Neurocomputing 508 (2022) 1–18

Contents lists available at ScienceDirect

Neurocomputing

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

Survey paper

A survey on deep learning-based single image crowd counting: Network design, loss function and supervisory signal

Haoyue Bai [a], Jiageng Mao [b], S.-H. Gary Chan [a,*]

[a] The Hong Kong University of Science and Technology, Hong Kong
[b] The Chinese University of Hong Kong, Hong Kong

ARTICLE INFO

Article history:
Received 14 February 2022
Revised 11 July 2022
Accepted 3 August 2022
Available online 8 August 2022
Communicated by Zidong Wang

Keywords:
Crowd counting
Network design
Loss function
Supervisory signal

ABSTRACT

Single image crowd counting is a challenging computer vision problem with wide applications in public safety, city planning, traffic management, etc. With the recent development of deep learning techniques, crowd counting has aroused much attention and achieved great success in recent years. This survey is to provide a comprehensive summary of recent advances on deep learning-based crowd counting techniques via density map estimation by systematically reviewing and summarizing more than 200 works in the area since 2015. Our goals are to provide an up-to-date review of recent approaches, and educate new researchers in this field the design principles and trade-offs. After presenting publicly available datasets and evaluation metrics, we review the recent advances with detailed comparisons on three major design modules for crowd counting: deep neural network designs, loss functions, and supervisory signals. We study and compare the approaches using the public datasets and evaluation metrics. We conclude the survey with some future directions.

1. Introduction

Single image crowd counting is to estimate the number of objects (people, cars, cells, etc.) in an image of an unconstrained scene, i.e., an image without any restriction on the scene. Crowd counting has attracted much attention in recent years due to its important applications in public safety, traffic management, consumer behavior, cell counting, etc. [131,73,12]. In this survey, we mainly focus on people as the crowd, though the techniques discussed may be extended to other domains.

Due to the importance of crowd counting, extensive research have been done in the area, especially with the use of deep learning, which has demonstrated superior performances on various applications, such as computer vision [50,117,118], image classification [69], and multi-dimensional time series [5]. Deep learning achieves success for single image crowd counting with large-scale publicly available benchmarks [60,185] in recent years. This may be due to its data-driven properties [228,80] and capability of self-learning from raw data [103,148] for deep learning-based methods. In this work, we mainly discuss recent advanced deep learning-based single image crowd counting approaches due to its superiority in comparison to machine learning models.

Early approaches to count people are based on detection-based computer vision techniques, which are to detect individual objects, heads, or body parts and then count the total number in the image [135,86,76]. However, its accuracy deteriorates quickly for crowded scenes where objects have severe occlusions. To overcome it, the regression-based approach has been recently proposed, which directly estimates the count by relating it with the image. While achieving higher accuracy than the detection-based approach for crowded scenes, it lacks adequate spatial information of the people and is less interpretable [14,177,13], hindering its extension to localization study.

Most recently, crowd counting via density map estimation has emerged as a promising approach with encouraging results, where the input image is processed to a crowd density map, which is simply integrated to obtain the number of people in a pixel of the image [73,133,7,11,228,97,80,58]. Such approaches achieve high accuracy for crowded scenes and preserve spatial information of people distribution. Besides, there are some emerging approaches such as S-DCNet [186] which classifies the features into a predefined count range for crowd estimation.

We summarize by comparing the four major crowd counting approaches in Table 1. All of them require image annotation through labeling in the training step. For detection-based approach, each object has to be fully identified and outlined, which incurs the highest labeling cost. On the other hand,

[*] Corresponding author.

https://doi.org/10.1016/j.neucom.2022.08.037
0925-2312/© 2022 Published by Elsevier B.V.

Table 1
Summary of crowd counting approaches on four major categories: detection-based, regression-based, density map estimation, and emerging approaches.

Category	Principles	Crowd Counting Accuracy	Location Accuracy	Annotation Complexity	Limitations	Examples
Detection based	Detect then count; early approach	Low	High	High (object framing)	Low accuracy for highly owded scenes	[135,86,76]
Regression based	Directly learn rightarrow regress the count	Medium	N/A	Low (image-level)	Less interpretable; lacks location information	[14,177,13]
Density map estimation	Compute number of people per pixel	High	Medium	Medium (head indication)	Low accuracy in low crowd scenes	[73,133,11,228,97,80]
Emerging approaches	Classify the features into a predefined count range	High	Low	Medium (head indication)	Not flexible rightarrow wide count range	[205]

regression-based approach does not need to annotate individual objects but the total object count, and hence its annotation cost is the lowest. Density estimation has an intermediate labeling cost between the two because only the heads of the people need to be indicated.

We focus in this survey on crowd counting via density map estimation. With the development of deep learning approaches in the field of computer vision, counting accuracy has been greatly improved with the use of deep learning-based models as compared with approaches based on handcrafted features. We overview in Fig. 2 (a) the major design components for CNN-based crowd counting via density map estimation. An input image of a crowd scene is fed into a deep neural network which estimates the density map of the image (the upper branch). Here the critical issue is the *network design* so that the sum of the density value in all the pixels closely matches with the crowd count in the input image. For training (the lower branch), an image is first annotated with *supervisory signal*, which may range from fully to pseudo labeled, to generate the ground truth (given by the number of people per pixel in the image). The ground truth is used to adjust the node parameters of the deep neural network through minimizing a *loss function* between the network-generated density map and the ground truth.

We present recent advances on deep learning-based crowd counting. Our goals are to educate the new researchers state-of-the-arts and equip them with insights, tools, and principles to design novel networks. We survey and compare the available datasets, performance metrics, network design, loss function, and supervisory signal. Our survey is timely and unique.

We discuss related work as follows (see Table 2). Teixeira et al. [167] is an early survey on human sensing. However, it has not focused on crowd scene analysis. Li et al. [77] reviews crowd scene analysis in terms of crowd behavior, activity analysis, and anomaly detection, with crowd counting playing a small role. Ahuja et al. [2] covers different crowd estimation methods. Though Zitouni et al. [239] evaluate different crowd analysis methods, is not mainly for CNN-based approach via density map estimation, which has become the mainstream for crowd counting in recent years. Chrysler et al. [26] discusses the methods to tackle the challenges of the lack of training data, perspective distortion faced by the crowd counting system. The work [158] surveys on CNN-based approach for a single image, but it only roughly discussed the recent advances on CNN-based methods. It has not discussed the advanced convolutional operations and attention-based model, loss function, and supervisory signal, and only up to the year 2017.

In contrast with previous papers, our work comprehensively summarizes more than two hundred deep learning-based crowd counting algorithms in the recent five years. Our work is of current interest and value, because it is more comprehensive, summarizing the more recent, popular, and critical design components of this active field and provide an in-depth illustration of the representative schemes in the area. Through this survey, we expect to offer an up-to-date summary of recent advances in this field and educate new researchers on the design principles and trade-offs.

Table 2
A comprehensive analysis of other counting related survey papers. Compared with previous related works, our work is of current interest and value, because it is timely, more comprehensive and provide an in-depth analysis of the representative approaches in this active area.

Paper	Year	Venue	Comparison of Other Crowd Counting Surveys
Approaches on Crowd Counting and Density Estimation: A Review [74]	2021	PAA	This paper focus on elaborating deep learning-based counting methods, which is board-based and mainly focus on the network design considerations without discussing loss functions and supervisory signals.
A Literature Review of Crowd-counting System on Convolutional Neural Network [26]	2021	IOPCS	This survey discusses the challenges faced by crowd counting systems and focuses on developing a more robust crowd counting methodology. However, this survey is a short paper. The network design discussion misses some important recent approaches such as DM-Count, SASNet. It also lacks unsupervised learning counting approaches.
A Survey of Recent Advances in Crowd Density Estimation using Image Processing [2]	2019	ICCES	This is a short paper, which mainly discusses the traditional approaches with hand-crafted features. Deep learning-based approaches only play a small part.
A Survey of Techniques for Automatically Sensing the Behavior of a Crowd [33]	2018	ACMCS	This paper surveys practical solutions for sensing pedestrian behavior, and also combining privacy, transparency, scalability, and ease of deployment. However, this paper is for traditional methods with hand-crafted features.
A Survey of Recent Advances in CNN-based Single Image Crowd Counting and Density Estimation [158]	2018	PRL	This paper surveys CNN-based crowd counting approaches for a single image, but it only roughly discussed the recent advances on CNN-based methods. It has not discussed the advanced convolutional operations and attention-based model, loss function and supervisory signal, and only up to the year 2017.-
Crowded Scene Analysis: A Survey [77]	2014	T-CSVT	This paper reviews crowd scene analysis in terms of crowd behavior, activity analysis, and anomaly detection, with crowd counting playing a small role.
Advances and Trends in Visual Crowd Analysis: A Systematic Survey and Evaluation of Crowd Modelling Techniques [239]	2016	NC	This paper evaluates different crowd analysis methods, is not mainly for CNN-based approach via density map estimation, which has become the mainstream for crowd counting in recent years.
A Survey of Human-Sensing: Methods for Detecting, Presence, Count, Location, Track, and Identity [167]	2010	CS	An early survey on human sensing. However, it has not focused on crowd scene analysis but on the study of presence, count, location, track, and identification.

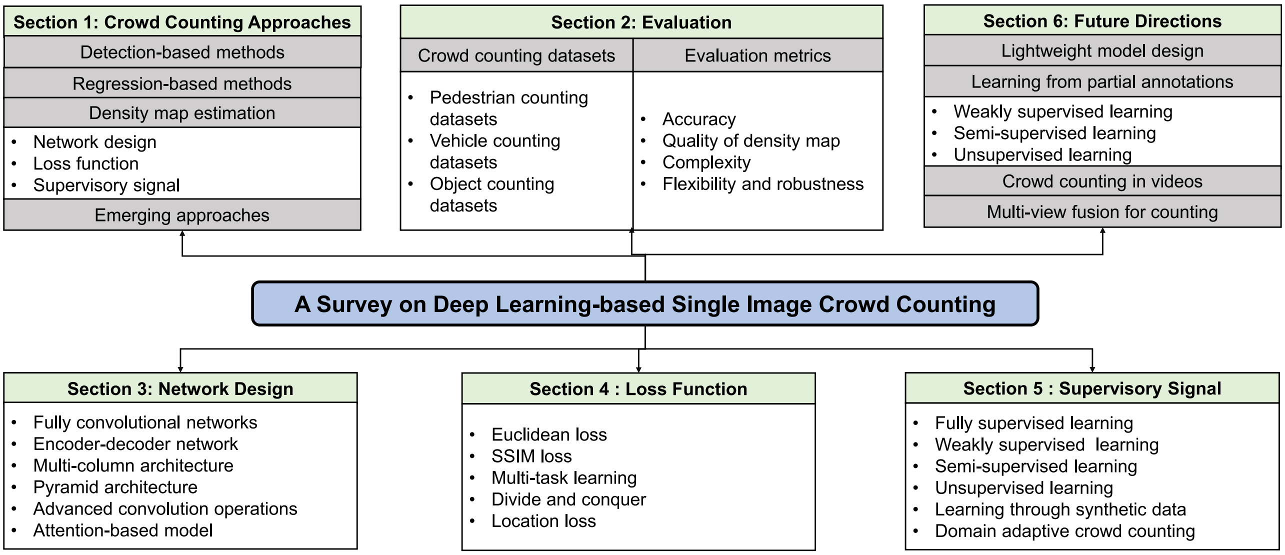

Section 1: Crowd Counting Approaches
Detection-based methods
Regression-based methods
Density map estimation
• Network design
• Loss function
• Supervisory signal
Emerging approaches
Section 2: Evaluation
Crowd counting datasets
Evaluation metrics
• Pedestrian counting datasets
• Vehicle counting datasets
• Object counting datasets
• Accuracy
• Quality of density map
• Complexity
• Flexibility and robustness
Section 6: Future Directions
Lightweight model design
Learning from partial annotations
• Weakly supervised learning
• Semi-supervised learning
• Unsupervised learning
Crowd counting in videos
Multi-view fusion for counting
A Survey on Deep Learning-based Single Image Crowd Counting
Section 3: Network Design
• Fully convolutional networks
• Encoder-decoder network
• Multi-column architecture
• Pyramid architecture
• Advanced convolution operations
• Attention-based model
Section 4 : Loss Function
• Euclidean loss
• SSIM loss
• Multi-task learning
• Divide and conquer
• Location loss
Section 5 : Supervisory Signal
• Fully supervised learning
• Weakly supervised learning
• Semi-supervised learning
• Unsupervised learning
• Learning through synthetic data
• Domain adaptive crowd counting

Fig. 1. The structure of this survey. First, we overview the four main categories of deep learning-based crowd counting methods. Second, we present publicly available counting datasets and evaluation metrics. Then, we review recent advances on crowd counting schemes, which is mainly pertain to deep neural network design, loss function and supervisory signal. We conclude the survey with future directions.

Fig. 1 shows the main design components for crowd counting we will discuss in this paper. For network design, we describe the basic principles of major techniques such as fully convolutional network, encoder-decoder architecture, multi-column, and pyramid network, etc. For loss function, we discuss the widely used Euclidean loss and the recently advanced schemes such as SSIM loss, and multi-task learning. For supervisory signal, we introduce different ground truth generation methods for fully supervised setting and compare it with weakly supervised and semi-supervised learning, and self-supervised learning, and automatic labeling through synthetic data. Typical representative schemes are summarized and compared in each section.

The rest of the paper is organized as follows. In Section 2, we summarize the publicly available crowd counting datasets, evaluation metrics, and design considerations. We present in Section 3 the details of deep neural network design. Section 4 discusses the loss functions, and Section 5 reviews supervisory signal to train crowd counting network. We conclude with future directions in Section 6.

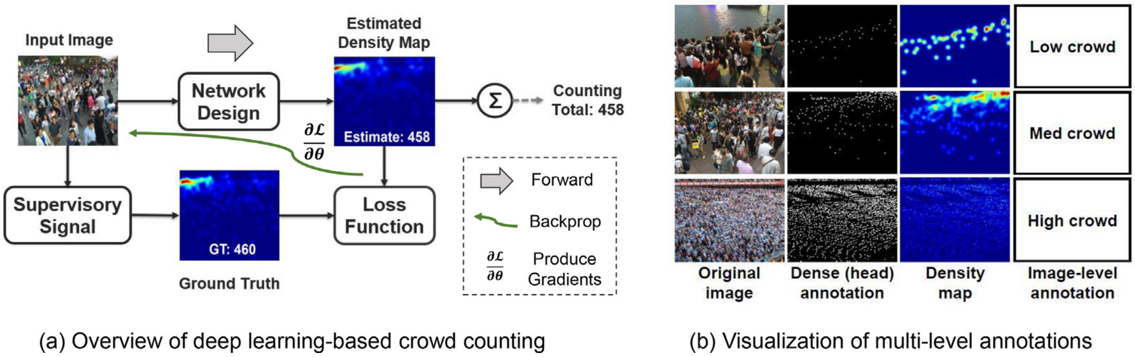
2. Datasets and Performance Evaluation

In this section, we first summarize the most widely used crowd counting datasets in Section 2.1. Then we discuss the design considerations and performance metrics to study crowd counting in Section 2.2.

2.1. Datasets

Public datasets are used as benchmarks to evaluate crowd counting models. In choosing a dataset, the following metrics are often considered:

- *Image resolution:* Datasets with high resolutions usually show better visual quality. Furthermore, due to their higher pixel density, they often achieve higher count accuracy.
- *Number of images:* Datasets with a large number of images often cover more diverse scenes, backgrounds, view angles, and lighting conditions. Large and diverse datasets are beneficial to optimize deep learning-based models and mitigate over-fitting problems.
- *Object count:* The number of objects in a dataset is an important consideration for crowd analysis. The minimum, maximum, and average counts shed light on crowd density in the dataset. Datasets with a large crowd density level coverage and the number of objects is usually more challenging for crowd counting.



