A Survey of Crowd Counting CNN-based

hui gao

Henan University of Technology

miaolei deng

Henan University of Technology

wenjun zhao

Henan University of Technology

dexian zhang (✉ zdx@haut.edu.cn)

Henan University of Technology

Research Article

Keywords: Deep Learning, Crowd Counting, Convolutional Neural Network, Density Estimation

Posted Date: October 15th, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-957081/v1>

License: This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

A Survey of Crowd Counting CNN-based

Hui Gao^{1,2}, Miaolei Deng^{1,3}, Wenjun Zhao^{1,3}, *Dexian Zhang^{1,3}

¹ Key Laboratory of Intelligent Information Processing, Ministry of Education, Henan University of Technology, Zhengzhou 450001, China

² School of Mechanical and Electrical Engineering, Henan University of Technology, Zhengzhou 450001, China

³ School of Information Science and Engineering, Henan University of Technology, Zhengzhou 450001, China

Abstract—Benefiting from the powerful feature representation ability of deep learning, Convolutional Neural Network (CNN) provides a better solution to estimate accurately the number of people in a crowded scene, but it still faces many problems that need to be solved urgently. It is one of the key and difficult points in the field to reduce the complexity of the network and to improve the real-time performance of the network, so as to improve the accuracy of crowd counting. Firstly, this paper introduces the research background and application of crowd counting. Secondly, it focuses on the commonly used counting model, loss function, and dataset and evaluation method. Then compare the performance structure, advantages and disadvantages of different algorithms horizontally on several published datasets. Finally, it summarizes the shortcomings of the existing crowd counting, put forward to the future research direction of crowd counting.

Index Terms—Deep Learning; Crowd Counting; Convolutional Neural Network; Density Estimation

I. INTRODUCTION

Over the last few years, target counting technologies such as crowd counting^[1], vehicle counting^[2] and cell counting^[3] have been widely used in the field of computer vision, and accurate counting of people has always been a hot and difficult research topic in target counting. For example, analyze the

crowd flow in shopping malls. Marketing activities can be better carried out according to consumers' preferences; Grasping the crowd distribution in places where people gather, such as stadiums and concerts, can not only evacuate people in time to prevent stampede accidents, but also help to crack down on illegal processions and assemblies. In a scene where people gather rapidly. Crowd counting plays a vital role in social security and control management.

The early crowd counting method is based on detection^[4-7]. This kind of method uses a sliding window to detect the crowd in the scene, and counts the number of people according to the result. For example, the detectors such as R-CNN^{[8]-[10]}, YOLO^[11] and SSD^[12], which appeared in recent years, have amazing detection accuracy in sparse scenes, but the detection effect is not satisfactory in the face of high density crowd and occlusion. In order to solve the above problems, regression-based methods^{[13][14]} first extract the low-level features of the scene, and then learn a regression model to give an estimate of the approximate number of people, but these methods ignore the spatial information. Therefore, Lemptisky et al.^[15] first adopted the method based on density estimation by learning the linear mapping between local features and corresponding density maps. Random forest^[16] trains two different forests by introducing congestion prior, and obtains satisfactory performance. Although these methods consider spatial information, they only use manual extraction of low-level

information, and can not obtain high-quality density maps to estimate more accurate counts.

In summary, the contributions of this paper are mainly in the following folds:

1) From different aspects such as network structure, loss function and performance evaluation index, this paper systematically and comprehensively summarizes and deeply discusses the research status in the field of crowd counting. This comprehensive combing can help researchers quickly understand the research status and key technologies of crowd counting algorithm based on deep learning.

2) Based on the dataset, the counting effects of different models are compared, and the reasons for the advantages and disadvantages of the counting models are analyzed, which provides reference for future researchers to design more optimized counting models.

3) Summarize and analyze the problems in model design and loss function definition, and point out the direction for future research in this field.

