncertainty. In: Paper Presented at the Proceedings of the

AAAI Conference on Artificial Intelligence (2020). https://doi.org/10.1609/aaai.v34i07.6852

104. Liu, L., et al.: DENet: a universal network for counting crowd with varying densities and scales. IEEE Transactions on Multimedia 23, 1060–1068 (2021). https://doi.org/10.1109/TMM.2020.2992979
105. Basalamah, S., Khan, S.D., Ullah, H.: Scale driven convolutional neural network model for people counting and localization in crowd scenes. IEEE Access 7, 71576–71584 (2019). https://doi.org/10.1109/ACCESS.2019.2918650
106. Babu Sam, D., et al.: Locate, size and count: accurately resolving people in dense crowds via detection. IEEE Transactions on Pattern Analysis and Machine Intelligence 43(8), 2739–2751 (2019). https://doi.org/10.1109/TPAMI.2020.2974830
107. Liu, Y., et al.: Crowd counting via cross-stage refinement networks. IEEE Transactions on Image Processing 29, 6800–6812 (2020). https://doi.org/10.1109/TIP.2020.2994410
108. Chen, Z., et al.: Deep density-aware count regressor. In: Paper Presented at the the 24th European Conference on Artificial Intelligence (ECAI 2020) (2019). https://arxiv.org/abs/1908.03314
109. Chen, X., et al.: Relevant region prediction for crowd counting. Neurocomputing 407, 399–408 (2020). https://doi.org/10.1016/j.neucom.2020.04.117
110. Chen, J., Su, W., Wang, Z.: Crowd counting with crowd attention convolutional neural network. Neurocomputing 382, 210–220 (2020). https://doi.org/10.1016/j.neucom.2019.11.064
111. Wu, X., et al.: Counting crowds with varying densities via adaptive scenario discovery framework. Neurocomputing 397, 127–138 (2020). https://doi.org/10.1016/j.neucom.2020.02.045
112. Dong, Z., et al.: Scale-Recursive Network with point supervision for crowd scene analysis. Neurocomputing 384, 314–324 (2020). https://doi.org/10.1016/j.neucom.2019.12.070
113. Ding, X., et al.: Crowd density estimation using fusion of multi-layer features. IEEE Transactions on Intelligent Transportation Systems 22(8), 1–12 (2020). https://doi.org/10.1109/TITS.2020.2983475
114. Jiang, X., et al.: Density-aware multi-task learning for crowd counting. IEEE Transactions on Multimedia 23, 443–453 (2021). https://doi.org/10.1109/TMM.2020.2980945
115. Zhao, M., et al.: Scale-aware crowd counting via depth-embedded convolutional neural networks. IEEE Transactions on Circuits and Systems for Video Technology 30(10), 3651–3662 (2020). https://doi.org/10.1109/TCSVT.2019.2943010
116. Sajid, U., et al.: ZoomCount: a zooming mechanism for crowd counting in static images. IEEE Transactions on Circuits and Systems for Video Technology 30(10), 3499–3512 (2020). https://doi.org/10.1109/TCSVT.2020.2978717
117. Wang, Q., et al.: Density-aware curriculum learning for crowd counting. IEEE Transactions on Cybernetics 52(6), 1–13 (2020). https://doi.org/10.1109/TCYB.2020.3033428
118. Zou, Z., et al.: Crowd counting via hierarchical scale recalibration network. In: Paper Presented at the ECAI 2020, the 24th European Conference on Artificial Intelligence (2020). http://ecai2020.eu/papers/424_paper.pdf
119. Mo, H., et al.: Background noise filtering and distribution dividing for crowd counting. IEEE Transactions on Image Processing 29, 8199–8212 (2020). https://doi.org/10.1109/TIP.2020.3009030
120. Liu, L., et al.: Efficient crowd counting via structured knowledge transfer. In: Paper Presented at the the 28th ACM International Conference on Multimedia (2020). https://doi.org/10.1145/3394171.3413938
121. Tian, Y., et al.: PaDNet: pan-density crowd counting. IEEE Transactions on Image Processing 29, 2714–2727 (2020). https://doi.org/10.1109/TIP.2019.2952083
122. Sindagi, V.A., Patel, V.M.: HA-CCN: hierarchical attention-based crowd counting network. IEEE Transactions on Image Processing 29, 323–335 (2020). https://doi.org/10.1109/TIP.2019.2928634