(a) Overview of deep learning-based crowd counting (b) Visualization of multi-level annotations

Fig. 2. Overview of deep learning-based single image crowd counting methods via density map estimation. Figure (a) shows the major components for deep learning-based crowd counting via density estimation. Figure (b) presents visualization of original image, labor-intensive dense annotation, ground truth density maps, and image-level weak annotation. The annotation paradigms are from [159].

We identify some common datasets used in the research community including pedestrian counting and object datasets, and extract and present some typical images from the datasets in Fig. 3. There are also some other works focus on counting from remote scenes [237,34,230,38,121] and indoor crowd counting [87]. We also compare these datasets in Table 3. These datasets are elaborated below:

- *RGBT-CC* contains RGB-thermal data captured in different scenarios in urban scenes with various densities, e.g., malls, streets, playgrounds, train stations, etc. $1,013$ pairs are in light and $1,017$ pairs are in darkness. RGBT-CC is randomly divided into three sets: 1030 pairs for training, 200 pairs for validation, 800 pairs for testing.
- *NWPU-Crowd* [185] consists of $5,109$ images and $2,133,375$ annotated instances with points and boxes. Compared with other real-world crowd counting datasets, the NWPU-Crowd dataset has the largest density range of the annotated objects from 0 to $20,033$ per image. The average resolution of this dataset is 2191 × 3209, which is generally larger than other widely used 2D single image crowd counting datasets.
- *JHU-Crowd* [162] is collected under diverse scenarios, environmental, and weather conditions include images with weather-based degradations and illumination variations. This dataset contains a rich set of labels: blur-level, occlusion-level, size-level, and other image-level annotations.
- *CrowdSurveillance* [210] is a large scale crowd counting dataset with high-resolution images captured under challenging scenarios. The dataset is built by both online crawling and real-world surveillance video which covers more challenging scenarios with complicated backgrounds and varying crowd counts.
- *DroneCrowd* [198] is formed by 112 video clips with 33,600 high resolution frames with large variations in scale, viewpoint, and background clutters, which captured under 70 different scenarios across 4 cities. The video clips are recorded at 25 frames per seconds with 1920 × 1080 resolution. This dataset also provides 20,800 people trajectories with head annotations and several video-level attributes in sequences, i.e., illumination, altitude, and density. DroneCrowd is divided into training and testing sets, with 82 and 30 video sequences respectively.
- *UCF-QNRF* [60] contains 1,535 challenging images and a total of 1,251,642 annotations. The minimum and the maximum number of objects within an image are 49 and 12,865. The training and testing sets are selected by sorting the images according to the counts and picking every 5th image as the test set (1201 images for training and 334 images for testing). Besides, this large-scale dataset covers different locations, viewpoints, perspective effects, and different times of the day.
- *GCC* [148] [187] is a large-scale diverse synthetic crowd dataset, which was generated based on a computer game, Grand Theft Auto V. GTA V Crowd Counting (GCC) dataset consists of

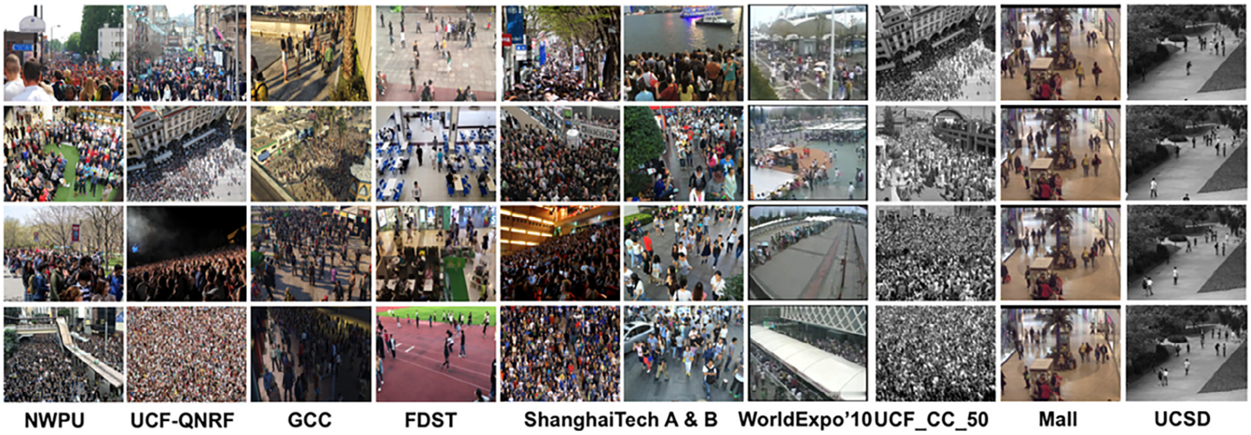


Fig. 3. Some typical crowd scene images of publicly available datasets. Different columns represents different crowd counting datasets and we visualize four typical images for each dataset. The NWPU [185], UCF-QNRF [60], ShanghaiTech A & B [228], WorldExpo'10 [221], and UCF_CC_50 [59] are image-based datasets. The FDST [37], Mall [17], and UCSD [12] are video-based datasets. The GCC [148] is a diverse synthetic crowd dataset.

Table 3
An overview of datasets statistics for crowd counting. **Image Number** is the number of images; **Total** is total number of labeled objects; **Min Count** is the minimal crowd count; **Max Count** is the maximum crowd count; **Ave Count** is the average crowd count.

Category	Dataset	Year	Average Resolution	Image Number	Total	Min Count	Max Count	Avg Count
Pedestrian Counting	RGBT [92]	2021	640×480	2,030	138,389	-	-	68
	NWPU-Crowd [185]	2020	2191×3209	5,109	2,133,375	0	20,033	418
	JHU-Crowd [162]	2019	1450×900	4250	1,114,785	0	7286	262
	Crowd Surveillance [210]	2019	1342×840	13,945	386,513	-	-	35
	DroneCrowd [198]	2019	1920×1080	33,600	4,864,280	25	455	145
	UCF-QNRF [60]	2019	2013×2902	1,535	1,251,642	49	12,865	815
	GCC [148]	2019	1080×1920	15,212	7,625,843	0	3,995	501
	Fudan-ST [37]	2019	1080×1920	15,000	394,081	9	57	27
	ST Part A [228]	2016	589×868	482	241,677	33	3,139	501
	ST Part B [228]	2016	768×1024	716	88,488	9	578	124
	WorldExpo'10 [221]	2015	576 × 720	3,980	199,923	1	253	50
	UCF_CC_50 [59]	2013	2101×2888	50	63,974	94	4,543	1,280
	Mall [17]	2012	240×320	2,000	62,325	13	53	31
	UCSD [12]	2008	158×238	2,000	49,885	11	46	25
	VisDrone Vehicle [238]	2019	991×1511	5303	198,984	10	349	38
	Penguin [4]	2016	700×700	8200	72160	-	5	8.8
	TRANCOS [45]	2015	640×480	1244	46796	-	-	38

15,212 images, with a resolution of 1080 × 1920, containing more than 7,625,843 people annotation. GCC is more diverse than other real-world datasets. It captures 400 different crowd scenes in the GTA C game, which includes multiple types of locations.

- *Fudan-ShanghaiTech* [37] contains 100 videos captured from 13 different scenes. FDST includes 150,000 frames and 394,081 annotated heads, which is larger than previous video crowd counting datasets in terms of frames. The training set of the FDST dataset consists of 60 videos, 9000 frames, and the testing set contains the remaining 40 videos, 6000 frames. The number of frames per second (FPS) for FDST is 30.
- *ShanghaiTech A & B* [228] consists of two parts: Part A and Part B, which contains 482 images (300 images for training, 182 images for testing), and 716 images (400 images for training, 316 images for testing), respectively. Part A includes high-density crowds that are collected from the Internet. Part B is captured from the busy streets of urban areas in Shanghai, which are less crowded than the scenes from Part A.
- *WorldExpo'10* [221] focus on cross-scene counting. It consists of 1132 video sequences captured by 108 surveillance cameras during the Shanghai 2010 WorldExpo. WorldExpo'10 dataset is randomly selected from the video sequences, which has 3,980 frames with 199,923 head annotations. The training set of WorldExpo'10 contains 3,380 frames from 103 scenes, and the remaining 600 frames are sampled from five other different scenes with each scene being 120 frames for testing.
- *UCF_CC_50* [59] has 50 black and white crowd images and 63974 annotations, with the object counts ranging from 94 to 4543 and an average of 1280. The original average resolution of the dataset is 2101 × 2888. This challenging dataset is crawled from the Internet. For experiments, UCF_CC_50 were divided into 5 subsets and performed fivefold cross-validation. The maximum resolution was reduced to 1024 for efficient computation.
- *Mall* [17] was captured by a public surveillance camera in a shopping mall, which contains more challenging lighting conditions and more severe perspective distortion than the UCSD dataset [12]. The Mall dataset consists of 2000 video frames with fixed resolution (320 × 240) and 62,325 total pedestrian instances. The first 800 frames were used for training and the remaining 1200 frames for testing.
- *UCSD* [12] consists of an hour of video with 2000 annotated frames and in a total of 49,885 pedestrian instances, which was captured from a pedestrian walkway of the UCSD campus by a stationary camera. The original video was recorded at 30fps with a frame size of 480 × 740 and later downsampled to 10fps with dimension 158 × 238. The 601–1400 frames were used for training and the remaining 1200 frames for testing. The ROI of the walkway and the traveling direction are also provided.
- *VisDrone Vehicle* [7] is modified from the original VisDrone2019 detection dataset [238] with bounding boxes of targets to crowd counting annotations. The new vehicle annotation location is the center point of the original bounding box. This dataset consists of 3953 training samples, 364 validation samples, and 986 test samples.
- *Penguin* [4] is a large and challenging dataset of penguins in the wild with high-degree of object occlusion and scale variation. The collected images are compounded by many factors, e.g., adversarial weather conditions, variability of vantage points of the cameras, extreme crowding, and inter-occlusion between penguins. The Penguin dataset is divided into two subsets for 70% and 30% of the total images respectively.
- *TRANCOS* [45] is a vehicle crowd counting dataset which is to estimate the number of vehicles in an image of a traffic congestion situation. TRANCOS consists of 1244 traffic jam images with 46796 annotated vehicles. All the collected images contain traffic congestions with a variety of different scenes and viewpoints, covering different lighting conditions, different levels of overlap, and crowdedness. This dataset is divided into three parts: 403 images for training, 420 images for validation, and 421 images for testing.

2.2. Performance Evaluation and Metrics

In evaluating crowd counting networks, the following performance metrics are often used:

- *Accuracy:* Accuracy refers to counting accuracy and location accuracy.

  -Counting accuracy is affected by scale variation and isolated clusters of objects [7]. Scale variation means the same object would appear as a different size in an image due to its perspective and distance from the camera. Besides, an image may have isolated object clusters, and models properly capturing such contextual information usually perform better than others. To quantitatively evaluate counting accuracy, Mean Absolute Error (MAE), Mean Squared Error (MSE) and mean Normalized Absolute Error (NAE) are commonly used, defined respectively as: $MAE = \frac{1}{N}\sum_{i=1}^{N}|C_i - \widehat{C}_i|$, $MSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}|C_i - \widehat{C}_i|^2}$, $NAE = \frac{1}{N}\sum_{i=1}^{N}\frac{|C_i - \widehat{C}_i|}{C_i}$, where N is total number of test images, C_i the ground truth of the i-th image, and $\widehat{C}_i$ the estimated count.

  -Location accuracy is related to the spatial information preserved in the density map. Models with higher quality density map generated usually contains more spatial information for localization tasks.
- *Quality of density map:* Density map can be evaluated in terms of resolution and visual quality.

  -High-resolution density maps usually show better location accuracy and preserve more spatial information for localization tasks (e.g., detection and tracking).

  -To quantitatively evaluate the visual quality of the generated density maps, Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity in Images (SSIM) [194].
- *Complexity:* Complexity consists of computational complexity and annotation complexity.

  -Computational complexity is evaluated based on measures such as the number of model parameters, floating-point operations (FLOPs), and inference time.

  -Annotation complexity, as shown in Table 1, refers to data labeling cost. In general, object-level annotation as conducted in the detection-based approach has high complexity. Density map estimation requires point-level (head) annotation, which is relatively less costly. If unlabeled or synthetic data are used, the complexity can be further reduced.
- *Flexibility and robustness:*

  -The flexibility of models is evaluated based on the sensitivity of processing images with arbitrary sizes and the ability to model different kinds of objects (e.g., non-rigid objects).

  -Robustness refers to distribution shift robustness. It is evaluated in terms of out-of-distribution accuracy, where the test data come from another distribution (w.r.t. the training one).

3. Deep Neural Network Design

Network design is one of the most important parts for density map estimation. In this section, we present the major deep networks for crowd counting: fully convolutional networks

(Section 3.1), encoder-decoder architecture (Section 3.2), multi-column network (Section 3.3), pyramid structure (Section 3.4), advanced operations (Section 3.5), attention-based model (Section 3.6), vision transformer (Section 3.7), and neural architecture search (Section 3.8). We compare these approaches in Section 3.9, and remark on some other emerging approaches in Section 3.10.

3.1. Fully Convolutional Network

An early CNN-based density map estimation approach is based on a fully convolutional network (FCN) [119], which is modified from the existing CNN architecture (VGG16) and replaces all the fully-connected layers with fully convolutional layers in order to analyze images of arbitrary sizes. As shown in Fig. 4 (a), FCN learns an end-to-end mapping from an input image to the corresponding density map and produces a proportionally sized density map output gave the input image. The FCN structure is simple but accurate, which has been widely used.

However, the FCN crowd counting method has some limitations. The resolution of the generated density map is only $1/4$ of the input width and $1/4$ of the input height due to the max pooling operations (extract high-level features but reduce resolutions) in FCN, which lacks fine details and spatial information for localization tasks, compared with high-resolution density maps. Besides, the FCN crowd counting model is susceptible to scale variation problems in crowd scene images, which limits its applicability in the general environment.

3.2. Encoder-Decoder Architecture

The Encoder-decoder model is proposed to align the resolution of the produced density map with the input image. As shown in Fig. 4 (b), the encoder-decoder network consists an encoder and a decoder: an encoder network takes the input image and output high-level features, which hold the information and represents the input; a decoder network takes the features from the encoder and generate high-resolution density map. The encoder gradually downsamples the image resolution with convolutional or pooling layers, and the decoder progressively upsamples the feature maps from the encoder with deconvolutional layers or interpolations. The skip connections are applied on the feature maps from the encoder and decoder respectively.

Some of the deep learning-based crowd counting approaches are following the encoder-decoder structure in recent years (see, for examples, [220,62,11,96,164,20,168,29]). SANet [11] proposed a novel encoder-decoder network, called scale aggregation Network, which achieves accurate and efficient crowd estimation. The decoder generates high-resolution density maps with a set of transposed convolutions. Furthermore, encoder-decoder based architecture can significantly reduce the number of parameters compared with other architectures due to the downsample operations in the encoder. However, such architecture has not addressed the scale variation problem and has not considered the local and global contextual information.

3.3. Multi-Column Network

Multi-column and pyramid network is the most prominent models in recent crowd counting algorithms to extract the multi-scale features and tackle the scale variation problem [75,201,202,99,212,28,215,181].