The remainder of this paper is organized as follows. Section II introduces systematically the structure and loss function of these three types of networks. Section III introduces several loss function. Section IV analyzes public datasets, evaluation criteria, and compares several models. Section V discusses the challenges and problems to be solved in the current crowd counting methods, and section VI puts forward to the possible development direction in the future.

II. CROWD COUNTING MODELS

Compared with the traditional manual feature extraction, the CNN has a strong feature extraction capability, which is more effective and efficient in dealing with the problems of scene adaptation and scale diversity. Since 2016, CNN has become the mainstream network framework for density estimation and crowd counting. Fig.1 shows a crowd counting method, include traditional approaches and modern approaches of crowd counting.

Fig.1 Crowd Counting Approaches

Crowd counting methods CNN-based are mainly divided into regression method and density map method. Regression method can directly output the number of people by sending pictures of people to CNN, which is suitable for scenes with sparse crowds. In the

density map method, CNN outputs a crowd density map, and then calculates the number of people by mathematical integration and summation. The performance of this method depends on the quality of density map to some extent. In order to generate a high-quality

density map, a new loss function^[17] will be introduced to improve the clarity and accuracy of the density map. Different models have different levels of supervision and learning paradigms, and some models are designed across scenes and fields. Table I shows the milestone of crowd counting, including traditional approaches, density estimation and crowd counting methods CNN-based.

No matter which method is adopted, feature extraction is needed first. In order to improve the robustness of features, multi-scale

prediction, context awareness, hole convolution, deformable convolution and other methods are often used to improve the feature extraction process to enhance the ability of distinguishing features. According to the network structure, Crowd counting models can be divided into three categories: multi-branch structure, single-branch structure and special structure. This section will make a comprehensive analysis of CNN-based crowd counting from the network structure.

TABLE I A TABLE OF MILESTONE ON CROWD COUNTING

Class	Method	Model	Year
Traditional Approaches	Detection-based	MLR	2006
		KRR	2007
	Regression-based	Chan et al.	2008
		Lempitsky et al.	2010
	Density estimation based	RR	2012
		CA-RR	2013
		Count Forest	2015
		MCNN	2016
		Hvdr-CNN	2016
		MSCNN ^[62]	2017
		CP-CNN	2017
		CMTL	2017
		Switch-CNN	2017
		DecideNet	2018
		CSRNet	2018
		SA-Net	2018
		ACSCP	2018
		PSSDN	2019
		LSF-CNN	2019
		PGCNet ^[63]	2019
Modern Approaches	CNN-based	DUBNet ^[64]	2020
		ADSCNet ^[65]	2020

2.1 Multi-branch Structure Network

To solve the problem of target scale change, Boominathan et al.^[18] proposed a CrowdNet with double-branch structure CNN-based, as shown in Fig.2. A shallow network and a deep network are used to extract feature information of different scales and fuse them to predict the crowd density map. This combination can capture both high-level and

low-level semantic information, So as to adapt to the uneven scaling of the crowd and the change of the visual angle, thus being beneficial to the crowd counting in different scenes and different scales. Multi-scale problems can be effectively solved by introducing multi-channel networks and extracting different scale features with different receptive fields. A series of crowd counting algorithms based on multi-column convolutional neural networks are derived.

Fig.2 The Architecture of Two-column Crowd Counting Network^[18]

Zhang et al.^[19], inspired by multi-branch deep convolutional neural network^[20], proposed a multi-column convolutional neural network (MCNN) for crowd counting, and its structure is shown in Fig.3. Each branch network uses convolution kernels of different sizes to extract the feature information of targets of different scales, thus reducing the counting error caused by different sizes of targets due to the change of viewing angle. MCNN establishes the nonlinear relationship between the image and the crowd density map, and makes the model deal with any size of input images by replacing the full connection layer with the full convolution layer. In order

to further correct the influence caused by the change of viewing angle, MCNN does not use fixed Gaussian kernel when generating density map, but uses adaptive Gaussian kernel to calculate density map, which improves the quality of density map. Another contribution of this work is to collect and label ShanghaiTech crowd counting dataset, which is composed of 1198 labeled images, including various scenes where the crowd distribution changes from sparse to dense. At present, this dataset has become one of the benchmark datasets in the field of crowd counting.

