CNN-based Single Image Crowd Counting: Network Design, Loss Function and Supervisory Signal

HAOYUE BAI, The Hong Kong University of Science and Technology
S.-H. GARY CHAN, The Hong Kong University of Science and Technology

Single image crowd counting is a challenging computer vision problem with wide applications in public safety, city planning, traffic management, etc. This survey is to provide a comprehensive summary of recent advanced crowd counting techniques based on Convolutional Neural Network (CNN) via density map estimation. 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 conclude the survey with some future directions.

Additional Key Words and Phrases: Crowd counting, recent advances, network design, loss function, supervisory signal

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. [4, 29, 55]. In this survey, we mainly focus on people as the crowd, though the techniques discussed may be applicable in other domains.

An early approach to count people is based on detection-based computer vision techniques, which is to detect individual objects, heads, or body-part and then count the total number in the image [31, 35, 58]. 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 [5, 6, 76], 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 the number of people in a pixel of the image [2, 3, 29, 34, 41, 56, 97]. Such approaches achieve high accuracy for crowded scenes and preserve spatial information of people distribution.

We summarize by comparing the aforementioned three major crowd counting approaches in Table 1. All of them require image annotation through labeling in the training step. For the detection-based approach, each object has to be fully identified and outlined, which incurs the highest labeling cost. On the other hand, the regression-based approach does not need to annotate individual objects but the total object count, and hence its annotation cost is the lowest. Density map 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 Convolutional Neural Networks (CNN) based models as compared with approaches based on handcrafted features. We overview in

Authors' addresses: Haoyue Bai, The Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong; S.-H. Gary Chan, The Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong.

Table 1. Summary of crowd counting approaches.

Category	Principles	Crowd Counting Accuracy	Location Accuracy	Annotation Complexity	Limitations
Detection-based	Detect then count; early approach	Low	High	High (object framing)	Low accuracy for highly crowded scenes
Regression-based	Directly learn to regress the count	Medium	N/A	Low (image-level count)	Less interpretable; lacks location information
Density map estimation	Compute number of people per pixel	High	Medium	Medium (head indication)	Low accuracy in low crowd scenes

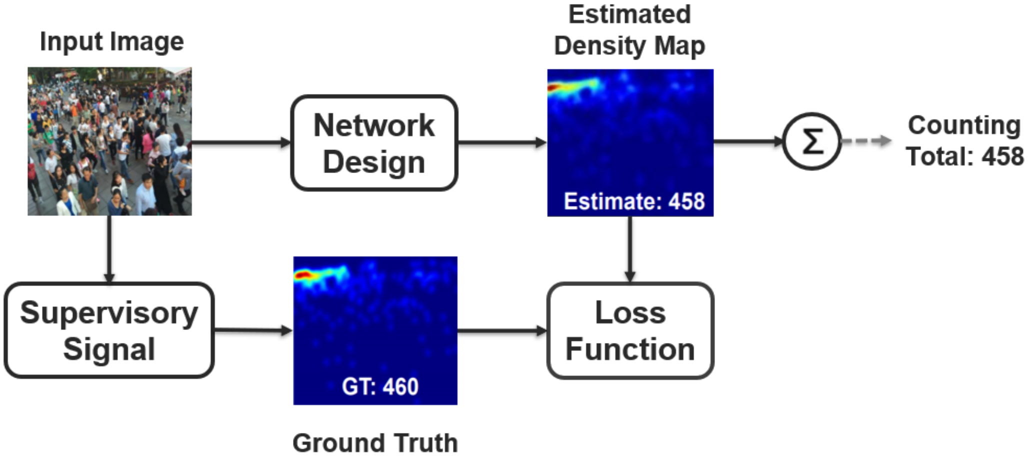

Fig. 1. Overview of CNN-based single image crowd counting methods via density map estimation.

Fig. 1 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 density map (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 CNN-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 previous related work as follows. Teixeira *et al.* [72] is 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. Li *et al.* [32] reviews crowd scene analysis in terms of crowd behavior, activity analysis, and anomaly detection, with crowd

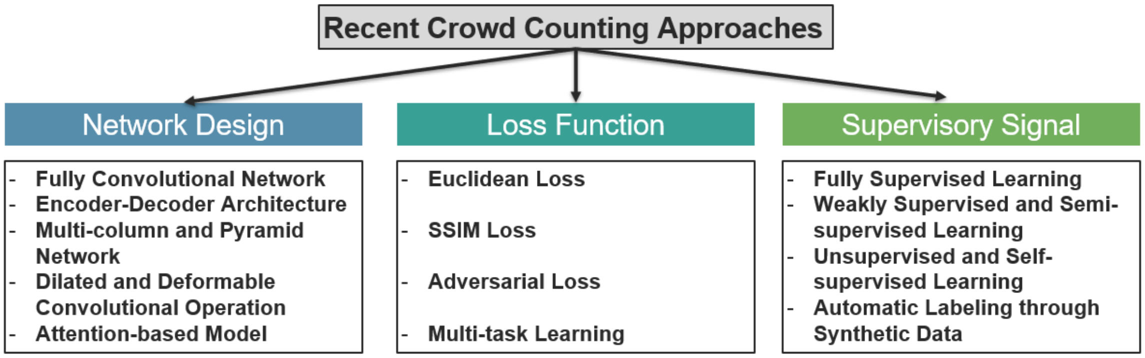
Recent Crowd Counting Approaches
Network Design
- Fully Convolutional Network
- Encoder-Decoder Architecture
- Multi-column and Pyramid Network
- Dilated and Deformable Convolutional Operation
- Attention-based Model
Loss Function
- Euclidean Loss
- SSIM Loss
- Adversarial Loss
- Multi-task Learning
Supervisory Signal
- Fully Supervised Learning
- Weakly Supervised and Semi-supervised Learning
- Unsupervised and Self-supervised Learning
- Automatic Labeling through Synthetic Data

Fig. 2. The structure of this survey. Recent advances on crowd counting schemes mainly pertain to deep neural network design, loss function and supervisory signal.

counting playing a small role. Though Zitouni *et al.* [99] 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. Sindagi *et al.* [67] surveys on CNN-based approach for crowd counting, but has not covered the hot recent advances on network design (e.g., advanced convolutional operations and attention-based model), loss function, and supervisory signal as we discuss here. In contrast with previous papers, our work comprehensively summarizes more than a hundred CNN-based crowd counting algorithms in the recent five years. Our work is of value, because we survey the popular and critical design components of this active field and provide an in-depth illustration of the representative schemes in the area.

Figure 2 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, adversarial 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, unsupervised 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 in the dataset:* 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 the 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.

We identify some common datasets used in the research community, and extract and present some typical images from the datasets in Fig. 3. We also compare them in Table 2. These datasets are elaborated below:

- *NWPU-Crowd* [77] 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.
- *UCF-QNRF* [23] contains 1,535 challenging images and a total of 1,251,642 annotations. The minimum and 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* [61] [79] 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 $15,212$ 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* [14] 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* [97] 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* [91] 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* [22] 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 five-fold cross-validation. The maximum resolution was reduced to 1024 for efficient computation.

Fig. 3. Some typical crowd scenes of publicly available datasets.

Table 2. An overview of datasets statistics for crowd counting.