The multi-column architecture incorporates multi-column architecture with different kernel sizes to extract different scale features in order to achieve accurate counting accuracy such as MCNN [228] and McML [25]. As shown in Fig. 4 (c), multi-column neural network (MCNN) consists of multiple branches with different kernel sizes (e.g., $5 \times 5, 7 \times 7$ and 9×9). The different branches accommodate different receptive fields, thus sensitive to multi-scale features. Finally, the features extracted by different columns are fused together to generate density maps. However, the accommodated scale diversity is restricted by the number of columns.

3.4. Pyramid Structure

Image pyramid and feature pyramid architectures are yet another approach to address scale variations (e.g., AFP [66], CP-CNN [157,3,206]), which mainly consists of two subgroups, image pyramid, and feature pyramid pooling. For the image pyramid-based model, as Fig. 4 (d) shows, different scale of the image pyramid (scale 1, ..., scale S) is feed into an FCN to predict the density map of that scale. Then, the final estimation is produced by adap-

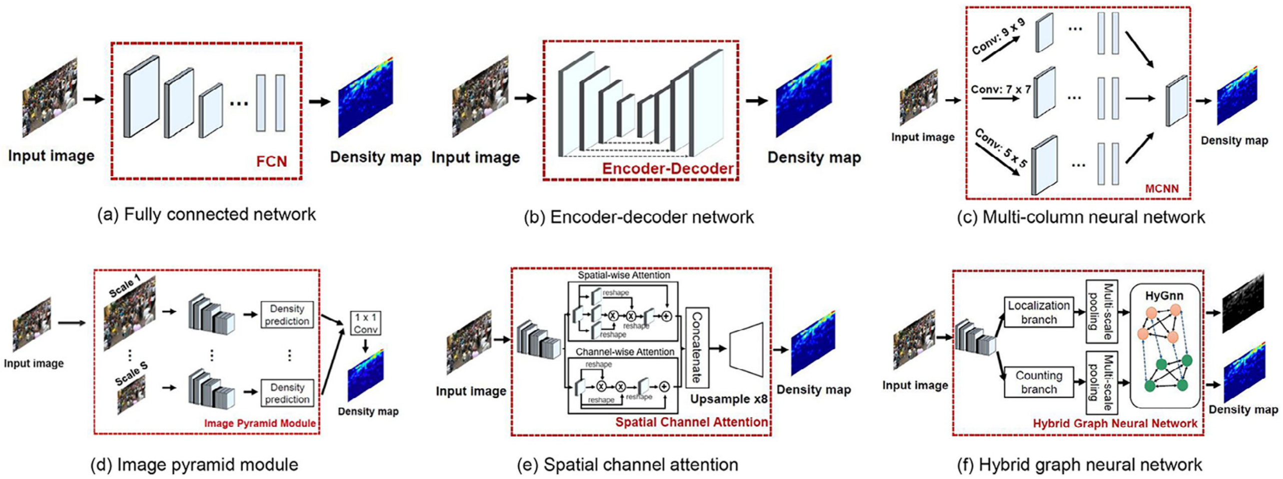


Fig. 4. A summary of the diverse range of network architectures used for deep learning-based single image crowd counting: (a) fully connected network [119]; (b) encoder-decoder architecture [11]; (c) multi-column network [228]; (d) pyramid structure [66]; (e) attention-based model [42]; (f) graph neural network [110]. The order of the networks according their presentation in this paper. (Better viewed in the zoom-in mode).

tive fusing the prediction from different scales. However, this kind of architecture remains a high computational complexity.

Besides, some relevant techniques are usually used together with the multi-column and pyramid networks to enhance the multi-scale feature extraction process such as skip-connections [162,160,195,30,120,108] and dense blocks [126,134,111,63,60,27].

3.5. Advanced Convolution Operations

There is a trend to leverage advanced convolutional operations to facilitate accurate crowd counting models and better CNN feature learning [233,207,56]. The deep learning-based single image crowd counting model benefits a lot from the advanced convolution such as dilated and deformable convolution, adaptive dilated convolution, and perspective-guided convolution. This can replace the traditional convolutional operations in the counting models.

There are four important advanced operations:

- *Dilated convolution* introduces the dilated rate to the convolutional layers, which defines a spacing between the weights of the kernel. Traditional convolutional operation is more focused on extracting local features. For the dilated convolution, three subfigures represent dilated operations with the same kernel size (3×3) but different dilated rates (Dilation = 1, Dilation = 2, and Dilation = 3), which enlarges the receptive field without increasing the computational cost and also preserves the resolution of the feature maps. Dilated convolution facilitate real-time applications and is popular in many recent crowd counting models: Dynamic Region Division (DRD) [49], Scale Pyramid Network (SPN) [19], Atrous convolutions spatial pyramid network (ACSPNet) [111], DENet [93], Dilated Convolutional Neural Networks (CSRNet) [80] and An Aggregated Multicolumn Dilated Convolution Network (AMDCNet) [28]. But this kind of operations not consider the multi-scale features and cannot fully capture the non-rigid objects.
- *Deformable convolution* is a kind of spatial sampling location augmenting schemes in the modules with additional offsets and learning the offsets from the target tasks, without additional supervision. This can model non-rigid objects with additional learnable offsets. Some recent literatures replace the traditional convolutions with the deformable convolutions and achieves superiors performance: Dilated-Attention-Deformable ConvNet (DADNet) [46], An Attention-injective Deformable Convolutional Network (ADCrowdNet) [97]. However, the deformable convolutional operations require high computational complexity.
- *Adaptive dilated convolution* is formed to predicts a continuous value of dilation rate for each location in order to effectively match the scale variation at different locations, which is better than fixed and discrete dilate rates. ADSCNet [8] is formulated based on adaptive dilated convolution, which is also able to preserve the strong consistency between the density and feature of each location.
- *Perspective-guided convolution* aims to tackle the continuous scale variation issue with perspective information. The perspective information contains instance information between camera and a scene, which is a reasonable prior for people scale estimation. Concretely, the perspective information functions areleveraged to guide the spatially variant smoothing of feature maps before feeding to the successive convolutions. PGCNet [210] is built by stacking multiple Perspective-guided convolutions (PGC) blocks based on a CNN backbone, which is a single-column CNN target to tackle the scale variation issues with a moderate increase in computation.

3.6. Attention-based Model

Attention mechanisms can be roughly divided into two subgroups: hard attention and soft attention [85,178,184,203,145,64,35,54,16,18]. Such mechanisms have been explicitly explored in recent years, and we summarize several recent algorithms applied with the attention mechanism: AFPNet [66], MRA-CNN [227], SAAN [52], DADNet [46], Relational Attention Network [219], Hierarchical Scale Recalibration Network [241], ACM-CNN [240], HA-CNN [159], Shallow Feature-based Dense Attention Network [123] and Multi-supervised Parallel Network [196].

SCAR [42] is one of the typical models to make use of attention schemes. SCAR proposes a spatial-/channel-wise attention regression module for crowd counting. As shown in Fig. 4 (e), the top half branch (spatial-wise attention) captures large-range contextual information and the change of density distribution, which the output feature map is weighted sum of attention map and original local feature map. The bottom half branch shows the channel-wise attention, which leverages both local and global contextual information for crowd counting. The features extracted by these two branches are late fused by concatenation and upsample post-processing to generate density maps. However, most of the methods discussed above are relying on pixel-wise loss functions for optimizing the model. We will discuss advanced loss functions to better capture spatial correlations between pixels and to generate high-quality density maps.

3.7. Vision Transformer

The mainstream crowd estimation approaches usually leverage the convolution neural network to extract features and significant progress has been achieved by incorporating larger context information into CNNs, which indicates that long-range context is essential. The self-attention mechanisms of transformers, which explicitly model all pairwise interactions between elements in a sequence, which is particularly suitable to extract the semantic crowd information.

TransCrowd [82] proposes two different kinds of approaches for single image crowd counting: TransCrowd-Token and TransCrowd-GAP, which can generate reasonable attention weight and achieve high counting performance.

3.8. Neural Architecture Search

Most of the recent advances in counting network design are based on hand-designed neural networks, which require large design efforts and strong domain knowledge. To extract multi-level features, convolutions with various receptive fields are designed by hand. Recently, automatic and lightweight network design has drawn much attention. Automated Machine Learning and Neural Architecture Search (NAS) techniques can be used to automatically design effective and efficient crowd counting architectures [193]. And the NAS-based approach is able to automatically discover the task-specific multi-scale crowd estimation models.

NAS-Count [55] automates the design of crowd counting models with NAS and proposes an end-to-end searched encoder-decoder architecture, where multi-scale features can be leveraged to tackle the scale variation problem. The first attempt in NAS needs hundreds of GPUs to run. However, NAS-Count leverage a differential one-shot search strategy to achieve fast search speed, where network parameters and architecture parameters are jointly optimized via gradient descent. In addition, NAS-Count is enabled by the compositional nature of CNN and is guided by task-specific search space and strategies. The architectures searched by the counting-oriented NAS framework achieve superior performance without demanding expert-involvement.

3.9. Comparisons

We compare the different networks discussed above in Table 4, and present their performance on three challenging crowd counting datasets in Table 5. We also provide a comprehensive performance analysis of state-of-the-art crowd counting approaches in Table 10. By analyzing the data, we find some intriguing observations.

As Tables 4,5 show, SANet achieves better counting performance on datasets with different crowd levels, compared with FCN. The generated density maps of FCN are only $1/4 \times 1/4$ of the original input image, which SANet is able to generate high-resolution density maps. The computational complexity for both the FCN and SANet is low (e.g., 0.91 M for SANet), which indicates that the encoder-decoder architecture is lightweight.

MCNN and CP-CNN consider scale variation problem, which is able to capture multi-scale features. MCNN extracts multi-scale features with multi-column architecture and CP-CNN extracts multi-scale features with pyramid architecture. CP-CNN achieves better counting accuracy and visual quality than MCNN, while for the computational complexity, the number of parameters for CP-CNN (68.4 M) is much larger than MCNN (0.13 M). This further demonstrates the effectiveness of multi-column architecture and pyramid architecture, while image pyramid architecture (e.g., CP-CNN) is of high computational complexity.

CSRNet and ADCrowdNet achieve better counting accuracy and visual quality than MCNN and CP-CNN on most of the datasets. CSRNet relies on dilated convolutional operations, which enlarge the receptive field without increase the computational cost. ADCrowdNet incorporates deformable convolutional operations, which are based on learnable additional offsets for better modeling non-rigid objects such as people. In addition, ADCrowdNet achieves better counting accuracy and visual quality than CSRNet but requires higher computational complexity.

SCAR shows better counting accuracy and visual quality than MCNN and CP-CNN, which is able to capture local and global contextual information based on spatial-wise attention and channel-wise attention schemes. The experimental results confirm the

Table 4
Comparisons of network design considerations for crowd counting. **Computational complexity** is evaluated based on the number of model parameters. The representative schemes of each network design category are analyzed thoroughly in terms of **advantages** and **limitations**.

Category		Representative Scheme	Advantages	Computational Complexity	Limitations
Fully convolution neural networks		FCN [119]	Can analyze images of arbitrary size	Low	Low-resolution density maps
Encoder-decoder architecture		SANet [11]	Able to generate high-resolution density maps	Low(0.9 M)	Not consider scale variation
Multi-column architecture		MCNN [228]	Extract multi-scale features with multi-column architecture	Low(0.1 M)	The scale diversity is restricted by the number of columns
Pyramid architecture		CP-CNN [157]	Extract multi-scale features with pyramid architecture	High(68.4 M)	High computational complexity
Advanced convolution operations	Dilated convolution operations	CSRNet [80]	Enlarge receptive field without increase the computational cost	Medium(16.3 M)	Not consider the non-rigid objects
	Deformable convolution operations	ADCrowdNet [97]	Learnable additional offsets for better modeling non-rigid objects	High	High computational complexity
	Adaptive dilated convolution	ADSCNet [8]	Learn continuous dilation rate	Medium	Not flexible rightarrow non-rigid objects
	Perspective- guided convolution	PGCNet [210]	Perspective information facilitate people scale estimation	Medium	Requires additional perspective information
Attention-based Model		SCAR [42]	Capture local and global contextual information	Medium	Rely on pixel-wise loss function
Vision Transformer		TransCrowd [82]	Able to modelrange context information	Medium	Computational expensive
Neural Architecture Search		NAS-Count [55]	Automate crowd counting model design	Medium	Computational expensive

Table 5
Quantitative comparisons of different network design considerations on widely used crowd counting datasets. The counting accuracy is evaluated based on **MAE** and **MSE**. The visual quality of the generated density maps is evaluated based on **PSNR** and **SSIM**. **ST PartA** and **ST partB** denotes ShanghaiTech A and ShanghaiTech B dataset [228], respectively.

Representative Schemes			ST PartA				ST PartB		UCF_CC_50	
Methods	Year	Column	MAE	MSE	PSNR	SSIM	MAE	MSE	MAE	MSE
FCN [119]	2016	Single	126.5	173.5	-	-	23.8	33.1	338.6	424.5
MCNN [228]	2016	Multi	110.2	173.2	21.40	0.52	26.4	41.3	377.6	509.1
CP-CNN [157]	2017	Multi	73.6	106.4	21.72	0.72	20.1	30.1	295.8	320.9
SANet [11]	2018	Single	67.0	104.5	-	-	8.4	13.6	258.4	334.9
CSRNet [80]	2018	Single	68.2	115.0	23.79	0.76	10.6	16.0	266.1	397.5
ADCrowd [97]	2019	Single	63.2	98.9	24.48	0.88	8.2	15.7	266.4	358.0
PGCNet [210]	2019	Single	57.0	86.0	-	-	8.8	13.7	244.6	361.2
SCAR [42]	2019	Double	66.3	114.1	23.93	0.81	9.5	15.2	259.0	374.0
ADSCNet [8]	2020	Single	60.7	100.6	-	-	6.4	11.3	198.4	267.3
MS-GAN [236]	2020	Single	-	-	-	-	18.7	30.5	345.7	418.3
HyGnn [110]	2020	Double	60.2	94.5	-	-	7.5	12.7	184.4	270.1
TransCrowd [82]	2021	Single	66.1	105.1	-	-	9.3	16.1	-	-
NAS-Count [55]	2021	Single	56.7	93.4	-	-	6.7	10.2	208.4	297.3
STNet [180]	2022	Single	52.9	83.6	-	-	6.3	10.3	162.0	230.4

effectiveness of attention mechanism variations for crowd counting. HyGnn shows good counting performance on different crowd counting datasets, which demonstrates the effectiveness of graph-based models to distill rich relations among multi-scale features for crowd counting.

The multi-path encoder-decoder network searched by NAS-Count demonstrates better performance than tedious hand-designing crowd counting models on four challenging datasets, which achieves a multi-scale model automatically without strong domain knowledge. This clearly demonstrates the potential to automatically design effective and efficient crowd counting architectures.