Fig.3 The Architecture of The Multi-column Crowd Counting Network^[19]

As mentioned earlier, the counting performance mainly depends on the quality of density map. In order to generate a density map of higher quality, Sindagi et al.^[21] proposed a context pyramid convolution neural network counting model CP-CNN, whose structure is shown in Fig.4. Scene context information of different scales is obtained through multiple CNN, and these context information are explicitly embedded

into the density map generation network. Improve the accuracy of density estimation. CP-CNN is composed of four parts, among which global context estimator(GCE) and local context estimator(LCE) respectively extract the global and local context information of the image, that is, predict the density level of the image from the global and local perspectives.

Fig.4 The Architecture of CP-CNN^[21]

Density map estimator(DME) does not generate density map directly, but uses MCNN to generate high-dimensional feature map. Fusion Convolutional Neural Network (F-CNN) fuses the outputs of the first three parts to generate density maps. In order to make up for the missing details in DME, F-CNN used a series of fractional steps to help reconstruct the details of density map. The existing CNN counting network mainly uses the pixel-level Euclidean distance loss function to train the network, which leads to the fuzzy density map. Therefore, CP-CNN introduces adversarial loss, and overcomes the deficiency of Euclidean distance loss function by using Generative adversary net (GAN) ^[22].

In 2017, Sam et al.^[23] proposed a multi-column selective convolution neural

network (Switch-CNN) for crowd counting, and its structure is shown in Fig.5. Different from MCNN, Switch-CNN adopts multi-column network structure, but each column network handles different areas independently. Before being sent to the network, the image is divided into 3*3 areas. Then, each area is divided into density grades by using a specific Switch module, and corresponding branches are selected according to the density grades to count. By using regression networks with different scales to estimate the density of people with different densities, the final counting results are more accurate. Switch-CNN also has drawbacks that can't be ignored. If the branch selection is wrong, the counting accuracy will be greatly affected.

Fig.5 The Architecture of Switch-CNN^[23].

Switch-CNN selects the appropriate branch network to estimate the crowd density according to the content of the image block, which provides a new idea for designing the multi-column counting network. However, Switch-CNN divides the density level into three levels, which makes it difficult to deal with the scene with a wide range of crowd density changes. To this end, Sam et al.^[24] improved Switch-CNN and put forward incremental growing CNN (IG-CNN). The hierarchical training process is shown in Fig. 6. Starting from a Base CNN model, through continuous iteration, a CNN binary tree is finally generated, and the leaf nodes are the regressions used for density estimation,

wherein each regressive device corresponds to a specific density grade. The first layer divides the training set D_0 into D_{00} and D_{01} by clustering, and then R_{00} and R_{01} are copied from R_0 . Then R_{00} and R_{01} are trained on the corresponding training sets D_{00} and D_{01} , respectively, and the construction of other layers is similar. Finally, through hierarchical clustering, the original training set is divided into several subsets, each subset corresponds to a density level, and the corresponding density estimator is responsible for counting. In the test stage, the corresponding density estimator is selected according to the density level of the picture.

Fig. 6. Hierarchical Differential Training in IG-CNN^[24].

In the existing crowd counting models, it is usually simply assumed that the crowd distribution in the scene is sparse or dense. For sparse scenes, the detection method is used to count^[25]; For dense scenes, the regression method is used to estimate the crowd density. Such models are often difficult to deal with the counting of crowd scenes with a wide range of density changes. To solve this

problem, Liu et al.^[26] put forward a crowd counting model, DecideNet, which combines detection and regression, and its structure is shown in Fig.7. The model is also a multi-column counting network, in which RegNet module directly estimates the crowd density from the image by regression method, and DetNet module adds a Gaussian convolution layer behind Faster-RCNN. The

detection result is directly converted into a crowd density map, and then QualityNet is introduced into the attention module to automatically judge the crowd density, and the weights of the detection and regression methods are adaptively adjusted according to the judgment result, and then the two density

maps are fused according to this weight, so as to obtain a better optimal solution. However, because RegNet and DetNet both use large receptive fields and have too many model parameters, the training complexity of this model is high.