Dataset	Year	Average Resolution	Number of Images	Total	Min Count	Max Count	Average Count
NWPU-Crowd [77]	2020	2191 × 3209	5,109	2,133,375	0	20,033	418
UCF-QNRF [23]	2019	2013 × 2902	1,535	1,251,642	49	12,865	815
GCC [61]	2019	1080×1920	15,212	7,625,843	0	3,995	501
Fudan-ShanghaiTech [14]	2019	1080 × 1920	15,000	394,081	9	57	27
ShanghaiTech Part A [97]	2016	589 × 868	482	241,677	33	3,139	501
ShanghaiTech Part B [97]	2016	768 × 1024	716	88,488	9	578	124
WorldExpo'10 [91]	2015	576 × 720	3,980	199,923	1	253	50
UCF_CC_50 [22]	2013	2101 × 2888	50	63,974	94	4,543	1,280
Mall [7]	2012	240 × 320	2,000	62,325	13	53	31
UCSD [4]	2008	158 × 238	2,000	49,885	11	46	25

- *Mall* [7] 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 [4]. 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* [4] 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.

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 [2]. 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) and Mean Squared Error (MSE) are commonly used, defined respectively below:

    $$MAE = \frac{1}{N}\sum_{i=1}^{N}|C_i - \hat{C}_i|, \tag{1}$$

    $$MSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}|C_i - \hat{C}_i|^2}, \tag{2}$$

    where N is the total number of test images, C_i the ground truth count of the i-th image, and $\hat{C}_i$ 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) are often used:

    $$PSNR = 10\lg(\frac{MAX_I^2}{MSE}), \tag{3}$$

    $$SSIM(x,y) = \left[l(x,y)^{\alpha} \cdot c(x,y)^{\beta} \cdot s(x,y)^{\gamma}\right], \tag{4}$$

    where MAX_I is the maximum pixel value of Image I, $l(x,y)$ is luminance given by

    $$l(x,y) = \frac{2\mu_x\mu_y + c_1}{\mu_x^2 + \mu_y^2 + c_1};$$

    $c(x,y)$ is contrast given by

    $$c(x,y) = \frac{2\sigma_x\sigma_y + c_2}{\sigma_x^2 + \sigma_x^2 + c_2};$$

    $s(x,y)$ is structure given by

    $$s(x,y) = \frac{\sigma_{xy} + c_3}{\sigma_x\sigma_y + c_3}.$$

    Setting $\alpha = \beta = \gamma = 1$ results in a specific form of SSIM index:

    $$SSIM(x,y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}. \tag{5}$$

- *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

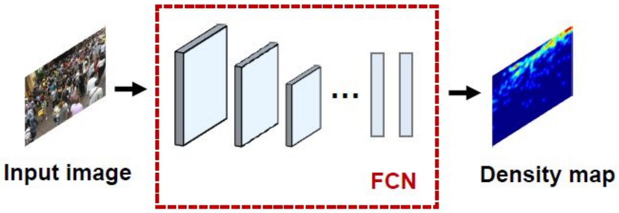

Fig. 4. Illustration of Fully Convolutional Networks for crowd counting.

(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 object).
  - 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 and pyramid network (Section 3.3), dilated and deformable convolution operations (Section 3.4), and attention-based model (Section 3.5). We compare these approaches in Section 3.6, and remark on some other emerging approaches in Section 3.7.

3.1 Fully Convolutional Network

An early CNN-based density map estimation approach is based on a fully convolutional network (FCN) [48], 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, FCN learns an end-to-end mapping from an input image to the corresponding density map and produces a proportionally sizeddensity 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. 5, the encoder-decoder network consists an encoder and a decoder: an encoder network takes the input

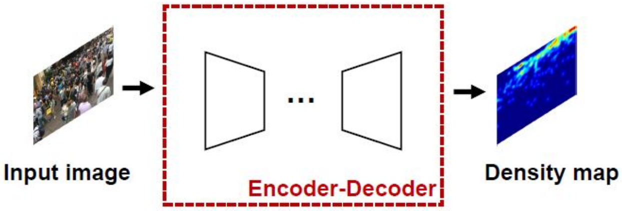

Fig. 5. The architecture of Encoder-Decoder based model.

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.

Recently, many CNN-based crowd counting approaches are following encoder-decoder structure (see, for examples, [3, 24, 40, 90]). SANet [3] 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 and Pyramid 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 [1, 10, 30, 42, 83, 84, 87]. Besides, some relevant techniques usually used together with the multi-column and pyramid networks to enhance the multi-scale feature extraction process such as skip-connections [11, 49, 69, 70, 80] and dense blocks [23, 25, 46, 51, 57].

There are two important elements in the architecture:

- *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 [97] and McML [9]. As shown in Fig. 6, multi-column neural network (MCNN) consists of multiple branches with different kernel sizes (e.g., 5×5, 7×7 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.
- *Pyramid architecture* is yet another approach to address scale variations (e.g., AFP [26] and CP-CNN [66]), which mainly consists of two subgroups, image pyramid, and feature pyramid pooling. For the image pyramid-based model, as Fig. 7 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 adaptive fusing the prediction from different scales. However, this kind of architecture remains a high computational complexity.

3.4 Dilated and Deformable Convolutional Operations

There is a trend to leverage advanced convolutional operations to facilitate accurate crowd counting model and better CNN feature learning. The CNN-based single image crowd counting model benefits a lot from the dilated and deformable convolutional operations. This can replace the traditional convolutional operations in the architectures discussed above.

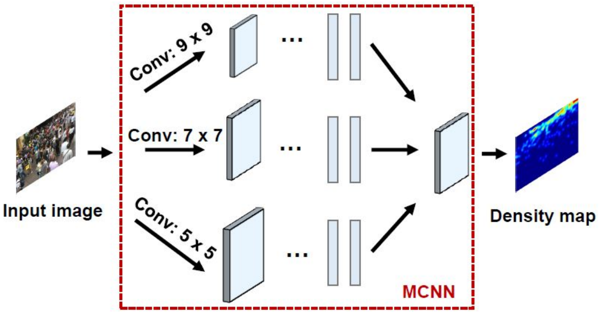

Fig. 6. Multi-column based architecture for crowd counting.

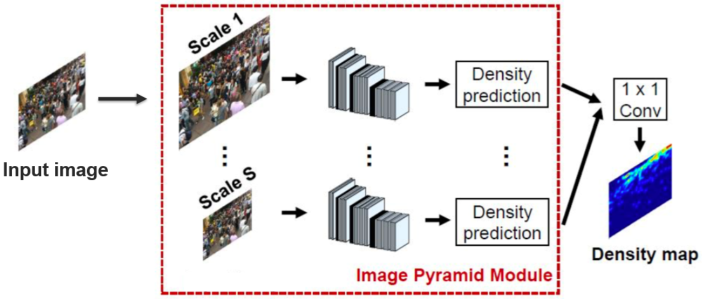

Fig. 7. Image pyramid-based model for crowd counting.

There are two important advanced convolutional 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, as Fig. 8 shows, 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) [19], Scale Pyramid Network (SPN) [8], Atrous convolutions spatial pyramid network (ACSPNet) [46], DENet [38], Dilated Convolutional Neural Networks (CSRNet) [34] and An Aggregated Multicolumn Dilated Convolution Network (AMDCNet) [10]. 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 locations 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. As shown in Fig. 9, the first subfigure illustrates the traditional convolutional operation and the second subfigure represents the deformable convolution. The red arrows present the additional offsets. Some recent literatures replace the traditional convolutions with the deformable convolutions

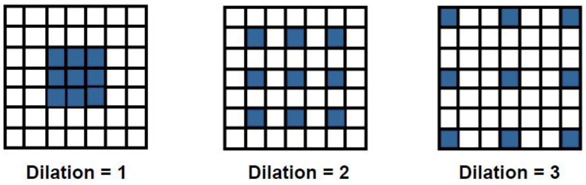

Fig. 8. Dilated convolutions to enlarge the receptive field for crowd counting models.