3.10. Others

There are also some other emerging network designs for crowd counting, discussed below:

- *Generative Adversarial Networks* Generative Adversarial Networks (GAN) has been applied to a wide range of tasks in computer vision, and also have been adopted to crowd counting tasks such as GAN-MTR [128], MS-GAN [213,236], ACSCP [111] and CODA [79]. Generative adversarial networks can be used to improve the visual quality of the generated density maps, but usually degrades counting accuracy. For example, MS-GAN [213,236] proposed multi-scale GAN, which incorporates the inception module in the generation part. This paper investigated GAN as an effective solution to the crowd counting problem, to generate high-quality crowd density maps of arbitrary crowd density scenes. Besides, Adversarial Cross-Scale Consistency Pursuit (ACSCP) [111] designed a novel scale-consistency regularizer that enforces that the sum up of the crowd counts from local patches. The authors further boosted density estimation performance by further exploring the collaboration between both objectives.
- *Graph neural networks* based method distills rich relations among multi-scale features for crowd counting. As shown in Fig. 4 (f), HyGnn [110] exploits useful information from the auxiliary task (localization branch). The HyGnn module in the red box jointly represents the task-specific feature maps of different scales as nodes, multi-scale relations as edges, counting, and localization relations as edges, which distilled rich relations between the nodes to obtain more powerful representations, leading to robust and accurate results.
- *Recurrent neural networks* based Deep Recurrent Spatial-Aware Network (DSRNet) [96] utilize a learnable spatial transform module with a region-wise refinement process to adaptively enlarge the varied scales coverage. Researchers in [150] decoded the features into local counts using an LSTM decoder, finally predicts the image global count. The local counts and global count are all learning targets.
- *Prior-guided modules* help enhancing counting performance, as discussed in recent literature [81,141,211,216,143,132,124,229,61,179]. Multi-stage density map regression network is a scale-aware convolutional neural network (MMNet) [31], which not only captures multi-scale features generated by various sizes of filters but also integrates multi-scale features generated by different stages to handle scale variation problems.
- *Local counting network* proposes an adaptive mixture regression framework [104] in a coarse-to-fine manner to improve counting accuracy, which fully utilizes the context and multi-scale information from different convolutional features. Besides, local counting networks perform more precise counting regression on local patches of images.
- *Multi-model fusion* is another class of techniques for crowd counting [142,92,166]. Recently, most of the current works for crowd counting with state-of-the-art performance are density-map estimation-based approaches. Some researchers tried to improve the existing framework with both point and box annotation such as LCFCN [71], PSDDN [106], BSAD [57], DecideNet [90] and DRD [49]. DecideNet [90] is one of the typical methods, which proposed a separate decide subnet to combine detection and density estimation. Combining detection with density map estimation usually utilizes detection for the low crowd and density estimation for the high crowd. However, these kinds of methods require high computational complexity and high annotation complexity.

4. Loss Function

The loss functions are used to optimize the model. Early works usually adopt the pixel-wise Euclidean loss (Section 4.1), later different advanced loss functions are utilized for better density estimation. In this section, we discuss some recent advances on loss functions for crowd counting: SSIM loss (Section 4.2), and multi-task learning (Section 4.3). We compare them in Section 4.4 and present some other emerging considerations in Section 4.5.

4.1. Euclidean Loss

Most of the early crowd counting approaches use Euclidean loss to optimize the models. The Euclidean loss is a pixel-wise estimation error:

$$L_E = \frac{1}{N}||F(x_i;\theta) - y_i||_2^2, \tag{1}$$

where θ indicates the model parameters, N means the number of pixels, x_i denotes the input image, and y_i is ground truth and $F(x_i;\theta)$ is the generated density map. The total crowd counting result can be summarized over the estimated crowd density map. The pixel-wise L2 loss is a flexible and widely used loss function for crowd counting. However, this pixel-wise loss does not take local and global contextual information as well as the visual quality of the generated density maps into account. Thus, this kind of loss function cannot produce satisfactory high-quality density maps and highly accurate crowd estimation.

4.2. SSIM Loss

Some variants of structure similarity (SSIM) loss are proposed for crowd counting to force the network to learn the local correlation within regions of various sizes, thereby producing locally consistent estimation results such as SSIM loss [11], multi-scale SSIM loss [134], DMS-SSIM loss [95] and DMSSIM loss [75]. Then the local pattern consistency can be formulated as:

$$L_s = 1 - \frac{1}{N}\sum_x SSIM(x). \tag{2}$$

The pixel-wise Euclidean loss usually assumes that adjacent pixels are independent and ignores the local correlation in the density maps, the Euclidean loss can be fused with the SSIM loss to leverage local correlations among pixels for generating high-quality density maps and accurate crowd estimation.

For example, the Cross-Level Parallel Network [75] fused the difference of mean structural similarity index (DMSSIM) with the MSE loss to optimize the module. Besides, Multi-View Scale Aggregation Networks [134] proposed a multi-scale SSIM for multi-view crowd counting. However, SSIM loss is hard to learn local correlations with a large spectrum of varied scales.

4.3. Multi-task Learning

The main task of crowd counting is the total counting accuracy, thus the direct global count constraints may benefit the counting accuracy. The headcount loss can be defined as:

$$L_c = \frac{1}{N}\sum_{i=1}^{N}||\frac{F_c(x_i;\theta) - y_i}{y_i + 1}||, \tag{3}$$

where $F_c(x_i;\theta)$ is the estimated head count, and y_i is the ground truth head count. Then the total loss function is formulated as follow:

$$L_{total} = L_E + \alpha L_c, \tag{4}$$

where α is the weight to balance the pixel-wise Euclidean loss and the total head counting loss. BL [114] stated that the original GT density map is imperfect due to occlusions, perspective effects, variations in object shapes and proposed Bayesian loss to constructs a density contribution probability model from the point annotations and addressed the above issues. The proposed Bayesian loss adopted more reliable supervision on the count expectation at each annotated point.

SaCNN [222] proposed to combine density map loss with the relative count loss. The relative count loss helps to reduce the variance of the prediction errors and improve the network generalization on very sparse crowd scenes. CFF [153] fused segmentation map loss, density map loss and global density loss. Plug-and-Play Rescaling [144] combined regression loss with classification loss. Shallow Feature-based Dense Attention Network [123] proposed to use MSE loss with counting loss and stated that counting loss not only accelerates the convergence but also improves the counting accuracy. Multi-supervised Parallel Network [196] combined MSE loss, cross-entropy loss, and L1 loss. Besides, there is also some paper to use a kind of combination loss to enforce similarities in local coherence and spatial correlation between maps [62,136] [60]. Multi-task learning based framework is widely used in recent papers [165,88,48,70,41,227,156]. However, this kind of framework is sensitive to hyper-parameters.

4.4. Comparisons

We summarize the advantages and limitations of the above loss functions in Table 6. We compare in Table 7 the performance of several state-of-the-arts with different loss functions.

CSRNet and ADCrowdNet are based on the same Euclidean loss but with different deep neural network designs and show different counting accuracy, which shows that the Euclidean loss is flexible and widely used in the early approaches. However, the Euclidean loss lacks contextual information and ignores the local correlation among pixels in the density maps.

The DSSINet achieves better performance than CSRNet and ADCrowdNet on different crowd counting datasets. These variants of structural similarity loss show counting improvements based on utilizing local correlation. However, these kinds of methods suffer in the situation of a large spectrum of various scales.

As Table 7 shows, DSSINet (SSIM loss) achieves better counting accuracy than ACSCP (Adversarial loss) with similar network design considerations (i.e., multi-scale scheme and dilated convolutional operations). The poor performance of ACSCP on ShanghaiTech A & B may probably be due to the adversarial loss. This further demonstrates that adversarial loss can help to generate high-quality density maps but may sacrifice counting accuracy.

HA-CNN shows better performance than ADCrowdNet even without deformable convolutional operations on two different crowd counting datasets. This demonstrates that multi-task learning with global counting constrain can work well in highly crowded scenes even without some advanced network operations. S-DCNet also achieves satisfactory counting accuracy on different crowd counting datasets, which confirms the effectiveness of the divide and conquer manner but is computationally expensive.

4.5. Others

There are some other loss optimization strategies to enhance crowd counting tasks [94,176,190,65,149]. CNN-Boosting [169] employed CNNs and incorporate two significant improvements: layered boosting and selective sampling. DAL-SVR [197] boosted deep attribute learning via support vector regression for fast-moving crowd counting. The paper learned superpixel

Table 6
Comparisons of recent advanced loss functions for crowd counting. The property of representative schemes for each loss functions category are summarized based on **advantages** and **limitations**.

Category	Representative Scheme	Advantages	Limitations
Euclidean loss	CSRNet [80]	Flexible; widely used	Not consider context information and visual quality
SSIM loss	SANet [11]	Variants of structural similarity loss to learn local correlation	Hard to learn the local correlation with various scales
Multi-task learning	MSPNet [196]	Varied and flexible rightarrow fuse different constrains	Sensitive to hyper-parameters
Others	S-DCNet [205]	Efficient divide and conquer manner	Computational expensive

Table 7
Comparisons of state-of-the-art crowd counting approaches with different loss functions. **Multi scale** is the multi-scale design considerations; **Dilated** is dilated convolutions; **Deform** is the deformable convolutions; **Atten** represents the attention-based scheme. **ST-A** denotes ShanghaiTech A dataset [228] and **ST-B** denotes ShanghaiTech B dataset [228].

Scheme	Multi scale	Dilated	Deform	Atten	Loss function	ST-A MAE	ST-A MSE	ST-B MAE	ST-B MSE
CSRNet [80]		√			Euclidean loss	68.2	115.0	10.6	16.0
ADCrowd [97]			√	√	Euclidean loss	63.2	98.9	8.2	15.7
DSSINet [95]	√	√			SSIM Loss	60.63	96.04	6.8	10.3
S-DCNet [205]	√				Divide-conquer	58.3	95.0	6.7	10.7
HA-CCN [159]				√	MSE loss with counting	62.9	94.9	8.1	13.4
GLoss [173]					Unbalanced optimal transport loss	61.3	95.4	7.3	11.7

segmentation-fast moving segmentation-feature extraction-motion features/appearance features/sift feature-features aggregation by PCA-regression learning SVR-data fusion and deeply learning cumulative attribute. D-ConvNet [154] used seep negative correlation learning, which is a successful ensemble learning technique for crowd counting. The authors extended D-ConvNet in [223], which proposed to regress via an efficient divide and conquer manner. D-ConvNet has been shown to work well for non-deep regression problems. Without extra parameters, the method controls the bias-variance–covariance trade-off systematically and usually yields a deep regression ensemble where each base model is both accurate and diversified. However, the whole framework is computationally expensive.

S-DCNet [205] designed a multi-stage spatial divide and conquer network. The collected images and labeled count values are limited in reality for crowd counting, which means that only a small closed set is observed. A dense region can always be divide until sub-region counts are within the previously observed closed set. S-DCNet only learns from a closed set but can generalize well to open-set scenarios. And avoid repeatedly computing sub-region convolutional features, this method is also efficient.

5. Supervisory Signal

In this section, we discuss different supervisory signals for crowd counting: fully supervised learning (Section 5.1), weakly supervised and semi-supervised learning (Section 5.2), unsupervised and self-supervised learning (Section 5.3), and automatic labeling through synthetic data (Section 5.4). We evaluate and compare them in Section 5.6.

5.1. Fully Supervised Learning

In the fully supervised crowd counting paradigm, the model is hard to optimize if we utilize the original discrete point-wise annotation maps as ground truth [83,39,218,109,89,1,192,234,172,147]. There are also some recent works study the problem of counting from scalar representations [182,163,84,116]. The continuous ground truth density map is usually generated from the original point-wise annotations via different ground truth generation methods such as applying an adaptive Gaussian kernel for each head annotation, which is important for accurate crowd estimation [174]. The fixed kernel or adaptive Gaussian kernel are widely used approaches to prepossess the original annotation and get the ground truth for density estimation and crowd counting [80]. The geometry-adaptive kernel is defined as follows:

$$F(x) = \sum_{i=1}^{N} \delta(x - x_i) \times G_{\sigma_i}(x), with\ \sigma_i = \beta \bar{d}_i, \tag{5}$$

where x denotes the pixel position in an image. For each target object, x_i in the ground truth, which is presented with a delta function $\delta(x - x_i)$. The ground truth density map $F(x)$ is generated by convolving $\delta(x - x_i)$ with a normalized Gaussian kernel based on parameter σ_i. And $\bar{d}_i$ shows the average distance of the k nearest neighbors.

GP [9] devises a Bayesian model that places a Gaussian process before a latent function whose square is the count density. Compared to different annotation methods concerning their difficulty for the annotator: dots or bounding box in all objects, GP is better in terms of accuracy and labeling effort. Besides, there are some recent advances to use a learned kernel to improve the prepossessing step and proposed an adaptive density map generator [170].

DM-Count [175] optimizes the network directly on the dot map, which can be considered as a special type of density map with 1×1 Gaussian blur. Most existing methods need to use an adaptive or fixed Gaussian to smooth each annotated dot or to estimate the likelihood of every pixel given the annotated point. DM-Count directly optimizes the original annotation and shows its generation error bound is tighter than that of Gaussian smoothed methods.

5.2. Weakly Supervised and Semi-supervised Learning

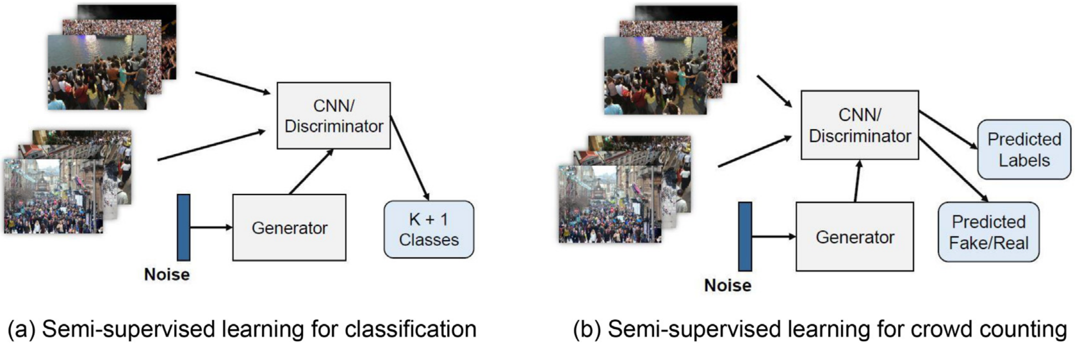
Recently, a number of works have emerged to make use of weakly labeled data for crowd counting [122,191,67,161,105,231,217,68,235] and the problem of learning from noisy annotations [171,78]. The original annotation process for crowd counting via density map estimation is point-level annotation, which is labor-intensive, HA-CCN [159] proposed a weakly supervised learning setup and leveraged the image-level labels instead of the densely point-wise annotation process to reduce label effort. As shown in Fig. 2 (b), the first column is the original image, the second column is the labor-intensive dense (head) annotation, the third column is the ground truth density maps, and the last column is the image-level weak annotation, which is used in the weakly supervised learning setting. This clearly shows that leveraging weakly labeled data (the last column) can largely reduce the annotation complexity compared with fully point-wise annotation (the second column). Besides, Scale-Recursive Network (SRN) with point supervision [32] is also a kind of weakly supervised framework based on SRN structure.

Typical semi-supervised GANs are unable to function in the regression regime due to biases introduced when using a single prediction goal. DG-GAN [127] generalized semi-supervised generative adversarial network (GANs) from classification problems to regression for use in dense crowd counting, refer to Fig. 5. This work allows the dual-goal GAN to benefit from unlabeled data in the training process. And [130] is an extension of DG-GAN, which proposed a novel loss function for feature contrasting and resulted in a discriminator that can distinguish between fake and real data based on feature statistics. However, weakly supervised crowd counting still requires annotations. Besides, it also requires task-specific knowledge to design effective neural networks and loss functions for leveraging weakly labeled data.

5.3. Unsupervised and Self-supervised Learning

Deep learning-based approaches are highly data-driven, i.e., they require a large amount of diverse labeled data in the training process. The labeling process for crowd counting is expensive, but the unlabeled data are cheap and widely available [146,107]. L2R [103] leveraged abundantly available unlabeled crowd images in learning to rank framework, refer to Fig. 6 (a), which is based on the observation that any sub-image of a crowded scene image is guaranteed to contain the same number or fewer persons than the super-image. The pixel-wise regression loss is fused with the ranking regularization to learn better representation for crowd counting tasks on unlabeled data.