123. Ranjan, V., et al.: Uncertainty estimation and sample selection for crowd counting. In: Paper Presented at the the 15th Asian Conference on Computer Vision (ACCV 2020) (2020). https://link.springer.com/chapter/10.1007/978-3-030-69541-5_23
124. Sajid, U., Wang, G.: Plug-and-Play rescaling based crowd counting in static images. In: Paper Presented at the 2020 IEEE Winter Conference on Applications of Computer Vision (WACV) (2020). https://ieeexplore.ieee.org/document/9093561
125. Krishna Reddy, M.K., et al.: Few-shot scene adaptive crowd counting using meta-learning. In: Paper Presented at the 2020 IEEE Winter Conference on Applications of Computer Vision (WACV) (2020). https://ieeexplore.ieee.org/document/9093409
126. Miao, Y., et al.: Shallow feature based dense attention network for crowd counting. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp. 11765–11772 (2020). https://doi.org/10.1609/aaai.v34i07.6848
127. Luo, A., et al.: Hybrid graph neural networks for crowd counting. In: Paper Presented at the Proceedings of the AAAI Conference on Artificial Intelligence (2020). https://ojs.aaai.org/index.php/AAAI/article/view/6839
128. Shi, X., et al.: A real-time deep network for crowd counting. In: Paper Presented at the ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (2020). https://ieeexplore.ieee.org/document/9053780
129. Han, T., et al.: Focus on semantic consistency for cross-domain crowd understanding. In: Paper Presented at the ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Barcelona (2020). https://ieeexplore.ieee.org/abstract/document/9054768
130. Wei, B., Yuan, Y., Wang, Q.: MSPNET: multi-supervised parallel network for crowd counting. In: Paper Presented at the ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (2020). https://ieeexplore.ieee.org/document/9054479
131. Wang, M., et al.: Stochastic multi-scale aggregation network for crowd counting. In: Paper Presented at the ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (2020). https://ieeexplore.ieee.org/document/9054238
132. Hou, Y., et al.: BBA-NET: a Bi-branch attention network for crowd counting. In: Paper Presented at the ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (2020). https://ieeexplore.ieee.org/document/9053955
133. Pan, X., et al.: Attention guided region division for crowd counting. In: Paper Presented at the ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (2020). https://ieeexplore.ieee.org/document/9053761
134. Kong, X., et al.: Weakly supervised crowd-wise attention for robust crowd counting. In: Paper Presented at the ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (2020). https://ieeexplore.ieee.org/document/9054258
135. Duan, H., Wang, S., Guan, Y.: SOFA-net: second-order and first-order attention network for crowd counting. In: Paper Presented at the the 31st British Machine Vision Virtual Conference (BMVC2020). (2020). https://www.bmvc2020-conference.com/assets/papers/0222.pdf
136. Chen, Y., et al.: Scale-aware rolling fusion network for crowd counting. In: Paper Presented at the 2020 IEEE International Conference on Multimedia and Expo (ICME) (2020). https://ieeexplore.ieee.org/document/9102854
137. Yang, Y., et al.: Reverse perspective network for perspective-aware object counting. In: Paper Presented at the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2020). https://ieeexplore.ieee.org/document/9156571
138. Bai, S., et al.: Adaptive dilated network with self-correction supervision for counting. In: Paper Presented at the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2020). https://ieeexplore.ieee.org/document/9157413
139. Deng, L., Wang, S.H., Zhang, Y.D.: Fully optimized convolutional neural network based on small-scale crowd. In: Paper Presented at the 2020 IEEE International Symposium on Circuits and Systems (ISCAS) (2020). https://ieeexplore.ieee.org/document/9180823