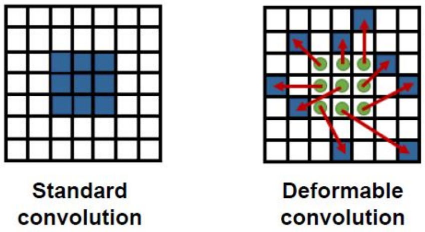

Fig. 9. The details of the deformable convolutional operations.

and achieves superiors performance: Dilated-Attention-Deformable ConvNet (DADNet) [17], An Attention-injective Deformable Convolutional Network (ADCrowdNet) [41], and the Deformation Aggregation Network (DA-Net) [103]. However, the deformable convolutional operations require high computational complexity.

3.5 Attention-based Model

Attention mechanisms can be roughly divided into two subgroups: hard attention and soft attention. Such mechanisms have been explicitly explored in recent years, and we summarize several recent algorithms applied with the attention mechanism: AFPNet [26], MRA-CNN [96], SAAN [20], DADNet [17], Relational Attention Network [89], Hierarchical Scale Recalibration Network [101], ACM-CNN [100], HA-CNN [68], Shallow Feature-based Dense Attention Network [50] and Multi-supervised Parallel Network [81].

Spatial-/Channel-wise Attention Regression Networks (SCAR) [16] 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. 10, the top half branch represents the spatial-wise attention and the bottom half branch shows the channel-wise attention, which can leverage 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 some advanced loss functions to better capture spatial correlations between pixels and to generate high-quality density maps.

3.6 Comparisons

We compare the different networks discussed above in Table 3, and present their performance on four challenging crowd counting datasets in Table 4.

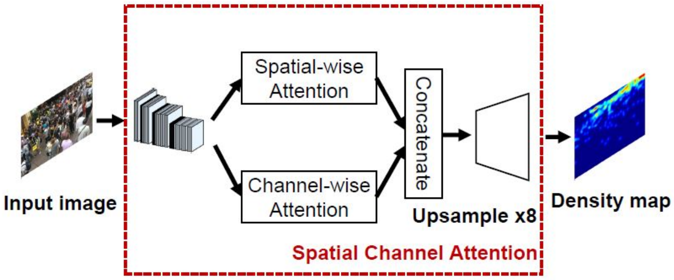

Fig. 10. Illustration of the spatial channel attention scheme for crowd counting.

As Tables 3 and 4 show, SANet achieves better counting performance on datasets with different crowd levels, compared with FCN. The generated density maps of FCN is only 1/4 × 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., 0.91M 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 (68.4M) is much larger than MCNN (0.13M). 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 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.

3.7 Others

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

- *Graph neural networks* based method distills rich relations among multi-scale features for crowd counting. As shown in Fig. 11, Hybrid Graph Neural Networks [45] 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.

Table 3. Comparisons of network design considerations for crowd counting.

Category		**Scheme**	**Advantages**	**Computational Complexity**	**Limitations**
Fully convolution neural networks		FCN [48]	Can analyze images of arbitrary size	Low	Low-resolution density maps
Encoder-Decoder Architecture		SANet [3]	Able to generate high-resolution density maps	Low (0.91M)	Not consider scale variation
Multi scale features	Multi-column architecture	MCNN [97]	Extract multi-scale features with multi-column architecture	Low (0.13M)	The scale diversity is restricted by the number of columns
	Pyramid architecture	CP-CNN [66]	Extract multi-scale features with pyramid architecture	High (68.4M)	High computational complexity
Advanced convolution operations	Dilated convolution operations	CSRNet [34]	Enlarge receptive field without increase the computational cost	Medium (16.26M)	Not consider the non-rigid objects
	Deformable convolution operations	ADCrowdNet [41]	Learnable additional offsets for better modeling non-rigid objects	High	High computational complexity
Attention-based Model		SCAR [16]	Capture local and global contextual information	Medium	Rely on pixel-wise loss function

Table 4. Analysis of different network design considerations.

Typical Schemes			**ST PartA**				**ST PartB**		**UCF_CC_50**		**UCSD**	
Methods	**Year**	**Column**	**MAE**	**MSE**	**PSNR**	**SSIM**	**MAE**	**MSE**	**MAE**	**MSE**	**MAE**	**MSE**
FCN [48]	2016	Single	126.5	173.5	N/A	N/A	23.8	33.1	338.6	424.5	N/A	N/A
MCNN [97]	2016	Multi	110.2	173.2	21.40	0.52	26.4	41.3	377.6	509.1	1.07	1.35
CP-CNN [66]	2017	Multi	73.6	106.4	21.72	0.72	20.1	30.1	295.8	320.9	N/A	N/A
SANet [3]	2018	Single	67.0	104.5	N/A	N/A	8.4	13.6	258.4	334.9	1.02	1.29
CSRNet [34]	2018	Single	68.2	115.0	23.79	0.76	10.6	16.0	266.1	397.5	1.16	1.47
ADCrowd [41]	2019	Single	63.2	98.9	24.48	0.88	8.2	15.7	266.4	358.0	1.10	1.42
SCAR [16]	2019	Double	66.3	114.1	23.93	0.81	9.5	15.2	259.0	374.0	N/A	N/A
HyGnn [45]	2020	Double	60.2	94.5	N/A	N/A	7.5	12.7	184.4	270.1	N/A	N/A

- *Recurrent neural networks* based Deep Recurrent Spatial-Aware Network (DSRNet) [40] utilize a learnable spatial transform module with a region-wise refinement process to adaptively enlarge the varied scales coverage. Researchers in [62] 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.

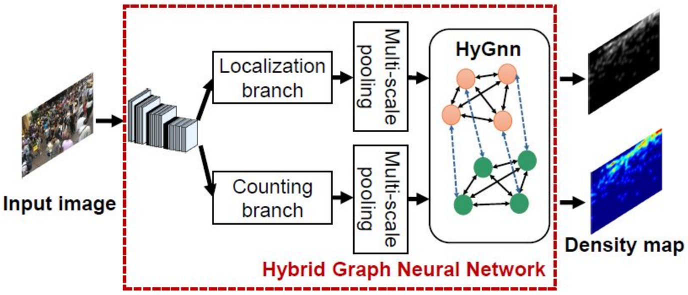

Fig. 11. The details of the Hybrid Graph Neural Networks for crowd counting.

- *Combining with detection* is another class of techniques for crowd counting. 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 [28], PSDDN [44], BSAD [21], DecideNet [37] and DRD [19]. DecideNet [37] 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 utilize 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), adversarial loss (Section 4.3), and multi-task learning (Section 4.4). We compare them in Section 4.5 and present some other emerging considerations in Section 4.6.

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:

$$L_E = \frac{1}{N} ||F(x_i;\theta) - y_i||_2^2, \tag{6}$$

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 [3], multi-scale SSIM loss [57], DMS-SSIM loss [39] and DMSSIM loss [30]. The SSIM index can be calculated point by point as Equation 2.2. Then the local pattern consistency can be formulated as:

$$L_s = 1 - \frac{1}{N} \sum_{x} SSIM(x). \tag{7}$$

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 [30] fused the difference of mean structural similarity index (DMSSIM) with the MSE loss to optimize the module. Besides, Multi-View Scale Aggregation Networks (MVSAN) [57] 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 Adversarial Loss

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 [53], MS-GAN [88], [98], ACSCP [46] and CODA [33]. The adversarial loss function can be defined as follow:

$$L_{adv} = -\log D(x_i, G(x_i, \theta)), \tag{8}$$

where x_i is the original input crowd image, $G(x_i)$ is the generated density map. $D(x_i, G(x_i, \theta))$ shows the probability that the generated density map is the real density map corresponding to the input crowd image. In the experiments, an adversarial loss can be used to improve the visual quality of the generated density maps, but usually degrades counting accuracy. For example, MS-GAN [88, 98] 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) [46] 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.