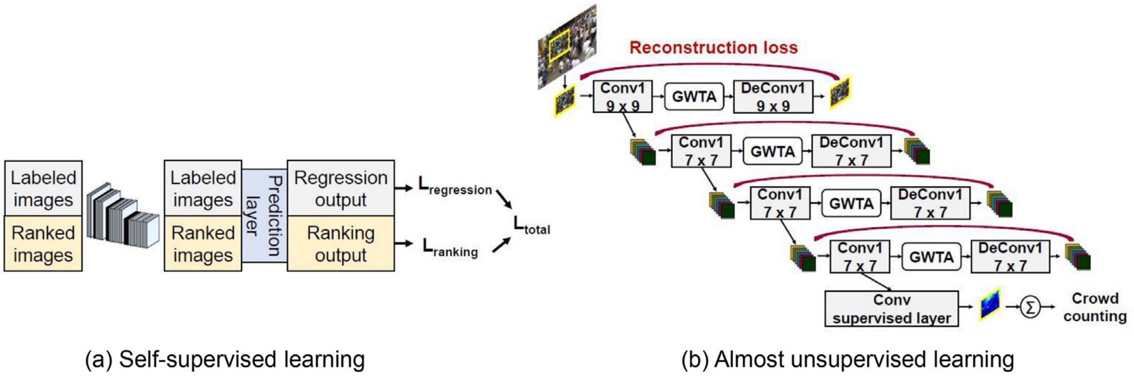
There is another potential direction to make use of unlabeled data such as the convolutional Winner-Take-All models, whose most parameters are obtained by unsupervised learning. GWTA-CCNN [148] utilized a Grid Winner-Take-All (GWTA) autoencoder to learn several layers of useful filters from unlabeled crowd images, refer to Fig. 6 (b). A small patch cropped from the original image is fed into the model. Most of the parameters are trained layer by layer based on the reconstruction loss. GWTA divides a convolution layer spatially into a grid of cells. Within each cell, only the maximumly activated neuron is allowed to update the filter. almost 99.9% of the parameters of the proposed model are trained without any labeled data, which the rest 0.1% are tuned with supervision. However, these kinds of self-supervised learning



(a) Semi-supervised learning for classification

(b) Semi-supervised learning for crowd counting

Fig. 5. The workflow of the original semi-supervised learning for classification problem (Figure a) and semi-supervised learning for single image crowd counting (Figure b) [127].



(a) Self-supervised learning

(b) Almost unsupervised learning

Fig. 6. The workflow of self-supervised learning and almost unsupervised learning for crowd counting. (a) The architecture of L2R: a self-supervised learning setup [103]. (b) The framework of GWTA-CCNN: an almost unsupervised learning method [148].

and almost unsupervised crowd counting approaches need a large amount of data to show effectiveness, which requires more training time and computational resources.

Lei et al. [72] proposed a weakly supervised crowd counting method to train the model from a small number of dot-map annotations and a large number of count-level annotations, with is used to reducing the annotation cost for crowd counting. The key idea is to enforce the consistency between density maps and total object count on weakly labeled images as regularization terms. The work of complete self-supervision [146] introduce a new training paradigm that does not need labeled data. This work reveals the power law nature for the distribution of crowds and adopt this signal for backpropagation in the optimal transport framework. This work achieves efficient crowd estimation.

5.4. Automatic Labeling through Synthetic Data

There are more challenges for crowd counting in the wild due to the changeable environment, large-range number of people cause the current methods can not work well. Due to scarce data, many methods suffer from over-fitting to a different extent. Some researchers attempt to tackle this problem through synthetic data [53,188]. CCWld [186] built a large-scale, diverse synthetic dataset, pretrain a crowd counter on the synthetic data, finetune on real data, propose a counting method via domain adaptation based cycle GAN, free humans from heavy data annotations. The authors in [48] based on the GCC dataset, designed a better domain adaptation scheme for reducing the counting noise in the background area. This paper pays more attention to the semantic consistency of the crowd and then could narrow the gap using a large-scale human detection dataset to train a crowd, semantic model. This method reduces the labeling effort, enhances accuracy, and improves robustness by making use of synthetic data. However, the synthetic data are still witnessed a larger domain gap compared with real data.

5.5. Domain Adaptive Crowd Counting

Most of the existing crowd counting methods are designed in a specific domain. Thus, designing crowd counting models that can achieve high counting performance in any domain is a challenging but meaningful problem. There is some robust crowd counting approaches against domain shifts proposed in recent years [226,51,98,214,232,138].

CVCS [226] proposes a cross-view cross-scene multi-view crowd counting paradigm, where the training and test set are from different scenes with arbitrary camera locations. CVCS are able to attentively selects and fuses multiple views using camera layout geometry, and a noise view regularization method to handle non-correspondence errors. CDCC [189] proposes a neural linear transformation method, which exploits domain factor and biases weights to learn the domain shift. AdaCrowd [139] makes use of a crowd counting network and a guiding network, which predicts some parameters in the counting network based on the unlabeled data from a particular scene and adapt to the new scene.

The work of [43] introduces a domain-adaptation-style crowd counting method by using multilevel feature-aware adaptation and structured density map alignment module, which is trained on generated data with ground-truth to the specific real-world scenes. The work [40] proposes to learn from synthetic crowd data and transferring knowledge to real data without ground truth. This DACC frame work adopt a high-quality image translation and density map reconstruction to enhance cross domain crowd counting quality. The work [10] propose a two-step approach that captures the intra-domain knowledge to facilitate unsupervised cross-domain crowd counting via synthetic datasets.

The scale or density gap among datasets is another type of domain gap for domain adaptive crowd counting [113,209,199,44]. For example, The work of [113] proposes a universal crowd counting model that can be applied across scenes and datasets via a scale alignment module. DCANet [209] introduces a domain-guided channel attention network to guide the extraction of domain-specific feature representation for multi-domain crowd counting. DKPNet [15] designs a domain-specific knowledge propagating network for extracking knowledge from multiple domains for improving crowd counting performance.

5.6. Comparisons

We summarize different supervisory signals for crowd counting with their representative schemes in Table 8. We compare them in Table 9.

BL achieves better performance on four different crowd counting datasets compared with CP-CNN, with a similar number of parameters for the backbone. The good performance of BL may be due to the Bayesian loss used to better model the non-rigid objects (e.g., people). The adaptive Gaussian kernel is widely used in crowd counting approaches, while the experimental results demonstrate the effectiveness of Bayesian loss, which is more reliable supervision.

CCWld shows much better accuracy than MCNN in Table 9 on various datasets with different backgrounds. We observe CCWld enhances the performance of counting accuracy and also improves the robustness, which is suitable for many real-world applications with diverse scenes, different view angles, and lighting conditions.

As shown in Table 9, the performance of HA-CNN is much better than other state-of-the-arts. After carefully designing the deep neural networks and loss functions, weakly supervised crowd counting achieves much better accuracy with relatively low annotation complexity.

The MAE and MSE of L2R (query by example) and L2R (query by keyword) is lower than CP-CNN. This confirms that leveraging the abundantly available unlabeled data improves counting performance. The experimental results further demonstrate that making use of unlabeled data is a promising direction for crowd counting.

5.7. Others

There are some other learning paradigms for crowd counting.

There is a typical training paradigm that is count from scalar representation. Some recent works achieve excellent results compared with density map regression method or learning from point map representation. TransCrowd [82] proposes to formulate crowd counting as a sequence-to-count paradigm based on transformers and achieves satisfactory performance. CrowdMLP [182] presents a multi-granularity MLP regressor for capturing global information and enchance crowd counting quality.

Recent research shows that the crowd localization can enhance the counting performance. FIDT [83] introduces a focal inverse distance transform map for crowd counting and crowd localization, which simultaneously conduct counting and crowd localization based on the FIDT map. IIM [39] presents an independent instance map segmentation for crowd localization by segmenting people crowds into non-overlapped independent components.

There is another series of counting works that achieve crowd counting from remote sensing data. The work [237] introduces a crowd counting benchmark from remote sensing perspective. The work [38] proposes a large-scale dense objects counting dataset based on remote sensing images. The work [230] proposes a flow-based Bi-path Network for remote sensing video sequences. IS-Count [121] presents a convariate-based importance sampling method for counting from remote sensing images. Compared with counting from normal perspective, the remote sensing images suffers more from small object recognition issues in designing the

Table 8
Comparisons of different supervisory signals for crowd counting. The representative schemes of different supervisory signals are analyzed based on their **advantages** and **limitations**.

Category	Schemes	Advantages	Limitations
Fully Supervised learning	CP-CNN [157]	Adaptive Gaussian kernel rightarrow accommodate different scales	Not flexible to non-rigid object
	BL [114]	Bayesian loss rightarrow model non-rigid objects	More reliable supervision but suffers in varied scales
Weakly supervised and semi-supervised learning	HA-CCN [159]	Low annotation complexity	Still requires weakly annotations and task specific knowledge
Unsupervised and self-supervised learning	L2R [103]	Low annotation cost; abundantly available	Large amount of data requires more training time
Automatic labeling through synthetic data	CCWld [186]	Reduce labeling effort; enahnce accuracy; improve robustness	Large domain gap from synthetic rightarrow real data

Table 9
Quantitative comparisons of state-of-the-art crowd counting approaches with different supervisory signals. **Column** shows the type and number of columns for counting model. **Multi** is the Multi-column network; **Double** represents two columns; **Single** is the single column network. **ST PartA** and **ST PartB** denotes ShanghaiTech A & B dataset [228], respectively. The evaluation metrics for counting accuracy is **MAE** and **MSE**.

Typical Schemes		ST PartA		ST PartB		UCF_CC_50		UCF-QNRF	
Methods	**Column**	**MAE**	**MSE**	**MAE**	**MSE**	**MAE**	**MSE**	**MAE**	**MSE**
CP-CNN [157]	Multi	73.6	106.4	20.1	30.1	295.8	320.9	-	-
L2R [103] (Query by example)	Double	72.0	106.6	14.4	23.8	291.5	397.6	-	-
L2R [103] (Query by keyword)	Double	73.6	112.0	13.7	21.4	279.6	388.9	-	-
BL [114]	Single	62.8	101.8	7.7	12.7	229.3	308.2	88.7	154.8
CCWld [186]	Single	64.8	107.5	7.6	13.0	-	-	102.0	171.4
URC [208]	Single	72.8	111.6	12.0	18.7	294.0	443.1	128.1	218.1
HA-CCN [159]	Single	58.3	95.0	6.7	10.7	256.2	348.4	118.1	180.4

Table 10
A comprehensive performance analysis of various categories of crowd counting methods across different datasets. Bold denotes the best performance and italic denotes the third best performance. **ST PartA** is the ShanghaiTech A dataset [228]. The evaluation metrics for the counting performance is **MAE** and **MSE**.

Typical Schemes			ST PartA		UCF_CC_50		UCF-QNRF		NWPU	
Methods	**Year**	**Column**	**MAE**	**MSE**	**MAE**	**MSE**	**MAE**	**MSE**	**MAE**	**MSE**
CSRNet [80]	2018	Single	68.2	115.0	266.1	397.5	-	-	104.8	433.4
SaCNN [222]	2018	Single	86.8	139.2	314.9	424.8	-	-	-	-
DADNet [46]	2019	Single	64.2	99.9	285.5	389.7	113.2	189.4	-	-
MRNet [165]	2019	Single	63.3	97.8	232.3	314.8	111.1	182.8	-	-
ADCNet [97]	2019	Single	70.9	115.2	273.6	362.0	-	-	-	-
HA-CNN [159]	2019	Single	62.9	94.9	256.2	348.4	118.1	180.4	-	-
PGCNet [210]	2019	Single	57.0	*86.0*	244.6	361.2	-	-	-	-
SDANet [123]	2020	Single	63.6	101.8	227.6	316.4	-	-	-	-
CTN [137]	2020	Single	61.5	103.4	210.0	305.4	86.0	146.0	78.0	448.0
DM-Count [175]	2020	Single	59.7	95.7	211.0	291.5	85.6	148.3	70.5	357.6
NAS-Count [55]	2020	Single	56.7	93.4	208.4	297.3	101.8	163.2	-	-
SRF-Net [20]	2020	Single	60.4	97.2	197.3	271.8	98.0	170.0	-	-
ADSCNet [8]	2020	Single	55.4	97.7	198.4	267.3	**71.3**	132.5	-	-
UEPNet [176]	2021	Single	54.6	91.2	165.2	275.9	81.1	131.7	-	-
S3 [84]	2021	Single	57.0	96.0	-	-	80.6	139.8	83.5	346.9
NDConv [218]	2022	Single	61.4	104.2	167.2	240.6	95.9	182.4	-	-
TransCrowd [82]	2022	Single	66.1	105.1	272.2	395.3	97.2	168.5	88.4	400.5
MAN [85]	2022	Single	56.8	90.3	-	-	77.3	131.5	76.5	*323.0*
CMTL [156]	2017	Double	101.3	152.4	322.8	397.9	-	-	-	-
ACSCP [151]	2018	Double	75.7	102.7	291.0	404.6	-	-	-	-
SDNet [113]	2021	Double	55.0	92.7	197.5	264.1	80.7	146.3	-	-
BM-Count [89]	2021	Double	57.3	90.7	-	-	81.2	138.6	83.4	358.4
BSCC [125]	2021	Double	58.3	100.1	-	-	86.3	153.1	-	-
P2PNet [163]	2021	Double	**52.7**	**85.1**	172.7	256.2	85.3	154.5	77.4	362.0
GauNet [23]	2022	Double	54.8	89.1	186.3	256.5	81.6	153.7	-	-
RAN [21]	2022	Double	57.9	99.2	**155.0**	**219.5**	83.4	141.8	*65.3*	432.9
MCNN [228]	2016	Multi	110.2	173.2	377.6	509.1	277	426	218.5	700.6
CP-CNN [157]	2017	Multi	73.6	106.4	295.8	320.9	-	-	-	-
Switching [6]	2017	Multi	90.4	135.0	318.1	439.2	-	-	-	-
SANet [11]	2018	Multi	67.0	104.5	258.4	334.9	-	-	-	-
DSSINet [95]	2019	Multi	60.6	96.0	216.9	302.4	99.1	159.2	-	-
CFF [153]	2019	Multi	65.2	109.4	-	-	93.8	146.5	-	-
S-DCNet [205]	2019	Multi	58.3	95.0	204.2	301.3	104.4	176.1	-	-
CAN [99]	2019	Multi	62.3	100.0	212.2	243.7	107.0	183.0	-	-
SPANet [24]	2019	Multi	59.4	92.5	232.6	311.7	-	-	-	-
DPN [115]	2020	Multi	58.1	91.7	183.2	284.5	84.7	147.2	-	-
AMRNet [104]	2020	Multi	61.6	98.4	184.0	265.8	86.6	152.2	-	-
ASNet [64]	2020	Multi	57.8	90.1	174.8	251.6	91.6	159.7	-	-
DeepCount [22]	2020	Multi	65.2	112.5	-	-	95.7	167.1	-	-
ikNN [129]	2020	Multi	68.0	117.7	237.8	305.7	104.0	172.0	-	-
M-SFANet [168]	2020	Multi	57.6	94.5	167.5	256.3	87.6	147.8	-	-
EPA [215]	2021	Multi	60.9	91.6	250.1	352.1	-	-	-	-
DKPNet [15]	2021	Multi	55.6	91.0	-	-	81.4	147.2	**61.8**	438.7
SASNet [164]	2021	Multi	*53.6*	88.4	*161.4*	*234.5*	85.2	147.3	-	-
MFDC [102]	2021	Multi	55.4	91.3	-	-	*76.2*	*121.5*	74.7	**267.9**
MPS [218]	2022	Multi	71.4	110.7	-	-	-	-	-	-
MNA [171]	2020	N/A	61.9	99.6	-	-	85.8	150.6	96.9	534.2
BL [114]	2019	N/A	62.8	101.8	229.3	308.2	88.7	154.8	-	-
UOT [116]	2021	N/A	58.1	95.9	-	-	83.3	142.3	87.8	387.5
BinLoss [155]	2021	N/A	61.3	88.7	-	-	85.9	**120.6**	71.7	376.4

counting networks but the problem of scale variation for counting from normal perspective is more serious.