Fig.7 The Architecture of DecideNet^[26]

Multi-column counting network uses convolution kernels of different sizes to extract multi-scale features of images, and its good effect shows the importance of multi-scale representation. However, the multi-column counting network also introduces new problems. Firstly, the performance of multi-scale expression usually depends on the number of branches of the network, that is, the diversity of scales is limited by the number of branches. Secondly, most of the existing works use Euclidean distance as the loss function, assuming that the pixels are independent of each other, which leads to the fuzzy density map.

To solve the above problems, Cao et al.^[27] proposed a scale aggregation network(SANet), whose structure is shown in Fig. 8. In this

model, MCNN is not adopted, but the architecture idea of Inception^[28] is used for reference. Different convolution kernels of different sizes are used at each convolution layer to extract features of different scales, and finally a high-resolution density map is generated by deconvolution. The whole model consists of FME (feature map encoder) and DME (density map estimator). FME aggregates to extract multi-scale features, and DME fuses features to generate high-resolution density maps. When measuring the similarity between predicted density map and ground-truth, SSIM is used to calculate the local consistency loss, and then Euclidean loss and local consistency loss are weighted to get the total loss.

Fig.8 The Architecture of SANet^[27]

Due to the large difference of target size in different depth of field, the modeling ability of crowd counting model is required. To solve this problem, Mohammad et al.^[29] introduced attention mechanism into the field of crowd counting for the first time, and proposed a multi-branch scale-aware attention network (SAAN), whose structure is shown in Fig. 9. The network consists of four modules, the multi-scale feature extractor (MFE) is responsible for extracting the multi-scale feature map from the input image. Inspired by MCNN, MFE is designed as a multi-column network with three branches. Each branch has different receptive fields and can capture features of different scales. In order to obtain the global density information of the image, three global density levels are defined corresponding to three branches with different

scales in MFE. Then, the global scale attentions (GSA) module is used to extract the global context information of the input image, calculate the scores corresponding to three global density levels, and normalize the three scores. GSA can only extract the global scale information of the image, but in the actual crowd counting image, there are density differences in different positions. Therefore, local scale attention (LSA) is added to extract fine-grained local context information at different positions of the image and generate three pixel-level attention maps to describe local scale information. Finally, the feature map extracted by MFE is weighted according to the global and local scale information, and then the weighted feature map is input into a fusion network (FN) generate the final density map.

Fig.9 The Architecture of SAAN.^[29]

The input image is passed simultaneously through three sub-networks: multi-scale feature extractor (MFE), global scale attentions (GSA), and local scale attentions (LSA). MFE extracts feature maps in three different scales. GSA and LSA produce three global scores and three pixel-wise local attention maps, respectively. Multi-scale feature maps are then weighted by the corresponding GSA and LSA outputs. The attention-weighted features are then used as the input to the fusion network to predict the density map. Finally, the crowd count can be obtained by summing over the entries in the predicted density map.

2.2 Single Branch Structure Network

With the deepening of the network, the multi-branch structure consumes a lot of training time, resulting in redundancy of the network structure. Therefore, Wang et al.^[30] first introduced CNN into the field of crowd counting, and proposed an end-to-end CNN regression model suitable for dense crowd scenes. The model improves AlexNet^[31], and replaces the last full connection layer with a single neuron layer to directly predict the number of people. Because there is no predicted crowd density map, it is impossible to count the distribution of people in the scene.

In addition, although the model automatically learns effective counting features through CNN, because AlexNet is narrow in width and shallow in depth, the robustness of features is not strong enough, the counting effect in crowded scenes is poor, and the effect is not ideal when counting across scenes. Lack of sufficient generalization.