4.4 Multi-task Learning

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

$$L_c = \frac{1}{N} \sum_{i=1}^{N} || \frac{F_c(x_i; \theta) - y_i}{y_i + 1} ||, \tag{9}$$

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:

$$L_{total} = L_E + \alpha L_c, \tag{10}$$

where α is the weight to balance the pixel-wise Euclidean loss and the total head counting loss. SaCNN [92] 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 [63] fused segmentation map loss, density map loss and global density loss. Plug-and-Play Rescaling [60] combined regression loss with classification

Table 5. Comparisons of recent advanced loss functions for crowd counting.

Category	Advantages	Limitations
Euclidean loss	Flexible; widely used	Not consider context information and visual quality
SSIM loss	Variants of structural similarity loss to learn local correlation	Hard to learn the local correlation with various scales
Adversarial loss	Generate high-quality crowd density maps	Degrade counting accuracy
Multi-task learning	Fuse Euclidean loss with direct global count constrain	Sensitive to hyper-parameters
Others	Efficient divide and conquer manner	Computational expensive

Table 6. Comparisons of state-of-the-arts crowd counting approaches with different loss functions.

Scheme	Year	Multi scale	Dilated	Deform	Atten	Loss function	ST-A MAE	ST-A MSE	ST-B MAE	ST-B MSE
CSRNet [34]	2018		√			Euclidean loss	68.2	115.0	10.6	16.0
ADCrowd [41]	2019			√	√	Euclidean loss	63.2	98.9	8.2	15.7
DSSINet [39]	2019	√	√			SSIM Loss	60.63	96.04	6.8	10.3
ACSCP [46]	2019	√	√			Adversarial loss	85.2	137.1	15.4	23.1
S-DCNet [86]	2019	√				Divide-conquer	58.3	95.0	6.7	10.7
HA-CCN [68]	2020				√	MSE loss with global counting	62.9	94.9	8.1	13.4

loss. Shallow Feature-based Dense Attention Network [50] proposed to use MSE loss with counting loss and stated that counting loss not only accelerates the convergence but also improve the counting accuracy. Multi-supervised Parallel Network [81] 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 [24], [59] [23]. Multi-task learning based framework is widely used in recent papers [71], [36], [18], [27], [15], [96], [65]. However, this kind of framework is sensitive to hyper-parameters.

4.5 Comparisons

We summarize the advantages and limitations of the above loss functions in Table 5. We compare in Table 6 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 ignore 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 6 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 and ShanghaiTech B may probably 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.

4.6 Others

There are some other loss optimization strategies to enhance crowd counting tasks. CNN-Boosting [74] employed CNNs and incorporate two significant improvements: layered boosting and selective sampling. DAL-SVR [82] boosted deep attribute learning via support vector regression for fast-moving crowd counting. The paper learned superpixel 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 [64] used seep negative correlation learning, which is a successful ensemble learning technique for crowd counting. The authors extended D-ConvNet in [93], which proposed to regress via an efficient divide and conquer manner. D-ConvNet has 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 [86] 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 5.5.

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. Thus, 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. The fixed kernel or adaptive Gaussian kernel are widely used approaches to prepossess the original annotation and get the ground truth

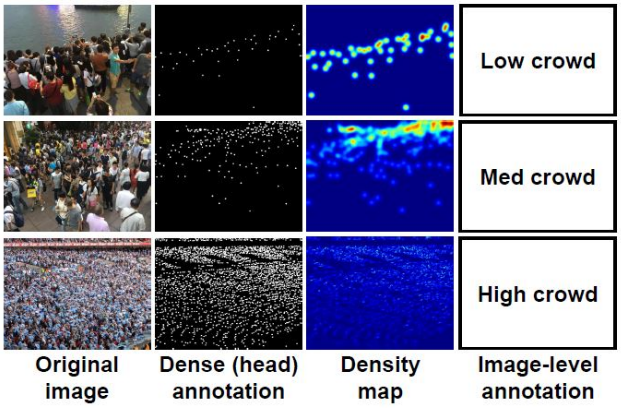

Fig. 12. The architecture of HA-CCN: a weakly supervised learning setup for crowd counting.

for density estimation and crowd counting [34]. The geometry-adaptive kernel is defined as follows:

$$F(x) = \sum_{i=1}^{N} \delta(x - x_i) \times G_{\sigma_i}(x), with\ \sigma_i = \beta \bar{d}_i, \tag{11}$$

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 [73] 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 [75].

BL [47] 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.

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. The original annotation process for crowd counting via density map estimation is point-level annotation, which is labor-intensive, HA-CCN [68] 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. 12, 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 [12] is also a kind of weakly supervised framework based on SRN structure.

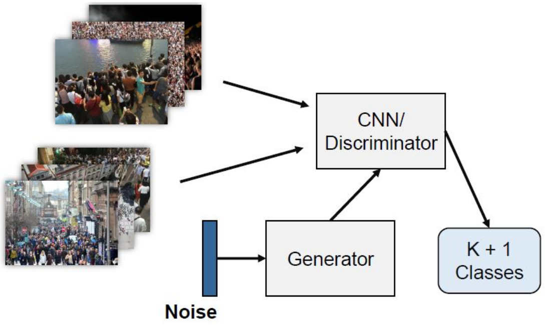

Fig. 13. The workflow of the original semi-supervised learning for classification problem.

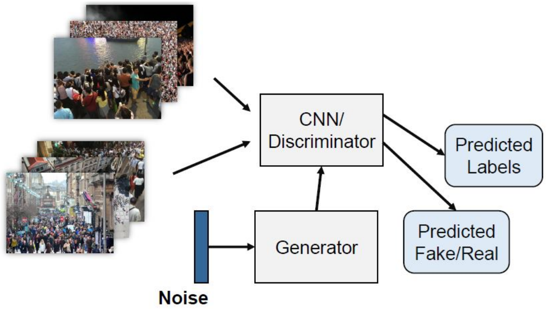

Fig. 14. The workflow of semi-supervised learning for density map estimation-based crowd counting.

Typical semi-supervised GANs are unable to function in the regression regime due to biases introduced when using a single prediction goal. DG-GAN [52] generalized semi-supervised generative adversarial network (GANs) from classification problems to regression for use in dense crowd counting, refer to Fig. 13 and Fig. 14. This work allows the dual-goal GAN to benefit from unlabeled data in the training process. And [54] 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

CNN-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. L2R [43] leveraged abundantly available unlabeled crowd images in learning to rank framework, refer to Fig. 15, 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 [61] utilized a Grid Winner-Take-All

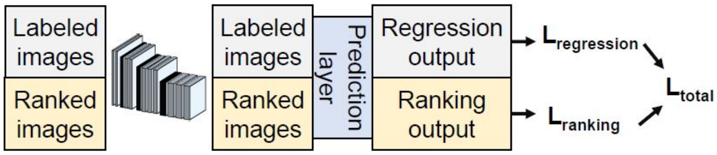

Fig. 15. The architecture of L2R: a self-supervised learning setup for crowd counting.

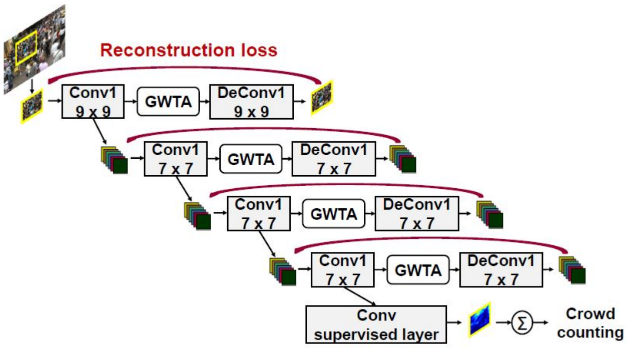

Fig. 16. The architecture of the almost unsupervised learning framework for crowd counting.