140. Yang, Y., et al.: Weakly-supervised crowd counting learns from sorting rather than locations. In: Paper Presented at the ECCV 2020 : 16th European Conference on Computer Vision (2020). https://link.springer.com/chapter/10.1007/978-3-030-58598-3_1

141. Zhao, Z., et al.: Active crowd counting with limited supervision. In: Paper Presented at the ECCV 2020 : 16th European Conference on Computer Vision (2020). https://link.springer.com/chapter/10.1007/978-3-030-58565-5_34

142. Liu, Y., et al.: Semi-supervised crowd counting via self-training on surrogate tasks. In: Paper Presented at the ECCV 2020 : 16th European Conference on Computer Vision (2020). https://link.springer.com/chapter/10.1007/978-3-030-58555-6_15

143. Sindagi, V.A., et al.: Learning to count in the crowd from limited labeled data. In: Paper Presented at the ECCV 2020 : 16th European Conference on Computer Vision (2020). https://link.springer.com/chapter/10.1007/978-3-030-58621-8_13

144. Liu, L., et al.: Weighing counts: sequential crowd counting by reinforcement learning. In: Paper Presented at the ECCV 2020 : 16th European Conference on Computer Vision (2020). https://link.springer.com/chapter/10.1007/978-3-030-58607-2_10

145. Liu, X., Yang, J., Ding, W.: Adaptive mixture regression network with local counting map for crowd counting. In: Paper Presented at the ECCV 2020 : 16th European Conference on Computer Vision (2020). https://rd.springer.com/chapter/10.1007/978-3-030-58586-0_15

146. Hu, Y., et al.: NAS-count: counting-by-density with neural architecture search. In: Paper Presented at the ECCV 2020: 16th European Conference on Computer Vision (2020). https://dl.acm.org/doi/abs/10.1007/978-3-030-58542-6_45

147. Liu, Y., et al.: Towards unsupervised crowd counting via regression-detection Bi-knowledge transfer. In: Paper Presented at the Proceedings of the 28th ACM International Conference on Multimedia Seattle WA USA (2020). https://doi.org/10.1145/3394171.3413825

148. Wang, Q., et al.: Pixel-wise crowd understanding via synthetic data. International Journal of Computer Vision 129(1), 225–245 (2021). https://doi.org/10.1007/s11263-020-01365-4

149. Wan, J., Wang, Q., Chan, A.B.: Kernel-based density map generation for dense object counting. IEEE Transactions on Pattern Analysis and Machine Intelligence 44(3), 1–1370 (2020). https://doi.org/10.1109/TPAMI.2020.3022878

150. Wan, J., Chan, A.: Modeling noisy annotations for crowd counting. In: Paper Presented at the Neural Information Processing Systems (NeurIPS 2020), Online Conference (2020). https://proceedings.neurips.cc/paper/2020/file/22bb543b251c39ccdad8063d486987bb-Paper.pdf

151. Wang, Q., et al.: Neuron linear transformation: modeling the domain shift for crowd counting. IEEE Transactions on Neural Networks and Learning Systems 33(8), 1–13 (2020). https://doi.org/10.1109/TNNLS.2021.3051371

152. Thanasutives, P., et al.: Encoder-decoder based convolutional neural networks with multi-scale-aware modules for crowd counting. In: 2020 25th International Conference on Pattern Recognition (ICPR), pp. 2382–2389 (2021). https://doi.org/10.1109/ICPR48806.2021.9413286

153. Mazzeo, P.L., et al.: MH-MetroNet—a multi-head CNN for passenger-crowd attendance estimation. Journal of Imaging 6(7), 62 (2020). https://doi.org/10.3390/jimaging6070062

154. Krishna Reddy, M.K., et al.: AdaCrowd: unlabeled scene adaptation for crowd counting. IEEE Transactions on Multimedia 24, 1008–1019 (2022). https://doi.org/10.1109/TMM.2021.3062481

155. Lei, Y., et al.: Towards using count-level weak supervision for crowd counting. Pattern Recognition 109, 107616 (2021). https://doi.org/10.1016/j.patcog.2020.107616

156. Modolo, D., et al.: Understanding the impact of mistakes on background regions in crowd counting. In: Paper Presented at the 2021 IEEE Winter Conference on Applications of Computer Vision (WACV) (2020). https://ieeexplore.ieee.org/document/9423112

157. Khan, S.D., Basalamah, S.: Scale and density invariant head detection deep model for crowd counting in pedestrian crowds. The Visual Computer 37(8), 2127–2137 (2021). https://doi.org/10.1007/s00371-020-01974-7

158. Rong, L., Li, C.: Coarse- and fine-grained attention network with background-aware loss for crowd density map estimation. In: Paper Presented at the 2021 IEEE Winter Conference on Applications of Computer Vision (WACV) (2021). https://ieeexplore.ieee.org/document/9423141