In order to solve the cross-scene problem, Zhang et al.^[32] proposed a cross-scene counting model Crowd CNN based on AlexNet, which tried to output the crowd density map for the first time, and its overall structure is shown in Fig.10. Among them, Fig.10 (a) depicts the pre-trained process of counting network, and optimizes the model through the alternate training of crowd density map and crowd counts. Then, According to the characteristics of the target scene, the algorithm selects similar scenes to fine-tuning the counting model, as shown in Fig.10 (b), so as to achieve the purpose of cross-scene counting. In order to improve the counting accuracy, the author also puts forward the concept of perspective map, but it is difficult to obtain perspective map, which limits the popularization of this model. Another contribution of this work is the establishment of the classic crowd counting dataset WorldExpo'10, which provides data for the evaluation of the cross-scene crowd counting model.

Fig 10. The Cross-scene Crowd Counting Model^[32]

2.3 Special Structure Network

Although both multi-branch and single-branch networks have achieved good counting results, the network redundancy and poor scene adaptability greatly reduce the robustness of the counting network. In order to overcome these problems, some new CNN structures have been used for crowd counting, such as Dilated Convolutional Networks^[33], Deformable Convolutional Networks^[34], GAN^[35] and so on. It not only reduces the complexity of the model, but also improves the accuracy of crowd counting.

In 2018, Li et al.^[36] proposed a hollow convolution neural network model CSRNet suitable for dense crowd counting, and its network structure is shown in Fig.11. Instead of adopting the multi-branch network structure

widely used in the past, CSRNet takes VGG16, which abandons the full connection layer, as the front part of the network, while the back end adopts 6-layer hollow convolution neural network to form a single-channel counting network. The network parameters are greatly reduced and the training difficulty is reduced. At the same time, with the help of hole convolution, the advantages of receptive field can be enlarged while maintaining the resolution, and more image details can be retained, which makes the generated crowd distribution density map of higher quality. There are four different configurations of CSRNet backend, a to d, as shown in Table II, among which group b scheme performs best on ShanghaiTech PartA dataset.

TABLE II CONFIGURATION OF CSRNET.

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

The success of CSRNet provides a new way of thinking for dense crowd counting, and then many scholars began to emulate the study of crowd counting by hollow convolution^[37].

Different branches of multi-branch counting network lack mutual cooperation, and each branch only tries to optimize its own

estimation by minimizing Euclidean loss. Because each branch only performs well on a specific scale, the density map generated after averaging the results of each branch is vague, and because the pool layer is used in the network, the resolution of the density map is greatly reduced. Make the final counting result

produce errors. In addition, there is a problem of inconsistent statistics across scales. There is a difference between the total number of people obtained by dividing an image into multiple copies and inputting them into the network, and the number of people calculated by inputting the whole image.

To solve these problems, inspired by the successful application of GAN in image translation [38], literature[39] proposed an advanced cross-scale consistency pursuit network (ACSCP) based on GAN, and its structure is shown in Fig.11. The introduction of countermeasure loss makes the generated density map sharper, and the generator of

U-net architecture[40] ensures the high resolution of the density map. At the same time, the cross-scale consistency regularizer constrains the cross-scale error between images. Therefore, the model can finally generate a high-quality and high-resolution crowd distribution density map, thus obtaining higher crowd counting accuracy. The method of using GAN to improve the accuracy of crowd counting opens a new way of thinking. In the SFCN[41] counting network, Improved Cycle GAN[42] is used to generate pictures with similar dataset styles and contribute to GCC dataset. In DACC[43], Cycle GAN is also used for style transfer.

Fig.11 The Architecture of Adversarial Cross-Scale Consistency Pursuit Networks (ACSCP)[39].