6. Conclusion and Future Directions

Crowd counting is an important and challenging problem in computer vision. This survey paper covers the design considerations and recent advances with respect to single image crowd counting problem, and summarizes more than 200 crowd counting schemes using deep learning approaches proposed since 2015. We have discussed the major datasets, performance metrics, design considerations, techniques, and representative schemes to tackle the problem. We provide a comprehensive overview and comparison of three major design modules for deep learning in crowd counting, deep neural network design, loss function, and supervisory signal. The research field of crowd counting is rich and still evolving. We discuss some future trends and possible research directions below:

- *Automatic and lightweight network designing* has drawn much attention in recent years [91,152,183,200]. Currently, designing CNN-based crowd counting models still requires a manual network and feature selection with strong domain knowledge. Automated Machine Learning has been applied to image classification and object detection, which has the potential to automatically design efficient crowd counting architectures. Besides, CNN-based crowd counting models have increased in-depth with millions of parameters, which requires massive computation. Thus, there is also a need for model compression and acceleration techniques to deploy lightweight model.
- *Weakly supervised and unsupervised crowd counting* is able to reduce the labeling effort. With the performance saturation for some supervised learning scenarios, researchers devote efforts to make use of unlabeled and weakly labeled images for crowd counting Most of the state-of-the-art algorithms are based on fully supervised learning and trained with point-wise annotations, which has several limitations such as

labor-intensive labeling process, easily over-fitting, and not salable in the absence of densely labeled crowd images. Weakly-supervised and unsupervised learning has attracted much attention in vision applications, which has value for crowd counting tasks to reduce labeling effort, enhance counting accuracy and improve robustness.

- *Crowd counting in videos* is becoming an active research direction. A straightforward approach is to consider the video frames independently by making use of the crowd counting techniques proposed for still images. This is not satisfactory because it ignores the continuity or temporal correlation between frames, i.e., the motion information. Bidirectional ConvLSTM [204] is a recent attempt to leverage spatial–temporal information in video. There are some recent attempts to exploit the correlation in video data [242,36,47,100,140,112,101,43]. However, LSTM-based framework is not easy to train or to be extended to a general scenario. The 3D kernel is not effective in extracting the long-range contextual information. Effectively making use of the temporal correlation for accurate and efficient near real-time crowd counting systems is also a potential research direction.
- *Multi-view fusion for crowd counting* is important as a single camera cannot capture large and wide areas (e.g., parks, public squares). Multiple cameras with overlapping view are required to solve the wide-area counting task. There are some recent multi-view fusion approaches for crowd counting [224], which proposes a multi-camera fusion method to predict a ground-plane density map of the 3D world. There is also another approach based on a 2D-to-3D projection with 3D density map estimation and a 3D-to-2D projection consistency measure method [225]. Multi-view fusion for crowd counting provides a vivid visualization for the scenes, as well as the potentials for other applications like observing the scene in arbitrary view angles, which may contribute to better scene understanding. Therefore, crowd counting with multi-view fusion represents important research value.

Declaration of Competing Interest

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

Acknowledgment

This work was supported, in part, by Hong Kong Innovation and Technology Fund (under grant number ITP/050/20LP) and Hong Kong General Research Fund (under grant number 16200120).

References

[1] Abousamra, S., Hoai, M., Samaras, D., Chen, C.: Localization in the crowd with topological constraints. In: AAAI (2021).
[2] Ahuja, K.R., Charniya, N.N.: A survey of recent advances in crowd density estimation using image processing. In: ICCES (2019).
[3] S. Amirgholipour, X. He, W. Jia, D. Wang, L. Liu, Pdanet: Pyramid density-aware attention net for accurate crowd counting, NeuroComputing (2020).
[4] C. Arteta, V. Lempitsky, A. Zisserman, Counting in the wild, in: ECCV, 2016.
[5] S. Aydín, Deep learning classification of neuro-emotional phase domain complexity levels induced by affective video film clips, IEEE Journal of Biomedical and Health Informatics (2019).
[6] Babu Sam, D., Surya, S., Venkatesh Babu, R.: Switching convolutional neural network for crowd counting. In: CVPR (2017).
[7] Bai, H., Wen, S., Gary Chan, S.H.: Crowd counting on images with scale variation and isolated clusters. In: ICCV Workshops (2019).
[8] S. Bai, Z. He, Y. Qiao, H. Hu, W. Wu, J. Yan, Adaptive dilated network with self-correction supervision for counting, in: CVPR, 2020.
[9] von Borstel, M., Kandemir, M., Schmidt, P., Rao, M.K., Rajamani, K., Hamprecht, F.A.: Gaussian process density counting from weak supervision. In: ECCV (2016).
[10] Y. Cai, L. Chen, Z. Ma, C. Lu, C. Wang, G. He, Leveraging intra-domain knowledge to strengthen cross-domain crowd counting, in: ICME, 2021.
[11] Cao, X., Wang, Z., Zhao, Y., Su, F.: Scale aggregation network for accurate and efficient crowd counting. In: ECCV (2018).
[12] A.B. Chan, Z.S.J. Liang, N. Vasconcelos, Privacy preserving crowd monitoring: Counting people without people models or tracking, in: CVPR, 2008.
[13] Chan, A.B., Vasconcelos, N.: Bayesian poisson regression for crowd counting. In: ICCV (2009).
[14] A.B. Chan, N. Vasconcelos, Counting people with low-level features and bayesian regression, TIP (2012).
[15] Chen, B., Yan, Z., Li, K., Li, P., Wang, B., Zuo, W., Zhang, L.: Variational attention: Propagating domain-specific knowledge for multi-domain learning in crowd counting. In: ICCV (2021).
[16] J. Chen, W. Su, Z. Wang, Crowd counting with crowd attention convolutional neural network, Neurocomputing 382 (2020) 210–220.
[17] Chen, K., Loy, C.C., Gong, S., Xiang, T.: Feature mining for localised crowd counting. In: BMVC (2012).
[18] X. Chen, Y. Bin, C. Gao, N. Sang, H. Tang, Relevant region prediction for crowd counting, Neurocomputing 407 (2020) 399–408.
[19] Chen, X., Bin, Y., Sang, N., Gao, C.: Scale pyramid network for crowd counting. In: WACV (2019).
[20] Y. Chen, C. Gao, Z. Su, X. He, N. Liu, Scale-aware rolling fusion network for crowd counting, in: ICME, 2020.
[21] Y. Chen, J. Yang, D. Zhang, K. Zhang, B. Chen, S. Du, Region-aware network: Model human's top-down visual perception mechanism for crowd counting, Neural Networks (2022).
[22] Z. Chen, J. Cheng, Y. Yuan, D. Liao, Y. Li, J. Lv, Deep density-aware count regressor, ECAI (2019).
[23] Z.Q. Cheng, Q. Dai, H. Li, J. Song, X. Wu, A.G. Hauptmann, Rethinking spatial invariance of convolutional networks for object counting, in: CVPR, 2022.
[24] Cheng, Z.Q., Li, J.X., Dai, Q., Wu, X., Hauptmann, A.G.: Learning spatial awareness to improve crowd counting. In: ICCV (2019).
[25] Z.Q. Cheng, J.X. Li, Q. Dai, X. Wu, J.Y. He, A.G. Hauptmann, Improving the learning of multi-column convolutional neural network for crowd counting, in: ACM Multimedia, 2019.
[26] A. Chrysler, R. Gunarso, T. Puteri, H. Warnars, A literature review of crowd-counting system on convolutional neural network, in: IOP Conference Series: Earth and Environmental Science, 2021.
[27] F. Dai, H. Liu, Y. Ma, X. Zhang, Q. Zhao, Dense scale network for crowd counting, in: International Conference on Multimedia Retrieval, 2021.
[28] D. Deb, J. Ventura, An aggregated multicolumn dilated convolution network for perspective-free counting, in: CVPR Workshops, 2018.
[29] X. Ding, F. He, Z. Lin, Y. Wang, H. Guo, Y. Huang, Crowd density estimation using fusion of multi-layer features, IEEE Transactions on Intelligent Transportation Systems (2020).
[30] Ding, X., Lin, Z., He, F., Wang, Y., Huang, Y.: A deeply-recursive convolutional network for crowd counting. In: ICASSP (2018).
[31] L. Dong, H. Zhang, Y. Ji, Y. Ding, Crowd counting by using multi-level density-based spatial information: A multi-scale cnn framework, Information Sciences (2020).
[32] Z. Dong, R. Zhang, X. Shao, Y. Li, Scale-recursive network with point supervision for crowd scene analysis, Neurocomputing 384 (2020) 314–324.
[33] A. Draghici, M.V. Steen, A survey of techniques for automatically sensing the behavior of a crowd, ACM Computing Surveys (2018).
[34] D. Du, L. Wen, P. Zhu, H. Fan, Q. Hu, H. Ling, M. Shah, J. Pan, A. Al-Ali, A. Mohamed, et al., Visdrone-cc2020: The vision meets drone crowd counting challenge results, in: ECCV, 2020.
[35] H. Duan, S. Wang, Y. Guan, Sofa-net: Second-order and first-order attention network for crowd counting, BMVC (2020).
[36] Y. Fang, S. Gao, J. Li, W. Luo, L. He, B. Hu, Multi-level feature fusion based locality-constrained spatial transformer network for video crowd counting, Neurocomputing 392 (2020) 98–107.
[37] Y. Fang, B. Zhan, W. Cai, S. Gao, B. Hu, Locality-constrained spatial transformer network for video crowd counting, in: ICME, 2019.
[38] Gao, G., Liu, Q., Wang, Y.: Counting dense objects in remote sensing images. In: ICASSP (2020).
[39] Gao, J., Han, T., Yuan, Y., Wang, Q.: Learning independent instance maps for crowd localization. arXiv preprint arXiv:2012.04164 (2020).
[40] J. Gao, T. Han, Y. Yuan, Q. Wang, Domain-adaptive crowd counting via high-quality image translation and density reconstruction, TNNLS (2021).
[41] J. Gao, Q. Wang, X. Li, Pcc net: Perspective crowd counting via spatial convolutional network, TCSVT (2019).
[42] J. Gao, Q. Wang, Y. Yuan, Scar: Spatial-/channel-wise attention regression networks for crowd counting, Neurocomputing 363 (2019) 1–8.
[43] J. Gao, Y. Yuan, Q. Wang, Feature-aware adaptation and density alignment for crowd counting in video surveillance, in: IEEE transactions on cybernetics, 2020.
[44] Gong, S., Zhang, S., Yang, J., Dai, D., Schiele, B.: Bi-level alignment for cross-domain crowd counting. In: CVPR (2022).
[45] R. Guerrero-Gómez-Olmedo, B. Torre-Jiménez, R. López-Sastre, S. Maldonado-Bascón, D. Onoro-Rubio, Extremely overlapping vehicle counting, in: Iberian Conference on Pattern Recognition and Image Analysis, 2015.
[46] D. Guo, K. Li, Z.J. Zha, M. Wang, Dadnet: Dilated-attention-deformable convnet for crowd counting, in: ACM Multimedia, 2019.
[47] T. Han, L. Bai, J. Gao, Q. Wang, W. Ouyang, Dr. vic: Decomposition and reasoning for video individual counting, in: CVPR, 2022.

[48] Han, T., Gao, J., Yuan, Y., Wang, Q.: Focus on semantic consistency for cross-domain crowd understanding. In: ICASSP (2020).

[49] G. He, Z. Ma, B. Huang, B. Sheng, Y. Yuan, Dynamic region division for adaptive learning pedestrian counting, in: ICME, 2019.

[50] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016).

[51] He, Y., Ma, Z., Wei, X., Hong, X., Ke, W., Gong, Y.: Error-aware density isomorphism reconstruction for unsupervised cross-domain crowd counting. In: AAAI (2021).

[52] Hossain, M., Hosseinzadeh, M., Chanda, O., Wang, Y.: Crowd counting using scale-aware attention networks. In: WACV (2019).

[53] Hou, Y., Li, C., Lu, Y., Zhu, L., Li, Y., Jia, H., Xie, X.: Enhancing and dissecting crowd counting by synthetic data. In: ICASSP (2022).

[54] Hou, Y., Li, C., Yang, F., Ma, C., Zhu, L., Li, Y., Jia, H., Xie, X.: Bba-net: A bi-branch attention network for crowd counting. In: ICASSP (2020).

[55] Hu, Y., Jiang, X., Liu, X., Zhang, B., Han, J., Cao, X., Doermann, D.: Nas-count: Counting-by-density with neural architecture search. In: ECCV (2020).

[56] Huang, S., Li, X., Cheng, Z.Q., Zhang, Z., Hauptmann, A.: Stacked pooling for boosting scale invariance of crowd counting. In: ICASSP (2020).

[57] S. Huang, X. Li, Z. Zhang, F. Wu, S. Gao, R. Ji, J. Han, Body structure aware deep crowd counting, TIP (2017).

[58] Huberman-Spiegelglas, I., Fattal, R.: Single image object counting and localizing using active-learning. In: WACV (2022).

[59] Idrees, H., Saleemi, I., Seibert, C., Shah, M.: Multi-source multi-scale counting in extremely dense crowd images. In: CVPR (2013).

[60] H. Idrees, M. Tayyab, K. Athrey, D. Zhang, S. Al-Maadeed, N. Rajpoot, M. Shah, Composition loss for counting, density map estimation and localization in dense crowds, in: ECCV, 2018.

[61] S. Jiang, X. Lu, Y. Lei, L. Liu, Mask-aware networks for crowd counting, TCSVT (2019).

[62] X. Jiang, Z. Xiao, B. Zhang, X. Zhen, X. Cao, D. Doermann, L. Shao, Crowd counting and density estimation by trellis encoder-decoder networks, in: CVPR, 2019.

[63] X. Jiang, L. Zhang, P. Lv, Y. Guo, R. Zhu, Y. Li, Y. Pang, X. Li, B. Zhou, M. Xu, Learning multi-level density maps for crowd counting, TNNLS (2019).

[64] X. Jiang, L. Zhang, M. Xu, T. Zhang, P. Lv, B. Zhou, X. Yang, Y. Pang, Attention scaling for crowd counting, in: CVPR, 2020.

[65] X. Jiang, L. Zhang, T. Zhang, P. Lv, B. Zhou, Y. Pang, M. Xu, C. Xu, Density-aware multi-task learning for crowd counting, IEEE Transactions on Multimedia (2020).

[66] D. Kang, A. Chan, Crowd counting by adaptively fusing predictions from an image pyramid, BMVC (2018).

[67] S. Khaki, H. Pham, Y. Han, A. Kuhl, W. Kent, L. Wang, Deepcorn: A semi-supervised deep learning method for high-throughput image-based corn kernel counting and yield estimation, Knowledge-Based Systems (2021).