(GWTA) autoencoder to learn several layers of useful filters from unlabeled crowd images, refer to Fig. 16. 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 rest 0.1% are tuned with supervision. However, these kinds of self-supervised learning and almost unsupervised crowd counting approaches need a large amount of data to show effectiveness, which requires more training time and computational resources.

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. CCWld [78] 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 [18] based on 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, enhance accuracy, and improve robustness by making use of synthetic data. However, the synthetic data are still witnessed a larger domain gap compared with real data.

Table 7. Comparisons of different supervisory signals for crowd counting.

Category	Schemes	Advantages	Limitations
Fully Supervised learning	MCNN [97]	Adaptive Gaussian kernel to accommodate different scales	Not flexible to non-rigid object
	BL [47]	Bayesian loss to model non-rigid objects	More reliable supervision but suffers in large varied scales
Weakly supervised and semi-supervised learning	HA-CCN [68]	Low annotation complexity	Still requires weakly annotations and task specific knowledge
Unsupervised and self-supervised learning	L2R [43]	Low annotation cost; abundantly available	Large amount of data requires more training time
Automatic labeling through synthetic data	CCWld [78]	Reduce labeling effort; enahnce accuracy; improve robustness	Large domain gap from synthetic to real data

Table 8. Comparisons of state-of-the-arts crowd counting approaches with different supervisory signals.

Typical Schemes			ST PartA		ST PartB		UCF_CC_50		UCF-QNRF	
Methods	**Year**	**Columns**	**MAE**	**MSE**	**MAE**	**MSE**	**MAE**	**MSE**	**MAE**	**MSE**
MCNN [97]	2016	Multi	110.2	173.2	26.4	41.3	377.6	509.1	277	426
L2R [43] (Query by example)	2018	Double	72.0	106.6	14.4	23.8	291.5	397.6	N/A	N/A
L2R [43] (Query by keyword)	2018	Double	73.6	112.0	13.7	21.4	279.6	388.9	N/A	N/A
BL [47]	2019	Single	88.7	154.8	62.8	101.8	229.3	308.2	88.7	154.8
CCWld [78]	2019	Single	64.8	107.5	7.6	13.0	N/A	N/A	102.0	171.4
HA-CCN [68]	2020	Single	58.3	95.0	6.7	10.7	256.2	348.4	118.1	180.4

5.5 Comparisons

We summarize different supervisory signals for crowd counting with their representative schemes in Table 7. We compare them in Table 8.

BL achieves better performance on four different crowd counting datasets compared with MCNN. The good performance of BL may 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 counting accuracy than MCNN in Table 8 on different datasets with different backgrounds. This shows that the 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 8, 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 MCNN and BL. 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.

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 100 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 models 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. Currently, designing CNN-based crowd counting models still requires a manual network and feature selection with strong domain knowledge. Automated Machine Learning (AutoML) 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-arts 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 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 [85] is a recent attempt to leverage spatial-temporal information in video. There are some recent attempts to exploit the correlation in video data such as bidirectional ConvLSTM [85], E3D [102] and LST [13, 14]. 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 [94], which proposes a multi-camrea 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 [95]. 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.

REFERENCES

[1] Saeed Amirgholipour, Xiangjian He, Wenjing Jia, Dadong Wang, and Lei Liu. 2020. PDANet: Pyramid Density-aware Attention Net for Accurate Crowd Counting. arXiv:2001.0473

[2] Haoyue Bai, Song Wen, and S-H Gary Chan. 2019. Crowd counting on images with scale variation and isolated clusters. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*. IEEE, "https://dblp.uni-trier.de/db/conf/iccv/", 0–0.

[3] Xinkun Cao, Zhipeng Wang, Yanyun Zhao, and Fei Su. 2018. Scale aggregation network for accurate and efficient crowd counting. In *Proceedings of the European Conference on Computer Vision*. Springer, https://dblp.uni-trier.de/db/conf/eccv/, 734–750.

[4] Antoni B Chan, Zhang-Sheng John Liang, and Nuno Vasconcelos. 2008. Privacy preserving crowd monitoring: Counting people without people models or tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, IEEE, "https://dblp.uni-trier.de/db/conf/cvpr/", 1–7.

[5] Antoni B Chan and Nuno Vasconcelos. 2009. Bayesian poisson regression for crowd counting. In *Proceedings of the IEEE International Conference on Computer Vision*. IEEE, IEEE, "https://dblp.uni-trier.de/db/conf/iccv/", 545–551.

[6] Antoni B Chan and Nuno Vasconcelos. 2012. Counting people with low-level features and Bayesian regression. *IEEE Transactions on Image Processing* 21, 4 (2012), 2160–2177.

[7] Ke Chen, Chen Change Loy, Shaogang Gong, and Tony Xiang. 2012. Feature mining for localised crowd counting.. In *British Machine Vision Conference*, Vol. 1. British Machine Vision Association, https://dblp.uni-trier.de/db/conf/bmvc/, 3.

[8] Xinya Chen, Yanrui Bin, Nong Sang, and Changxin Gao. 2019. Scale pyramid network for crowd counting. In *IEEE Winter Conference on Applications of Computer Vision*. IEEE, IEEE, "https://dblp.org/db/conf/wacv/", 1941–1950.

[9] Zhi-Qi Cheng, Jun-Xiu Li, Qi Dai, Xiao Wu, Jun-Yan He, and Alexander G Hauptmann. 2019. Improving the learning of multi-column convolutional neural network for crowd counting. In *Proceedings of the ACM International Conference on Multimedia*. ACM, "https://dblp.org/db/conf/mm/", 1897–1906.

[10] Diptodip Deb and Jonathan Ventura. 2018. An aggregated multicolumn dilated convolution network for perspective-free counting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. IEEE, "https://dblp.uni-trier.de/db/conf/cvpr/", 195–204.

[11] Xinghao Ding, Zhirui Lin, Fujin He, Yu Wang, and Yue Huang. 2018. A deeply-recursive convolutional network for crowd counting. In *IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, IEEE, https://dblp.org/db/conf/icassp/, 1942–1946.

[12] Zihao Dong, Ruixun Zhang, Xiuli Shao, and Yumeng Li. 2020. Scale-Recursive Network with point supervision for crowd scene analysis. , 314–324 pages.

[13] Yanyan Fang, Shenghua Gao, Jing Li, Weixin Luo, Linfang He, and Bo Hu. 2020. Multi-level feature fusion based Locality-Constrained Spatial Transformer network for video crowd counting.

[14] Yanyan Fang, Biyun Zhan, Wandi Cai, Shenghua Gao, and Bo Hu. 2019. Locality-constrained spatial transformer network for video crowd counting. In *IEEE International Conference on Multimedia and Expo*. IEEE, IEEE, "https://dblp.org/db/conf/icmcs/", 814–819.

[15] Junyu Gao, Qi Wang, and Xuelong Li. 2019. Pcc net: Perspective crowd counting via spatial convolutional network.

[16] Junyu Gao, Qi Wang, and Yuan Yuan. 2019. SCAR: Spatial-/channel-wise attention regression networks for crowd counting. , 8 pages.

[17] Dan Guo, Kun Li, Zheng-Jun Zha, and Meng Wang. 2019. Dadnet: Dilated-attention-deformable convnet for crowd counting. In *Proceedings of the ACM International Conference on Multimedia*. ACM, "https://dblp.org/db/conf/mm/", 1823–1832.