159. Ma, Y.J., Shuai, H.H., Cheng, W.H.: Spatiotemporal dilated convolution with uncertain matching for video-based crowd estimation. IEEE Transactions on Multimedia 24, 261–273 (2022). https://doi.org/10.1109/TMM.2021.3050059

160. Wu, Z., et al.: CRANet: cascade residual attention network for crowd counting. In: Paper Presented at the 2021 IEEE International Conference on Multimedia and Expo (ICME) (2021)

161. Cai, Y., et al.: Leveraging intra-domain knowledge to strengthen cross-domain crowd counting. In: Paper Presented at the 2021 IEEE International Conference on Multimedia and Expo (ICME) (2021)

162. Wang, Y., et al.: A self-training approach for point-supervised object detection and counting in crowds. IEEE Transactions on Image Processing 30, 2876–2887 (2021). https://doi.org/10.1109/TIP.2021.3055632

163. Wan, J., Liu, Z., Chan, A.B.: A generalized loss function for crowd counting and localization. In: Paper Presented at the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2021)

164. Xu, Y., et al.: Crowd counting with partial annotations in an image. In: Paper Presented at the 2021 IEEE/CVF International Conference on Computer Vision (ICCV) (2021)

165. Liu, W., Durasov, N., Fua, P.: Leveraging self-supervision for cross-domain crowd counting. In: Paper Presented at the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022)

166. Ma, Z., et al.: Towards A universal model for cross-dataset crowd counting. In: Paper Presented at the 2021 IEEE/CVF International Conference on Computer Vision (ICCV) (2021)

167. Song, Q., et al.: To choose or to fuse? Scale selection for crowd counting. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, pp. 2576–2583 (2021). https://doi.org/10.1609/aaai.v35i3.16360

168. He, Y., et al.: Error-aware density isomorphism reconstruction for unsupervised cross-domain crowd counting. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, pp. 1540–1548 (2021). https://doi.org/10.1609/aaai.v35i2.16245

169. Song, Q., et al.: Rethinking counting and localization in crowds: a purely point-based framework. In: Paper Presented at the 2021 IEEE/CVF International Conference on Computer Vision (ICCV) (2021)

170. Ma, Z., et al.: Learning to count via unbalanced optimal transport. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, pp. 2319–2327 (2021). https://doi.org/10.1609/aaai.v35i3.16332

171. Abousamra, S., et al.: Localization in the crowd with topological constraints. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, pp. 872–881 (2021). https://doi.org/10.1609/aaai.v35i2.16170

172. Dai, F., et al.: Dense scale network for crowd counting. In: Paper Presented at the Proceedings of the 2021 International Conference on Multimedia Retrieval (2021)

173. Khan, S.D., Basalamah, S.: Sparse to dense scale prediction for crowd couting in high density crowds. Arabian Journal for Science and Engineering 46(4), 3051–3065 (2021). https://doi.org/10.1007/s13369-020-04990-w

174. Khan, S.D., et al.: A deep-fusion network for crowd counting in high-density crowded scenes. International Journal of Computational Intelligence Systems 14(1), 168 (2021). https://doi.org/10.1007/s44196-021-00016-x

175. Chen, J., et al.: SSR-HEF: crowd counting with multiscale semantic refining and hard example focusing. IEEE Transactions on Industrial Informatics 18(10), 6547–6557 (2022). https://doi.org/10.1109/TII.2022.3160634

176. Wang, F., et al.: Hybrid attention network based on progressive embedding scale-context for crowd counting. Information Sciences 591, 306–318 (2022)