[68] Kong, X., Zhao, M., Zhou, H., Zhang, C.: Weakly supervised crowd-wise attention for robust crowd counting. In: ICASSP (2020).

[69] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, NIPS (2012).

[70] A. Kumar, N. Jain, S. Tripathi, C. Singh, K. Krishna, Mtcnet: Multi-task learning paradigm for crowd count estimation, IEEE AVSS (2019).

[71] Laradji, I.H., Rostamzadeh, N., Pinheiro, P.O., Vazquez, D., Schmidt, M.: Where are the blobs: Counting by localization with point supervision. In: ECCV (2018).

[72] Y. Lei, Y. Liu, P. Zhang, L. Liu, Towards using count-level weak supervision for crowd counting, Pattern Recognition (2021).

[73] V. Lempitsky, A. Zisserman, Learning to count objects in images, in: NIPS, 2010.

[74] B. Li, H. Huang, A. Zhang, P. Liu, C. Liu, Approaches on crowd counting and density estimation: a review, Pattern Analysis and Applications (2021).

[75] J. Li, Y. Xue, W. Wang, G. Ouyang, Cross-level parallel network for crowd counting, IEEE Transactions on Industrial Informatics (2019).

[76] Li, M., Zhang, Z., Huang, K., Tan, T.: Estimating the number of people in crowded scenes by mid based foreground segmentation and head-shoulder detection. In: ICPR (2008).

[77] T. Li, H. Chang, M. Wang, B. Ni, R. Hong, S. Yan, Crowded scene analysis: A survey, TCSVT (2014).

[78] Li, W., Cao, Z., Wang, Q., Chen, S., Feng, R.: Learning error-driven curriculum for crowd counting. In: ICPR (2021).

[79] Li, W., Yongbo, L., Xiangyang, X.: Coda: Counting objects via scale-aware adversarial density adaption. In: ICME (2019).

[80] Y. Li, X. Zhang, D. Chen, Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes, in: CVPR, 2018.

[81] D. Lian, X. Chen, J. Li, W. Luo, S. Gao, Locating and counting heads in crowds with a depth prior, TPAMI (2021).

[82] D. Liang, X. Chen, W. Xu, Y. Zhou, X. Bai, Transcrowd: weakly-supervised crowd counting with transformers. Science China, Information Sciences (2022).

[83] Liang, D., Xu, W., Zhu, Y., Zhou, Y.: Focal inverse distance transform maps for crowd localization and counting in dense crowd. arXiv preprint arXiv:2102.07925 (2021).

[84] H. Lin, X. Hong, Z. Ma, X. Wei, Y. Qiu, Y. Wang, Y. Gong, Direct measure matching for crowd counting, IJCAI (2021).

[85] Lin, H., Ma, Z., Ji, R., Wang, Y., Hong, X.: Boosting crowd counting via multifaceted attention. In: CVPR (2022).

[86] Z. Lin, L.S. Davis, Shape-based human detection and segmentation via hierarchical part-template matching, TPAMI (2010).

[87] M. Ling, X. Geng, Indoor crowd counting by mixture of gaussians label distribution learning, TIP (2019).

[88] Liu, C., Weng, X., Mu, Y.: Recurrent attentive zooming for joint crowd counting and precise localization. In: CVPR (2019).

[89] Liu, H., Zhao, Q., Ma, Y., Dai, F.: Bipartite matching for crowd counting with point supervision. In: IJCAI (2021).

[90] J. Liu, C. Gao, D. Meng, A.G. Hauptmann, Decidenet: Counting varying density crowds through attention guided detection and density estimation, in: CVPR, 2018.

[91] L. Liu, J. Chen, H. Wu, T. Chen, G. Li, L. Lin, Efficient crowd counting via structured knowledge transfer, in: ACM Multimedia, 2020.

[92] L. Liu, J. Chen, H. Wu, G. Li, C. Li, L. Lin, Cross-modal collaborative representation learning and a large-scale rgbt benchmark for crowd counting, in: CVPR, 2021.

[93] L. Liu, W. Jia, J. Jiang, S. Amirgholipour, Y. Wang, M. Zeibots, X. He, Denet: A universal network for counting crowd with varying densities and scales, IEEE Transactions on Multimedia (2020).

[94] Liu, L., Lu, H., Zou, H., Xiong, H., Cao, Z., Shen, C.: Weighing counts: Sequential crowd counting by reinforcement learning. In: ECCV (2020).

[95] Liu, L., Qiu, Z., Li, G., Liu, S., Ouyang, W., Lin, L.: Crowd counting with deep structured scale integration network. In: ICCV (2019).

[96] L. Liu, H. Wang, G. Li, W. Ouyang, L. Lin, Crowd counting using deep recurrent spatial-aware network, IJCAI (2018).

[97] N. Liu, Y. Long, C. Zou, Q. Niu, L. Pan, H. Wu, Adcrowdnet: An attention-injective deformable convolutional network for crowd understanding, in: CVPR, 2019.

[98] Liu, W., Durasov, N., Fua, P.: Leveraging self-supervision for cross-domain crowd counting. In: CVPR (2022).

[99] W. Liu, M. Salzmann, P. Fua, Context-aware crowd counting, in: CVPR, 2019.

[100] W. Liu, M. Salzmann, P. Fua, Counting people by estimating people flows, TPAMI (2020).

[101] Liu, W., Salzmann, M., Fua, P.: Estimating people flows to better count them in crowded scenes. In: ECCV (2020).

[102] Liu, X., Li, G., Han, Z., Zhang, W., Yang, Y., Huang, Q., Sebe, N.: Exploiting sample correlation for crowd counting with multi-expert network. In: ICCV (2021).

[103] X. Liu, J. Van De Weijer, A.D. Bagdanov, Leveraging unlabeled data for crowd counting by learning to rank, in: CVPR, 2019.

[104] Liu, X., Yang, J., Ding, W., Wang, T., Wang, Z., Xiong, J.: Adaptive mixture regression network with local counting map for crowd counting. In: ECCV (2020).

[105] Liu, Y., Liu, L., Wang, P., Zhang, P., Lei, Y.: Semi-supervised crowd counting via self-training on surrogate tasks. In: ECCV (2020).

[106] Y. Liu, M. Shi, Q. Zhao, X. Wang, Point in, box out: Beyond counting persons in crowds, in: CVPR, 2019.

[107] Y. Liu, Z. Wang, M. Shi, S. Satoh, Q. Zhao, H. Yang, Towards unsupervised crowd counting via regression-detection bi-knowledge transfer, in: ACM Multimedia, 2020.

[108] Y. Liu, Q. Wen, H. Chen, W. Liu, J. Qin, G. Han, S. He, Crowd counting via cross-stage refinement networks, IEEE Transactions on Image Processing (2020).

[109] J.L. Louëdec, G. Cielniak, Gaussian map predictions for 3d surface feature localisation and counting, BMVC (2021).

[110] Luo, A., Yang, F., Li, X., Nie, D., Jiao, Z., Zhou, S., Cheng, H.: Hybrid graph neural networks for crowd counting. In: AAAI (2020).

[111] J. Ma, Y. Dai, Y.P. Tan, Atrous convolutions spatial pyramid network for crowd counting and density estimation, Neurocomputing 350 (2019) 91–101.

[112] Y.J. Ma, H.H. Shuai, W.H. Cheng, Spatiotemporal dilated convolution with uncertain matching for video-based crowd estimation, IEEE Transactions on Multimedia (2021).

[113] Ma, Z., Hong, X., Wei, X., Qiu, Y., Gong, Y.: Towards a universal model for cross-dataset crowd counting. In: ICCV (2021).

[114] Ma, Z., Wei, X., Hong, X., Gong, Y.: Bayesian loss for crowd count estimation with point supervision. In: ICCV (2019).

[115] Z. Ma, X. Wei, X. Hong, Y. Gong, Learning scales from points: A scale-aware probabilistic model for crowd counting, in: ACM Multimedia, 2020.

[116] Ma, Z., Wei, X., Hong, X., Lin, H., Qiu, Y., Gong, Y.: Learning to count via unbalanced optimal transport. In: AAAI (2021).

[117] Mao, J., Niu, M., Bai, H., Liang, X., Xu, H., Xu, C.: Pyramid r-cnn: Towards better performance and adaptability for 3d object detection. In: ICCV (2021).

[118] Mao, J., Niu, M., Jiang, C., Liang, H., Chen, J., Liang, X., Li, Y., Ye, C., Zhang, W., Li, Z., et al.: One million scenes for autonomous driving: Once dataset. arXiv preprint arXiv:2106.11037 (2021).

[119] M. Marsden, K. McGuinness, S. Little, N.E. O'Connor, Fully convolutional crowd counting on highly congested scenes, VISAPP (2016).

[120] Marsden, M., McGuinness, K., Little, S., O'Connor, N.E.: Resnetcrowd: A residual deep learning architecture for crowd counting, violent behaviour detection and crowd density level classification. In: AVSS (2017).

[121] C. Meng, E. Liu, W. Neiswanger, J. Song, M. Burke, D. Lobell, S. Ermon, Is-count: Large-scale object counting from satellite images with covariate-based importance sampling, AAAI (2021).

[122] Meng, Y., Zhang, H., Zhao, Y., Yang, X., Qian, X., Huang, X., Zheng, Y.: Spatial uncertainty-aware semi-supervised crowd counting. In: ICCV (2021).

[123] Miao, Y., Lin, Z., Ding, G., Han, J.: Shallow feature based dense attention network for crowd counting. In: AAAI (2020).