[18] Tao Han, Junyu Gao, Yuan Yuan, and Qi Wang. 2020. Focus on semantic consistency for cross-domain crowd understanding. In *IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, IEEE, https://dblp.org/db/conf/icassp/, 1848–1852.

[19] Gaoqi He, Zhenwei Ma, Binhao Huang, Bin Sheng, and Yubo Yuan. 2019. Dynamic region division for adaptive learning pedestrian counting. In *IEEE International Conference on Multimedia and Expo*. IEEE, IEEE, "https://dblp.org/db/conf/icmcs/", 1120–1125.

[20] Mohammad Hossain, Mehrdad Hosseinzadeh, Omit Chanda, and Yang Wang. 2019. Crowd counting using scale-aware attention networks. In *IEEE Winter Conference on Applications of Computer Vision*. IEEE, IEEE, "https://dblp.org/db/conf/wacv/", 1280–1288.

[21] Siyu Huang, Xi Li, Zhongfei Zhang, Fei Wu, Shenghua Gao, Rongrong Ji, and Junwei Han. 2017. Body structure aware deep crowd counting. *IEEE Transactions on Image Processing* 27, 3 (2017), 1049–1059.

[22] Haroon Idrees, Imran Saleemi, Cody Seibert, and Mubarak Shah. 2013. Multi-source multi-scale counting in extremely dense crowd images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, "https://dblp.uni-trier.de/db/conf/cvpr/", 2547–2554.

[23] Haroon Idrees, Muhmmad Tayyab, Kishan Athrey, Dong Zhang, Somaya Al-Maadeed, Nasir Rajpoot, and Mubarak Shah. 2018. Composition loss for counting, density map estimation and localization in dense crowds. In *Proceedings of the European Conference on Computer Vision*. Springer, https://dblp.uni-trier.de/db/conf/eccv/, 532–546.

[24] Xiaolong Jiang, Zehao Xiao, Baochang Zhang, Xiantong Zhen, Xianbin Cao, David Doermann, and Ling Shao. 2019. Crowd counting and density estimation by trellis encoder-decoder networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, "https://dblp.uni-trier.de/db/conf/cvpr/", 6133–6142.

[25] Xiaoheng Jiang, Li Zhang, Pei Lv, Yibo Guo, Ruijie Zhu, Yafei Li, Yanwei Pang, Xi Li, Bing Zhou, and Mingliang Xu. 2019. Learning multi-level density maps for crowd counting.

[26] Di Kang and Antoni Chan. 2018. Crowd counting by adaptively fusing predictions from an image pyramid. arXiv:1805.06115

[27] Abhay Kumar, Nishant Jain, Suraj Tripathi, Chirag Singh, and Kamal Krishna. 2019. MTCNET: Multi-task Learning Paradigm for Crowd Count Estimation. arXiv:1908.08652

[28] Issam H Laradji, Negar Rostamzadeh, Pedro O Pinheiro, David Vazquez, and Mark Schmidt. 2018. Where are the blobs: Counting by localization with point supervision. In *Proceedings of the European Conference on Computer Vision*. Springer, https://dblp.uni-trier.de/db/conf/eccv/, 547–562.

[29] Victor Lempitsky and Andrew Zisserman. 2010. Learning to count objects in images. In *Advances in Neural Information Processing Systems*. MIT press, "https://dblp.uni-trier.de/db/conf/nips/", 1324–1332.

[30] Jing Li, Yaokai Xue, Weiqun Wang, and Gaoxiang Ouyang. 2019. Cross-Level Parallel Network for Crowd Counting. *IEEE Transactions on Industrial Informatics* 16, 1 (2019), 566–576.

[31] Min Li, Zhaoxiang Zhang, Kaiqi Huang, and Tieniu Tan. 2008. Estimating the number of people in crowded scenes by mid based foreground segmentation and head-shoulder detection. In *International Conference on Pattern Recognition*. IEEE, IEEE, "https://dblp.uni-trier.de/db/conf/icpr/", 1–4.

[32] Teng Li, Huan Chang, Meng Wang, Bingbing Ni, Richang Hong, and Shuicheng Yan. 2014. Crowded scene analysis: A survey. *IEEE transactions on circuits and systems for video technology* 25, 3 (2014), 367–386.

[33] Wang Li, Li Yongbo, and Xue Xiangyang. 2019. CODA: Counting Objects via Scale-Aware Adversarial Density Adaption. In *IEEE International Conference on Multimedia and Expo*. IEEE, IEEE, "https://dblp.org/db/conf/icmcs/", 193–198.

[34] Yuhong Li, Xiaofan Zhang, and Deming Chen. 2018. Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, "https://dblp.uni-trier.de/db/conf/cvpr/", 1091–1100.

[35] Zhe Lin and Larry S Davis. 2010. Shape-based human detection and segmentation via hierarchical part-template matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 32, 4 (2010), 604–618.

[36] Chenchen Liu, Xinyu Weng, and Yadong Mu. 2019. Recurrent attentive zooming for joint crowd counting and precise localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, IEEE, "https://dblp.uni-trier.de/db/conf/cvpr/", 1217–1226.

[37] Jiang Liu, Chenqiang Gao, Deyu Meng, and Alexander G Hauptmann. 2018. Decidenet: Counting varying density crowds through attention guided detection and density estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, "https://dblp.uni-trier.de/db/conf/cvpr/", 5197–5206.

[38] Lei Liu, Wenjing Jia, Jie Jiang, Saeed Amirgholipour, Yi Wang, Michelle Zeibots, and Xiangjian He. 2020. Denet: A universal network for counting crowd with varying densities and scales.

[39] Lingbo Liu, Zhilin Qiu, Guanbin Li, Shufan Liu, Wanli Ouyang, and Liang Lin. 2019. Crowd counting with deep structured scale integration network. In *Proceedings of the IEEE International Conference on Computer Vision*. IEEE, "https://dblp.uni-trier.de/db/conf/iccv/", 1774–1783.

[40] Lingbo Liu, Hongjun Wang, Guanbin Li, Wanli Ouyang, and Liang Lin. 2018. Crowd counting using deep recurrent spatial-aware network. arXiv:1807.00601

[41] Ning Liu, Yongchao Long, Changqing Zou, Qun Niu, Li Pan, and Hefeng Wu. 2019. Adcrowdnet: An attention-injective deformable convolutional network for crowd understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, "https://dblp.uni-trier.de/db/conf/cvpr/", 3225–3234.

[42] Weizhe Liu, Mathieu Salzmann, and Pascal Fua. 2019. Context-aware crowd counting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, "https://dblp.uni-trier.de/db/conf/cvpr/", 5099–5108.

[43] Xialei Liu, Joost Van De Weijer, and Andrew D Bagdanov. 2018. Leveraging unlabeled data for crowd counting by learning to rank. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, "https://dblp.uni-trier.de/db/conf/cvpr/", 7661–7669.

[44] Yuting Liu, Miaojing Shi, Qijun Zhao, and Xiaofang Wang. 2019. Point in, box out: Beyond counting persons in crowds. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, "https://dblp.uni-trier.de/db/conf/cvpr/", 6469–6478.

[45] Ao Luo, Fan Yang, Xin Li, Dong Nie, Zhicheng Jiao, Shangchen Zhou, and Hong Cheng. 2020. Hybrid Graph Neural Networks for Crowd Counting. arXiv:2002.00092

[46] Junjie Ma, Yaping Dai, and Yap-Peng Tan. 2019. Atrous convolutions spatial pyramid network for crowd counting and density estimation. , 91–101 pages.