177. Wang, Q., Breckon, T.P.: Crowd counting via segmentation guided attention networks and curriculum loss. IEEE Transactions on Intelligent Transportation Systems 23(9), 15233–15243 (2022). https://doi.org/10.1109/TITS.2021.3138896
178. Liang, D., et al.: TransCrowd: weakly-supervised crowd counting with transformers. Science China Information Sciences 65(6), 160104 (2022). https://doi.org/10.1007/s11432-021-3445-y
179. Wang, M., et al.: STNet: scale tree network with multi-level auxiliator for crowd counting. IEEE Transactions on Multimedia, 1 (2022). https://doi.org/10.1109/TMM.2022.3142398
180. Zhong, X., et al.: An improved normed-deformable convolution for crowd counting. IEEE Signal Processing Letters 29, 1794–1798 (2022). https://doi.org/10.1109/LSP.2022.3198371
181. Liang, D., et al.: Focal inverse distance transform maps for crowd localization. IEEE Transactions on Multimedia, 1–13 (2022). https://doi.org/10.1109/TMM.2022.3203870
182. Mo, H., et al.: Attention-guided collaborative counting. IEEE Transactions on Image Processing 31, 6306–6319 (2022). https://doi.org/10.1109/TIP.2022.3207584
183. Dong, L., et al.: CLRNet: a cross locality relation network for crowd counting in videos. IEEE Transactions on Neural Networks and Learning Systems, 1–15 (2022). https://doi.org/10.1109/TNNLS.2022.3209918
184. Gu, C., et al.: HDNet: a hierarchically decoupled network for crowd counting. In: Paper Presented at the 2022 IEEE International Conference on Multimedia and Expo (ICME) (2022)
185. Zand, M., et al.: Multiscale crowd counting and localization by multitask point supervision. In: Paper Presented at the ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (2022)
186. Shaopeng, Y., Guo, W., Ren, Y.: CrowdFormer: an overlap patching vision transformer for top-down crowd counting. In: Paper Presented at the Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence (2022)
187. Gong, S., et al.: Bi-Level alignment for cross-domain crowd counting. In: Paper Presented at the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022)
188. Lin, H., et al.: Semi-supervised crowd counting via density agency. In: Paper Presented at the Proceedings of the 30th ACM International Conference on Multimedia (2022)
189. Liu, S., et al.: Harnessing perceptual adversarial patches for crowd counting. In: Paper Presented at the Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security (2022)
190. Cheng, Z.Q., et al.: Rethinking spatial invariance of convolutional networks for object counting. In: Paper Presented at the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022)
191. Shu, W., et al.: Crowd counting in the frequency domain. In: Paper Presented at the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022)
192. Xiong, H., Yao, A.: Discrete-constrained regression for local counting models. In: Paper Presented at the Computer Vision – ECCV 2022. Cham (2022)
193. Liang, D., Xu, W., Bai, X.: An end-to-end transformer model for crowd localization. In: Paper Presented at the Computer Vision – ECCV 2022. Cham (2022)
194. Wang, M., et al.: Dynamic mixture of counter network for location-agnostic crowd counting. In: Paper Presented at the Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) (2023)
195. Liu, W., Salzmann, M., Fua, P.: Estimating people flows to better count them in crowded scenes. In: Paper Presented at the ECCV 2020 : 16th European Conference on Computer Vision (2020). https://link.springer.com/chapter/10.1007/978-3-030-58555-6_43
196. Wang, B., et al.: Distribution matching for crowd counting. In: Paper Presented at the Proceedings of the 34th International Conference on Neural Information Processing Systems (2020). https://dl.acm.org/doi/10.5555/3495724.3495859
197. Zhou, Y., et al.: Adversarial learning for multiscale crowd counting under complex scenes. IEEE Transactions on Cybernetics 51(11), 1–10 (2020). https://doi.org/10.1109/TCYB.2019.2956091
198. Chen, Y., et al.: Region-aware network: model human's Top-Down visual perception mechanism for crowd counting. Neural Networks 148, 219–231 (2022). https://doi.org/10.1016/j.neunet.2022.01.015
199. Lin, H., et al.: Boosting crowd counting via multifaceted attention. In: Paper Presented at the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022)
200. Liu, W., et al.: SSD: single shot MultiBox detector. In: Paper Presented at the ECCV 2016: 14th European Conference on Computer Vision (2016). https://link.springer.com/chapter/10.1007/978-3-319-46448-0_2
201. Redmon, J., et al.: You only look once: unified, real-time object detection. In: Paper Presented at the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016). https://ieeexplore.ieee.org/document/7780460
202. Yang, Y., et al.: Embedding perspective analysis into multi-column convolutional neural network for crowd counting. IEEE Transactions on Image Processing 30, 1395–1407 (2021). https://doi.org/10.1109/TIP.2020.3043122
203. Tan, X., et al.: Crowd counting via multi-layer regression. In: Paper Presented at the Proceedings of the 27th ACM International Conference on Multimedia (2019). https://doi.org/10.1145/3343031.3350914
204. Jiang, X., et al.: Crowd counting and density estimation by trellis encoder-decoder networks. In: Paper Presented at the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2019). https://ieeexplore.ieee.org/document/8954254
205. Liu, N., et al.: ADCrowdNet: an attention-injective deformable convolutional network for crowd understanding. In: Paper Presented at the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
206. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Paper Presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2015)
207. Yilmaz, B., et al.: Perspective-aware loss function for crowd density estimation. In: Paper Presented at the 2019 16th International Conference on Machine Vision Applications (MVA) (2019). https://ieeexplore.ieee.org/document/8758034
208. Chen, G., Guo, P.: Enhanced Information Fusion Network for Crowd Counting (2021). arXiv e-prints. https://doi.org/10.48550/arXiv.2101.04279
209. Zou, Z., et al.: Enhanced 3D convolutional networks for crowd counting. In: Paper Presented at the BMVC 2019 : 30th British Machine Vision Conference (2019). https://bmvc2019.org/wp-content/uploads/papers/1082-paper.pdf
210. Yi, Q., et al. (2021). Scale-Aware Network with Regional and Semantic Attentions for Crowd Counting under Cluttered Background. arXiv e-prints, arXiv:2101.01479. https://doi.org/10.48550/arXiv.2101.01479
211. Mirza, M., Osindero, S.: Conditional generative adversarial nets. In: Paper Presented at the Proceedings of the 27th International Conference on Neural Information Processing Systems (2014). https://dl.acm.org/doi/10.5555/2969033.2969125
212. Girshick, R.: Fast R-CNN. In: Paper Presented at the 2015 IEEE International Conference on Computer Vision (ICCV) (2015). https://ieeexplore.ieee.org/document/7410526
213. Belagiannis, V., et al.: Robust optimization for deep regression. In: Paper Presented at the 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile (2015). https://ieeexplore.ieee.org/document/7410681
214. Zhu, L., et al.: Dual Path Multi-Scale Fusion Networks with Attention for Crowd Counting (2019). arXiv e-prints, abs/1902.01115