[124] H. Mo, W. Ren, Y. Xiong, X. Pan, Z. Zhou, X. Cao, W. Wu, Background noise filtering and distribution dividing for crowd counting, IEEE Transactions on Image Processing (2020).
[125] Modolo, D., Shuai, B., Varior, R.R., Tighe, J.: Understanding the impact of mistakes on background regions in crowd counting. In: WACV (2021).
[126] Oh, M.h., Olsen, P.A., Ramamurthy, K.N.: Crowd counting with decomposed uncertainty. In: AAAI (2020).
[127] Olmschenk, G., Chen, J., Tang, H., Zhu, Z.: Dense crowd counting convolutional neural networks with minimal data using semi-supervised dual-goal generative adversarial networks. In: CVPR Workshops (2019).
[128] Olmschenk, G., Tang, H., Zhu, Z.: Crowd counting with minimal data using generative adversarial networks for multiple target regression. In: WACV (2018).
[129] G. Olmschenk, H. Tang, Z. Zhu, Improving dense crowd counting convolutional neural networks using inverse k-nearest neighbor maps and multiscale upsampling, VISAPP (2019).
[130] G. Olmschenk, Z. Zhu, H. Tang, Generalizing semi-supervised generative adversarial networks to regression using feature contrasting, CVIU (2019).
[131] Onoro-Rubio, D., López-Sastre, R.J.: Towards perspective-free object counting with deep learning. In: ECCV (2016).
[132] Pan, X., Mo, H., Zhou, Z., Wu, W.: Attention guided region division for crowd counting. In: ICASSP (2020).
[133] Pham, V.Q., Kozakaya, T., Yamaguchi, O., Okada, R.: Count forest: Co-voting uncertain number of targets using random forest for crowd density estimation. In: ICCV (2015).
[134] Z. Qiu, L. Liu, G. Li, Q. Wang, N. Xiao, L. Lin, Crowd counting via multi-view scale aggregation networks, in: ICME, 2019.
[135] V. Rabaud, S. Belongie, Counting crowded moving objects, in: CVPR, 2006.
[136] V. Ranjan, H. Le, M. Hoai, Iterative crowd counting, In: ECCV (2018).
[137] Ranjan, V., Wang, B., Shah, M., Hoai, M.: Uncertainty estimation and sample selection for crowd counting. In: ACCV (2020).
[138] Reddy, M.K.K., Hossain, M., Rochan, M., Wang, Y.: Few-shot scene adaptive crowd counting using meta-learning. In: WACV (2020).
[139] M.K.K. Reddy, M. Rochan, Y. Lu, Y. Wang, Adacrowd: unlabeled scene adaptation for crowd counting, IEEE Transactions on Multimedia (2021).
[140] W. Ren, X. Wang, J. Tian, Y. Tang, A.B. Chan, Tracking-by-counting: Using network flows on crowd density maps for tracking multiple targets, IEEE Transactions on Image Processing (2020).
[141] Rong, L., Li, C.: Coarse-and fine-grained attention network with background-aware loss for crowd density map estimation. In: WACV (2021).
[142] Sajid, U., Chen, X., Sajid, H., Kim, T., Wang, G.: Audio-visual transformer based crowd counting. In: ICCV (2021).
[143] Sajid, U., Ma, W., Wang, G.: Multi-resolution fusion and multi-scale input priors based crowd counting. In: ICPR (2021).
[144] Sajid, U., Wang, G.: Plug-and-play rescaling based crowd counting in static images. In: WACV (2020).
[145] U. Sajid, G. Wang, Towards more effective prm-based crowd counting via a multi-resolution fusion and attention network, Neurocomputing 474 (2022) 13–24.
[146] Sam, D.B., Agarwalla, A., Joseph, J., Sindagi, V.A., Babu, R.V., Patel, V.M.: Completely self-supervised crowd counting via distribution matching. arXiv preprint arXiv:2009.06420 (2020).
[147] D.B. Sam, S.V. Peri, M.N. Sundararaman, A. Kamath, R.V. Babu, Locate, size, and count: accurately resolving people in dense crowds via detection, TPAMI (2020).
[148] Sam, D.B., Sajjan, N.N., Maurya, H., Babu, R.V.: Almost unsupervised learning for dense crowd counting. In: AAAI (2019).
[149] Servadei, L., Sun, H., Ott, J., Stephan, M., Hazra, S., Stadelmayer, T., Lopera, D.S., Wille, R., Santra, A.: Label-aware ranked loss for robust people counting using automotive in-cabin radar. In: ICASSP (2022).
[150] Shang, C., Ai, H., Bai, B.: End-to-end crowd counting via joint learning local and global count. In: ICIP (2016).
[151] Z. Shen, Y. Xu, B. Ni, M. Wang, J. Hu, X. Yang, Crowd counting via adversarial cross-scale consistency pursuit, in: CVPR, 2018.
[152] Shi, X., Li, X., Wu, C., Kong, S., Yang, J., He, L.: A real-time deep network for crowd counting. In: ICASSP (2020).
[153] Shi, Z., Mettes, P., Snoek, C.G.: Counting with focus for free. In: ICCV (2019).
[154] Z. Shi, L. Zhang, Y. Liu, X. Cao, Y. Ye, M.M. Cheng, G. Zheng, Crowd counting with deep negative correlation learning, in: CVPR, 2018.
[155] S.V. Shivapuja, M.P. Khamkar, D. Bajaj, G. Ramakrishnan, R.K. Sarvadevabhatla, Wisdom of (binned) crowds: A bayesian stratification paradigm for crowd counting, in: ACM Multimedia, 2021.
[156] Sindagi, V.A., Patel, V.M.: Cnn-based cascaded multi-task learning of high-level prior and density estimation for crowd counting. In: AVSS (2017).
[157] Sindagi, V.A., Patel, V.M.: Generating high-quality crowd density maps using contextual pyramid cnns. In: ICCV (2017).
[158] V.A. Sindagi, V.M. Patel, A survey of recent advances in cnn-based single image crowd counting and density estimation, Pattern Recognition Letters (2018).
[159] V.A. Sindagi, V.M. Patel, Ha-ccn: Hierarchical attention-based crowd counting network, TIP (2019).
[160] Sindagi, V.A., Patel, V.M.: Multi-level bottom-top and top-bottom feature fusion for crowd counting. In: ICCV (2019).
[161] Sindagi, V.A., Yasarla, R., Babu, D.S., Babu, R.V., Patel, V.M.: Learning to count in the crowd from limited labeled data. In: ECCV (2020).
[162] Sindagi, V.A., Yasarla, R., Patel, V.M.: Pushing the frontiers of unconstrained crowd counting: New dataset and benchmark method. In: ICCV (2019).
[163] Song, Q., Wang, C., Jiang, Z., Wang, Y., Tai, Y., Wang, C., Li, J., Huang, F., Wu, Y.: Rethinking counting and localization in crowds: A purely point-based framework. In: ICCV (2021).
[164] Song, Q., Wang, C., Wang, Y., Tai, Y., Wang, C., Li, J., Wu, J., Ma, J.: To choose or to fuse? scale selection for crowd counting. In: AAAI (2021).
[165] X. Tan, C. Tao, T. Ren, J. Tang, G. Wu, Crowd counting via multi-layer regression, in: ACM Multimedia, 2019.
[166] H. Tang, Y. Wang, L.P. Chau, Tafnet: A three-stream adaptive fusion network for rgb-t crowd counting, ISCAS (2022).
[167] T. Teixeira, G. Dublon, A. Savvides, A survey of human-sensing: Methods for detecting presence, count, location, track, and identity, ACM Computing Surveys (2010).
[168] Thanasutives, P., Fukui, K.i., Numao, M., Kijsirikul, B.: Encoder-decoder based convolutional neural networks with multi-scale-aware modules for crowd counting. In: ICPR (2021).
[169] E. Walach, L. Wolf, Learning to count with cnn boosting, In: ECCV (2016).
[170] Wan, J., Chan, A.: Adaptive density map generation for crowd counting. In: ICCV (2019).
[171] J. Wan, A. Chan, Modeling noisy annotations for crowd counting, NeurIPS (2020).
[172] J. Wan, N.S. Kumar, A.B. Chan, Fine-grained crowd counting, in: IEEE transactions on image processing, 2021.
[173] Wan, J., Liu, Z., Chan, A.B.: A generalized loss function for crowd counting and localization. In: CVPR (2021).
[174] J. Wan, Q. Wang, A.B. Chan, Kernel-based density map generation for dense object counting, TPAMI (2020).
[175] B. Wang, H. Liu, D. Samaras, M.H. Nguyen, Distribution matching for crowd counting, NeurIPS (2020).
[176] Wang, C., Song, Q., Zhang, B., Wang, Y., Tai, Y., Hu, X., Wang, C., Li, J., Ma, J., Wu, Y.: Uniformity in heterogeneity: Diving deep into count interval partition for crowd counting. In: ICCV (2021).
[177] C. Wang, H. Zhang, L. Yang, S. Liu, X. Cao, Deep people counting in extremely dense crowds, in: ACM Multimedia, 2015.
[178] F. Wang, J. Sang, Z. Wu, Q. Liu, N. Sang, Hybrid attention network based on progressive embedding scale-context for crowd counting, Information Sciences (2022).
[179] L. Wang, B. Yin, X. Tang, Y. Li, Removing background interference for crowd counting via de-background detail convolutional network, Neurocomputing 332 (2019) 360–371.
[180] M. Wang, H. Cai, X. Han, J. Zhou, M. Gong, Stnet: Scale tree network with multi-level auxiliator for crowd counting, IEEE Transactions on Multimedia (2022).
[181] M. Wang, H. Cai, J. Zhou, M. Gong, Interlayer and intralayer scale aggregation for scale-invariant crowd counting, Neurocomputing 441 (2021) 128–137.
[182] Wang, M., Zhou, J., Cai, H., Gong, M.: Crowdmlp: Weakly-supervised crowd counting via multi-granularity mlp. arXiv preprint arXiv:2203.08219 (2022).
[183] P. Wang, C. Gao, Y. Wang, H. Li, Y. Gao, Mobilecount: An efficient encoder-decoder framework for real-time crowd counting, Neurocomputing 407 (2020) 292–299.
[184] Q. Wang, T.P. Breckon, Crowd counting via segmentation guided attention networks and curriculum loss, IEEE Transactions on Intelligent Transportation Systems (2022).
[185] Q. Wang, J. Gao, W. Lin, X. Li, Nwpu-crowd: A large-scale benchmark for crowd counting and localization, TPAMI (2020).
[186] Q. Wang, J. Gao, W. Lin, Y. Yuan, Learning from synthetic data for crowd counting in the wild, in: CVPR, 2019.
[187] Q. Wang, J. Gao, W. Lin, Y. Yuan, Pixel-wise crowd understanding via synthetic data, IJCV (2020).
[188] Q. Wang, J. Gao, W. Lin, Y. Yuan, Pixel-wise crowd understanding via synthetic data, IJCV (2021).
[189] Q. Wang, T. Han, J. Gao, Y. Yuan, Neuron linear transformation: Modeling the domain shift for crowd counting, TNNLS (2021).
[190] Q. Wang, W. Lin, J. Gao, X. Li, Density-aware curriculum learning for crowd counting. IEEE Transactions on, Cybernetics (2020).
[191] Y. Wang, J. Hou, X. Hou, L.P. Chau, A self-training approach for point-supervised object detection and counting in crowds, IEEE Transactions on Image Processing (2021).
[192] Wang, Y., Hou, X., Chau, L.P.: Dense point prediction: A simple baseline for crowd counting and localization. In: ICMEW (2021).
[193] Y. Wang, Z. Ma, X. Wei, S. Zheng, Y. Wang, X. Hong, Eccnas: Efficient crowd counting neural architecture search, TOMM (2022).
[194] Z. Wang, A.C. Bovik, H.R. Sheikh, E.P. Simoncelli, Image quality assessment: from error visibility to structural similarity, TIP (2004).
[195] Z. Wang, Z. Xiao, K. Xie, Q. Qiu, X. Zhen, X. Cao, X.: In defense of single-column networks for crowd counting, BMVC (2018).
[196] Wei, B., Yuan, Y., Wang, Q.: Mspnet: Multi-supervised parallel network for crowd counting. In: ICASSP (2020).
[197] X. Wei, J. Du, M. Liang, L. Ye, Boosting deep attribute learning via support vector regression for fast moving crowd counting, Pattern Recognition Letters (2019).
[198] L. Wen, D. Du, P. Zhu, Q. Hu, Q. Wang, L. Bo, S. Lyu, Detection, tracking, and counting meets drones in crowds: A benchmark, in: CVPR, 2021.
[199] Q. Wu, J. Wan, A.B. Chan, Dynamic momentum adaptation for zero-shot cross-domain crowd counting, in: ACM Multimedia, 2021.

[200] X. Wu, B. Xu, Y. Zheng, H. Ye, J. Yang, L. He, Fast video crowd counting with a temporal aware network, Neurocomputing 403 (2020) 13–20.
[201] X. Wu, Y. Zheng, H. Ye, W. Hu, T. Ma, J. Yang, L. He, Counting crowds with varying densities via adaptive scenario discovery framework, Neurocomputing 397 (2020) 127–138.
[202] Wu, X., Zheng, Y., Ye, H., Hu, W., Yang, J., He, L.: Adaptive scenario discovery for crowd counting. In: ICASSP (2019).
[203] Z. Wu, J. Sang, Y. Shi, Q. Liu, N. Sang, X. Liu, X.: Cranet: Cascade residual attention network for crowd counting, in: ICME, 2021.
[204] Xiong, F., Shi, X., Yeung, D.Y.: Spatiotemporal modeling for crowd counting in videos. In: ICCV (2017).
[205] Xiong, H., Lu, H., Liu, C., Liu, L., Cao, Z., Shen, C.: From open set to closed set: Counting objects by spatial divide-and-conquer. In: ICCV (2019).
[206] C. Xu, D. Liang, Y. Xu, S. Bai, W. Zhan, X. Bai, M. Tomizuka, Autoscale: Learning to scale for crowd counting, IJCV (2022).
[207] W. Xu, D. Liang, Y. Zheng, J. Xie, Z. Ma, Dilated-scale-aware category-attention convnet for multi-class object counting, IEEE Signal Processing Letters (2021).
[208] Xu, Y., Zhong, Z., Lian, D., Li, J., Li, Z., Xu, X., Gao, S.: Crowd counting with partial annotations in an image. In: ICCV (2021).
[209] Z. Yan, P. Li, B. Wang, D. Ren, W. Zuo, Towards learning multi-domain crowd counting, IEEE Trans. Circuits Syst, Video Technol, 2021.
[210] Yan, Z., Yuan, Y., Zuo, W., Tan, X., Wang, Y., Wen, S., Ding, E.: Perspective-guided convolution networks for crowd counting. In: ICCV (2019).
[211] Z. Yan, R. Zhang, H. Zhang, Q. Zhang, W. Zuo, Crowd counting via perspective-guided fractional-dilation convolution, IEEE Transactions on Multimedia (2021).
[212] B. Yang, W. Zhan, N. Wang, X. Liu, J. Lv, Counting crowds using a scale-distribution-aware network and adaptive human-shaped kernel, Neurocomputing 390 (2020) 207–216.
[213] Yang, J., Zhou, Y., Kung, S.Y.: Multi-scale generative adversarial networks for crowd counting. In: ICPR (2018).
[214] Yang, S.D., Su, H.T., Hsu, W.H., Chen, W.C.: Class-agnostic few-shot object counting. In: WACV (2021).
[215] Y. Yang, G. Li, D. Du, Q. Huang, N. Sebe, Embedding perspective analysis into multi-column convolutional neural network for crowd counting, IEEE Transactions on Image Processing (2020).
[216] Y. Yang, G. Li, Z. Wu, L. Su, Q. Huang, N. Sebe, Reverse perspective network for perspective-aware object counting, in: CVPR, 2020.
[217] Yang, Y., Li, G., Wu, Z., Su, L., Huang, Q., Sebe, N.: Weakly-supervised crowd counting learns from sorting rather than locations. In: ECCV (2020).
[218] Zand, M., Damirchi, H., Farley, A., Molahasani, M., Greenspan, M., Etemad, A.: Multiscale crowd counting and localization by multitask point supervision. In: ICASSP (2022).
[219] Zhang, A., Shen, J., Xiao, Z., Zhu, F., Zhen, X., Cao, X., Shao, L.: Relational attention network for crowd counting. In: ICCV (2019).
[220] Zhang, A., Yue, L., Shen, J., Zhu, F., Zhen, X., Cao, X., Shao, L.: Attentional neural fields for crowd counting. In: ICCV (2019).
[221] Zhang, C., Li, H., Wang, X., Yang, X.: Cross-scene crowd counting via deep convolutional neural networks. In: CVPR (2015).
[222] Zhang, L., Shi, M., Chen, Q.: Crowd counting via scale-adaptive convolutional neural network. In: WACV (2018).
[223] L. Zhang, Z. Shi, M.M. Cheng, Y. Liu, J.W. Bian, J.T. Zhou, G. Zheng, Z. Zeng, Nonlinear regression via deep negative correlation learning, TPAMI (2019).
[224] Zhang, Q., Chan, A.B.: Wide-area crowd counting via ground-plane density maps and multi-view fusion cnns. In: CVPR (2019).
[225] Zhang, Q., Chan, A.B.: 3d crowd counting via multi-view fusion with 3d gaussian kernels. In: AAAI (2020).
[226] Zhang, Q., Lin, W., Chan, A.B.: Cross-view cross-scene multi-view crowd counting. In: CVPR (2021).
[227] Y. Zhang, C. Zhou, F. Chang, A.C. Kot, Multi-resolution attention convolutional neural network for crowd counting, Neurocomputing 329 (2019) 144–152.
[228] Y. Zhang, D. Zhou, S. Chen, S. Gao, Y. Ma, Single-image crowd counting via multi-column convolutional neural network, in: CVPR, 2016.
[229] M. Zhao, C. Zhang, J. Zhang, F. Porikli, B. Ni, W. Zhang, Scale-aware crowd counting via depth-embedded convolutional neural networks, TCSVT (2019).
[230] Zhao, Z., Han, T., Gao, J., Wang, Q., Li, X.: A flow base bi-path network for cross-scene video crowd understanding in aerial view. In: ECCV (2020).
[231] Zhao, Z., Shi, M., Zhao, X., Li, L.: Active crowd counting with limited supervision. In: ECCV (2020).
[232] Zheng, L., Li, Y., Mu, Y.: Learning factorized cross-view fusion for multi-view crowd counting. In: ICME (2021).
[233] X. Zhong, Z. Yan, J. Qin, W. Zuo, W. Lu, An improved normed-deformable convolution for crowd counting, SPL (2022).
[234] Zhou, J.T., Zhang, L., Du, J., Peng, X., Fang, Z., Xiao, Z., Zhu, H.: Locality-aware crowd counting. TPAMI (2021).
[235] Q. Zhou, J. Zhang, L. Che, H. Shan, J.Z. Wang, Crowd counting with limited labeling through submodular frame selection, IEEE Transactions on Intelligent Transportation Systems (2018).
[236] Y. Zhou, J. Yang, H. Li, T. Cao, S.Y. Kung, Adversarial learning for multiscale crowd counting under complex scenes, in: IEEE transactions on cybernetics, 2020.
[237] P. Zhu, T. Peng, D. Du, H. Yu, L. Zhang, Q. Hu, Graph regularized flow attention network for video animal counting from drones, IEEE Transactions on Image Processing (2021).
[238] Zhu, P., Wen, L., Du, D., Bian, X., Ling, H., Hu, Q., Wu, H., Nie, Q., Cheng, H., Liu, C., et al.: Visdrone-vdt2018: The vision meets drone video detection and tracking challenge results. In: ECCV Workshops (2018).
[239] M.S. Zitouni, H. Bhaskar, J. Dias, M.E. Al-Mualla, Advances and trends in visual crowd analysis: A systematic survey and evaluation of crowd modelling techniques, Neurocomputing 186 (2016) 139–159.
[240] Z. Zou, Y. Cheng, X. Qu, S. Ji, X. Guo, P. Zhou, Attend to count: Crowd counting with adaptive capacity multi-scale cnns, Neurocomputing (2019).
[241] Z. Zou, Y. Liu, S. Xu, W. Wei, S. Wen, P. Zhou, Crowd counting via hierarchical scale recalibration network, ECAI (2020).
[242] Z. Zou, H. Shao, X. Qu, W. Wei, P. Zhou, Enhanced 3d convolutional networks for crowd counting, BMVC (2019).

Haoyue Bai received the B.Eng degree in information engineering in 2018, from Zhejiang University, China. She is currently working as a postgraduate student at the department of computer science and engineering, the Hong Kong University of Science and Technology. Her research interests include machine learning, computer vision, and video analysis.

Jiageng Mao received the B.Eng degree in Information and Electronic Engineering in 2018, from Zhejiang University, China. He is currently working as a research assistant at the department of Electronic Engineering, the Chinese University of Hong Kong. His research interests include 3D vision, computer vision, and autonomous driving.

S.-H. Gary Chan (Senior Member, IEEE) received the B.S.E. degree (Hons.) in electrical engineering from Princeton University, Princeton, NJ, USA, in 1993, with certificates in applied and computational mathematics, engineering physics, and engineering and management systems, and the MSE and Ph.D. degrees in electrical engineering from Stanford University, Stanford, CA, USA, in 1994 and 1999, respectively, with a minor in business administration. He is currently a Professor with the Department of Computer Science and Engineering, The Hong Kong University of Science and Technology (HKUST), Hong Kong. His research interests include smart sensing and IoT, cloud and fog/edge computing, indoor positioning and mobile computing, video/location/user/data analytics, and IT entrepreneurship.