[47] Zhiheng Ma, Xing Wei, Xiaopeng Hong, and Yihong Gong. 2019. Bayesian loss for crowd count estimation with point supervision. In *Proceedings of the IEEE International Conference on Computer Vision*. IEEE, "https://dblp.uni-trier.de/db/conf/iccv/", 6142–6151.

[48] Mark Marsden, Kevin McGuinness, Suzanne Little, and Noel E O'Connor. 2016. Fully convolutional crowd counting on highly congested scenes. arXiv:1612.00220

[49] Mark Marsden, Kevin McGuinness, Suzanne Little, and Noel E O'Connor. 2017. ResnetCrowd: A residual deep learning architecture for crowd counting, violent behaviour detection and crowd density level classification. In *IEEE International Conference on Advanced Video and Signal Based Surveillance*. IEEE, IEEE, "https://dblp1.uni-trier.de/db/conf/avss/", 1–7.

[50] Yunqi Miao, Zijia Lin, Guiguang Ding, and Jungong Han. 2020. Shallow Feature Based Dense Attention Network for Crowd Counting.. In *AAAI Conference on Artificial Intelligence*. AAAI, "https://dblp.uni-trier.de/db/conf/aaai/", 11765–11772.

[51] Min-hwan Oh, Peder A Olsen, and Karthikeyan Natesan Ramamurthy. 2020. Crowd Counting with Decomposed Uncertainty.. In *AAAI Conference on Artificial Intelligence*. AAAI, "https://dblp.uni-trier.de/db/conf/aaai/", 11799–11806.

[52] Greg Olmschenk, Jin Chen, Hao Tang, and Zhigang Zhu. 2019. Dense Crowd Counting Convolutional Neural Networks with Minimal Data using Semi-Supervised Dual-Goal Generative Adversarial Networks.

[53] Greg Olmschenk, Hao Tang, and Zhigang Zhu. 2018. Crowd counting with minimal data using generative adversarial networks for multiple target regression. In *IEEE Winter Conference on Applications of Computer Vision*. IEEE, IEEE, "https://dblp.org/db/conf/wacv/", 1151–1159.

[54] Greg Olmschenk, Zhigang Zhu, and Hao Tang. 2019. Generalizing semi-supervised generative adversarial networks to regression using feature contrasting. *Computer Vision and Image Understanding* 186 (2019), 1–12.

[55] Daniel Onoro-Rubio and Roberto J López-Sastre. 2016. Towards perspective-free object counting with deep learning. In *Proceedings of the European Conference on Computer Vision*. Springer, Springer, https://dblp.uni-trier.de/db/conf/eccv/, 615–629.

[56] Viet-Quoc Pham, Tatsuo Kozakaya, Osamu Yamaguchi, and Ryuzo Okada. 2015. Count forest: Co-voting uncertain number of targets using random forest for crowd density estimation. In *Proceedings of the IEEE International Conference on Computer Vision*. IEEE, "https://dblp.uni-trier.de/db/conf/iccv/", 3253–3261.

[57] Zhilin Qiu, Lingbo Liu, Guanbin Li, Qing Wang, Nong Xiao, and Liang Lin. 2019. Crowd counting via multi-view scale aggregation networks. In *IEEE International Conference on Multimedia and Expo*. IEEE, IEEE, "https://dblp.org/db/conf/icmcs/", 1498–1503.

[58] Vincent Rabaud and Serge Belongie. 2006. Counting crowded moving objects. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Vol. 1. IEEE, IEEE, "https://dblp.uni-trier.de/db/conf/cvpr/", 705–711.

[59] Viresh Ranjan, Hieu Le, and Minh Hoai. 2018. Iterative crowd counting. In *Proceedings of the European Conference on Computer Vision*. Springer, https://dblp.uni-trier.de/db/conf/eccv/, 270–285.

[60] Usman Sajid and Guanghui Wang. 2020. Plug-and-play rescaling based crowd counting in static images. In *IEEE Winter Conference on Applications of Computer Vision*. IEEE, "https://dblp.org/db/conf/wacv/", 2287–2296.

[61] Deepak Babu Sam, Neeraj N Sajjan, Himanshu Maurya, and R Venkatesh Babu. 2019. Almost unsupervised learning for dense crowd counting. In *AAAI Conference on Artificial Intelligence*, Vol. 33. AAAI, "https://dblp.uni-trier.de/db/conf/aaai/", 8868–8875.

[62] Chong Shang, Haizhou Ai, and Bo Bai. 2016. End-to-end crowd counting via joint learning local and global count. In *IEEE International Conference on Image Processing*. IEEE, IEEE, "https://dblp.org/db/conf/icip/", 1215–1219.

[63] Zenglin Shi, Pascal Mettes, and Cees GM Snoek. 2019. Counting with focus for free. In *Proceedings of the IEEE International Conference on Computer Vision*. IEEE, "https://dblp.uni-trier.de/db/conf/iccv/", 4200–4209.

[64] Zenglin Shi, Le Zhang, Yun Liu, Xiaofeng Cao, Yangdong Ye, Ming-Ming Cheng, and Guoyan Zheng. 2018. Crowd counting with deep negative correlation learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, "https://dblp.uni-trier.de/db/conf/cvpr/", 5382–5390.

[65] Vishwanath A Sindagi and Vishal M Patel. 2017. Cnn-based cascaded multi-task learning of high-level prior and density estimation for crowd counting. In *IEEE International Conference on Advanced Video and Signal Based Surveillance*. IEEE, IEEE, "https://dblp1.uni-trier.de/db/conf/avss/", 1–6.

[66] Vishwanath A Sindagi and Vishal M Patel. 2017. Generating high-quality crowd density maps using contextual pyramid cnns. In *Proceedings of the IEEE International Conference on Computer Vision*. IEEE, "https://dblp.uni-trier.de/db/conf/iccv/", 1861–1870.

[67] Vishwanath A Sindagi and Vishal M Patel. 2018. A survey of recent advances in cnn-based single image crowd counting and density estimation. *Pattern Recognition Letters* 107 (2018), 3–16.

[68] Vishwanath A Sindagi and Vishal M Patel. 2019. Ha-ccn: Hierarchical attention-based crowd counting network. *IEEE Transactions on Image Processing* 29 (2019), 323–335.

[69] Vishwanath A Sindagi and Vishal M Patel. 2019. Multi-level bottom-top and top-bottom feature fusion for crowd counting. In *Proceedings of the IEEE International Conference on Computer Vision*. IEEE, "https://dblp.uni-trier.de/db/conf/iccv/", 1002–1012.

[70] Vishwanath A Sindagi, Rajeev Yasarla, and Vishal M Patel. 2019. Pushing the frontiers of unconstrained crowd counting: New dataset and benchmark method. In *Proceedings of the IEEE International Conference on Computer Vision*. IEEE, "https://dblp.uni-trier.de/db/conf/iccv/", 1221–1231.

[71] Xin Tan, Chun Tao, Tongwei Ren, Jinhui Tang, and Gangshan Wu. 2019. Crowd Counting via Multi-layer Regression. In *Proceedings of the ACM International Conference on Multimedia*. ACM, "https://dblp.org/db/conf/mm/", 1907–1915.

[72] Thiago Teixeira, Gershon Dublon, and Andreas Savvides. 2010. A survey of human-sensing: Methods for detecting presence, count, location, track, and identity. *Comput. Surveys* 5, 1 (2010), 59–69.

[73] Matthias von Borstel, Melih Kandemir, Philip Schmidt, Madhavi K Rao, Kumar Rajamani, and Fred A Hamprecht. 2016. Gaussian process density counting from weak supervision. In *Proceedings of the European Conference on Computer Vision*. Springer, Springer, https://dblp.uni-trier.de/db/conf/eccv/, 365–380.