215. Sindagi, V.A., Patel, V.M.: Generating high-quality crowd density maps using contextual pyramid CNNs. In: Paper Presented at the 2017 IEEE International Conference on Computer Vision (ICCV) (2017). https://doi.org/10.1109/ICCV.2017.206

216. Zhou, W., et al.: Image quality assessment: from error visibility to structural similarity. IEEE Transactions on Image Processing 13(4), 600–612 (2004). https://doi.org/10.1109/TIP.2003.819861

217. Mittal, S.: The Role and Potential of CPUs in Deep Learning (2021). Retrieved from https://www.intel.com/content/www/us/en/developer/articles/technical/the-role-and-potential-of-cpus-in-deep-learning.html

218. Kaufman, S., et al.: A learned performance model for tensor processing units. In: Paper Presented at the Proceedings of Machine Learning and Systems 3 (MLSys 2021), Virtual-Only Conference (2021). https://proceedings.mlsys.org/paper/2021/hash/85d8ce590ad8981ca2c8286f79f59954-Abstract.html

219. Wang, Y.E., Wei, G.-Y., Brooks, D.M.: Benchmarking TPU, GPU, and CPU Platforms for Deep Learning (2019). arXiv:1907.10701 https://arxiv.org/abs/1907.10701

220. Shawahna, A., Sait, S.M., El-Maleh, A.: FPGA-based accelerators of deep learning networks for learning and classification: a review. IEEE Access 7, 7823–7859 (2019). https://doi.org/10.1109/ACCESS.2018.2890150

221. Deierling, K.: What is a DPU? Retrieved from https://blogs.nvidia.com/blog/2020/05/20/whats-a-dpu-data-processing-unit/ (2020)

222. Maillard, P., et al.: Radiation tolerant deep learning processor unit (DPU) based platform using xilinx 20nm kintex UltraScale™ FPGA. IEEE Transactions on Nuclear Science 70(4), 1–721 (2022). https://doi.org/10.1109/TNS.2022.3216360

How to cite this article: Deng, L., et al.: Deep learning in crowd counting: a survey. CAAI Trans. Intell. Technol. 1–35 (2023). https://doi.org/10.1049/cit2.12241