[74] Elad Walach and Lior Wolf. 2016. Learning to count with cnn boosting. In *Proceedings of the European Conference on Computer Vision*. Springer, Springer, https://dblp.uni-trier.de/db/conf/eccv/, 660–676.

[75] Jia Wan and Antoni Chan. 2019. Adaptive density map generation for crowd counting. In *Proceedings of the IEEE International Conference on Computer Vision*. IEEE, "https://dblp.uni-trier.de/db/conf/iccv/", 1130–1139.

[76] Chuan Wang, Hua Zhang, Liang Yang, Si Liu, and Xiaochun Cao. 2015. Deep people counting in extremely dense crowds. In *Proceedings of the ACM International Conference on Multimedia*. ACM, "https://dblp.org/db/conf/mm/", 1299–1302.

[77] Qi Wang, Junyu Gao, Wei Lin, and Xuelong Li. 2020. Nwpu-crowd: A large-scale benchmark for crowd counting. arXiv:2001.03360

[78] Qi Wang, Junyu Gao, Wei Lin, and Yuan Yuan. 2019. Learning from synthetic data for crowd counting in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, "https://dblp.uni-trier.de/db/conf/cvpr/", 8198–8207.

[79] Qi Wang, Junyu Gao, Wei Lin, and Yuan Yuan. 2020. Pixel-Wise Crowd Understanding via Synthetic Data. , 21 pages.

[80] Ze Wang, Zehao Xiao, Kai Xie, Qiang Qiu, Xiantong Zhen, and Xianbin Cao. 2018. In defense of single-column networks for crowd counting. arXiv:1808.06133

[81] Bo Wei, Yuan Yuan, and Qi Wang. 2020. MSPNET: Multi-Supervised Parallel Network for Crowd Counting. In *IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, IEEE, https://dblp.org/db/conf/icassp/, 2418–2422.

[82] Xinlei Wei, Junping Du, Meiyu Liang, and Lingfei Ye. 2019. Boosting deep attribute learning via support vector regression for fast moving crowd counting. *Pattern Recognition Letters* 119 (2019), 12–23.

[83] Xingjiao Wu, Yingbin Zheng, Hao Ye, Wenxin Hu, Tianlong Ma, Jing Yang, and Liang He. 2020. Counting crowds with varying densities via adaptive scenario discovery framework.

[84] Xingjiao Wu, Yingbin Zheng, Hao Ye, Wenxin Hu, Jing Yang, and Liang He. 2019. Adaptive scenario discovery for crowd counting. In *IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, IEEE, https://dblp.org/db/conf/icassp/, 2382–2386.

[85] Feng Xiong, Xingjian Shi, and Dit-Yan Yeung. 2017. Spatiotemporal modeling for crowd counting in videos. In *Proceedings of the IEEE International Conference on Computer Vision*. IEEE, "https://dblp.uni-trier.de/db/conf/iccv/", 5151–5159.

[86] Haipeng Xiong, Hao Lu, Chengxin Liu, Liang Liu, Zhiguo Cao, and Chunhua Shen. 2019. From open set to closed set: Counting objects by spatial divide-and-conquer. In *Proceedings of the IEEE International Conference on Computer Vision*. IEEE, "https://dblp.uni-trier.de/db/conf/iccv/", 8362–8371.

[87] Biao Yang, Weiqin Zhan, Nan Wang, Xiaofeng Liu, and Jidong Lv. 2020. Counting crowds using a scale-distribution-aware network and adaptive human-shaped kernel. , 207–216 pages.

[88] Jianxing Yang, Yuan Zhou, and Sun-Yuan Kung. 2018. Multi-scale generative adversarial networks for crowd counting. In *International Conference on Pattern Recognition*. IEEE, IEEE, "https://dblp.uni-trier.de/db/conf/icpr/", 3244–3249.

[89] Anran Zhang, Jiayi Shen, Zehao Xiao, Fan Zhu, Xiantong Zhen, Xianbin Cao, and Ling Shao. 2019. Relational attention network for crowd counting. In *Proceedings of the IEEE International Conference on Computer Vision*. IEEE, "https://dblp.uni-trier.de/db/conf/iccv/", 6788–6797.

[90] Anran Zhang, Lei Yue, Jiayi Shen, Fan Zhu, Xiantong Zhen, Xianbin Cao, and Ling Shao. 2019. Attentional neural fields for crowd counting. In *Proceedings of the IEEE International Conference on Computer Vision*. IEEE, "https://dblp.uni-trier.de/db/conf/iccv/", 5714–5723.

[91] Cong Zhang, Hongsheng Li, Xiaogang Wang, and Xiaokang Yang. 2015. Cross-scene crowd counting via deep convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, "https://dblp.uni-trier.de/db/conf/cvpr/", 833–841.

[92] Lu Zhang, Miaojing Shi, and Qiaobo Chen. 2018. Crowd counting via scale-adaptive convolutional neural network. In *IEEE Winter Conference on Applications of Computer Vision*. IEEE, IEEE, "https://dblp.org/db/conf/wacv/", 1113–1121.

[93] Le Zhang, Zenglin Shi, Ming-Ming Cheng, Yun Liu, Jia-Wang Bian, Joey Tianyi Zhou, Guoyan Zheng, and Zeng Zeng. 2019. Nonlinear regression via deep negative correlation learning.

[94] Qi Zhang and Antoni B Chan. 2019. Wide-area crowd counting via ground-plane density maps and multi-view fusion cnns. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, "https://dblp.uni-trier.de/db/conf/cvpr/", 8297–8306.

[95] Qi Zhang and Antoni B Chan. 2020. 3D Crowd Counting via Multi-View Fusion with 3D Gaussian Kernels.. In *AAAI Conference on Artificial Intelligence*. AAAI, "https://dblp.uni-trier.de/db/conf/aaai/", 12837–12844.

[96] Youmei Zhang, Chunluan Zhou, Faliang Chang, and Alex C Kot. 2019. Multi-resolution attention convolutional neural network for crowd counting. , 144–152 pages.

[97] Yingying Zhang, Desen Zhou, Siqin Chen, Shenghua Gao, and Yi Ma. 2016. Single-image crowd counting via multi-column convolutional neural network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, "https://dblp.uni-trier.de/db/conf/cvpr/", 589–597.

[98] Yuan Zhou, Jianxing Yang, Hongru Li, Tao Cao, and Sun-Yuan Kung. 2020. Adversarial learning for multiscale crowd counting under complex scenes.

[99] M Sami Zitouni, Harish Bhaskar, J Dias, and Mohammed E Al-Mualla. 2016. Advances and trends in visual crowd analysis: A systematic survey and evaluation of crowd modelling techniques. *Neurocomputing* 186 (2016), 139–159.

[100] Zhikang Zou, Yu Cheng, Xiaoye Qu, Shouling Ji, Xiaoxiao Guo, and Pan Zhou. 2019. Attend to count: Crowd counting with adaptive capacity multi-scale CNNs. , 75–83 pages.

[101] Zhikang Zou, Yifan Liu, Shuangjie Xu, Wei Wei, Shiping Wen, and Pan Zhou. 2020. Crowd Counting via Hierarchical Scale Recalibration Network. arXiv:2003.03545

[102] Zhikang Zou, Huiliang Shao, Xiaoye Qu, Wei Wei, and Pan Zhou. 2019. Enhanced 3D convolutional networks for crowd counting. arXiv:1908.04121

[103] Zhikang Zou, Xinxing Su, Xiaoye Qu, and Pan Zhou. 2018. Da-net: Learning the fine-grained density distribution with deformation aggregation network. *IEEE Access* 6 (2018), 60745–60756.