Although the crowd counting solution based on deep neural network has achieved remarkable results, the counting effect will still be seriously affected by background noise, occlusion and inconsistent crowd distribution in highly crowded and noisy scenes. To solve this problem, Liu et al. proposed a deformable convolution network ADCrowdNet[44], which integrated attention mechanism, for crowd counting. As shown in Fig.12, the network model is mainly composed of two parts connected in series, in which the attention map generator(AMG) is used to detect candidate crowd areas and estimate the congestion degree of these areas, providing

refined prior knowledge for the subsequent generation of crowd density map. Irrelevant information such as complex background can be filtered out through attention mechanism, Make the follow-up work only focus on the crowd area, and reduce the interference of various noises. Density map estimator (DME) is a multi-scale deformable convolution network for generating high-quality density maps. Because attention is injected, the direction parameter is added to the deformable convolution, and the convolution kernel extends on the feature graph under the guidance of attention. It can model different shapes of crowd distribution, which is well

adapted to the distortion caused by camera angle distortion and diversity of crowd distribution in real scenes, and ensures the

accuracy of crowd density map in crowded scenes.

Fig.12 The Architecture of ADCrowdNet^[44].

Note that the network structure of the map generator AMG uses the first 10 convolution layers of the VGG16 network as the front end to extract the bottom features of the image, and the back end structure uses multiple hole convolution layers^[45] with different hole rates to expand the receptive field and deal with multi-scale problems. The back-end outputs two-channel feature maps, representing foreground (crowd) and background respectively. Then, the corresponding weights are obtained by pooling the global average of the feature graph, and then the results are classified by softmax to obtain the probability. Finally, the attention map is obtained by point multiplication of the feature map and probability.

The network structure of density map estimator DME, the front end still uses VGG16, and the back end architecture is still similar to the inception structure, but multi-scale deformable convolution which is more suitable for crowded and noisy scenes is adopted to adapt to the geometric deformation of crowd distribution.

In the same year, DADNet^[46] also used deformable convolution to count people, and achieved good counting results.

Background noise will greatly affect the performance of crowd counting algorithm. In order to reduce the interference of background noise, many scholars have tried. For example, ADCrowdNet filters out the background through attention mechanism, so that the model only pays attention to the crowd area.

In addition, some scholars try to apply the image segmentation technology MASK R-CNN^[47] to the field of crowd counting to remove background noise.

The difficulty of background and crowd segmentation lies in how to make ground truth for segmentation. Therefore, researchers have made various attempts. SFANet^[48] uses the original coordinate point ground truth to carry out Gaussian blur with a fixed Gaussian kernel size, and then selects a certain threshold to binarize it with 0 and 1, thus forming a segmented ground truth; MAN^[49] uses the fixed Gaussian kernel to process the original coordinate point ground truth, and sets all non-zero values to 1 to form a segmented ground truth; W-Net^[50] uses normalized Gaussian kernel method to blur the coordinate point map, and then sets a certain threshold to classify it twice; According to SGANet^[51], each head is represented by 25×25 squares, thus making ground truth.

In a word, how to reduce the interference of background noise is still a key issue in the field of crowd counting in the future. In addition to the above crowd counting algorithm combined with segmentation algorithm, CFF^[52] combines segmentation task, classification task and counting task, which provides us with the idea of multi-task combination.

From the above analysis, it can be seen that the structure of the counting model is constantly changing with the deepening of research. In order to solve the multi-scale

problem, the counting network evolved from a simple single-branch structure to a complex multi-branch structure, which improved the counting accuracy. However, the multi-branch structure will bring a lot of network parameters and high computational complexity, the efficiency of counting model is low. In order to overcome these problems, the researchers tried to regress to the simple single-branch network structure, and introduced various new CNN technologies to reduce the complexity of the model and improve the counting accuracy. Therefore, reducing the number of branches and making the counting model simple and effective will be the design direction of the future model network structure.

In addition, it can be seen from the analysis that CNN technologies such as attention mechanism, hollow convolution, confrontation generation network and deformable convolution can solve the problems of multi-scale and complex background interference in the counting field and help improve the quality of density map. Therefore, when designing the network in the future, we can consider combining these technologies to improve the counting accuracy.

III. LOSS FUNCTION

Loss function is an indispensable part of model training, which is used to evaluate the consistency between the predicted value and the true value. The smaller the value, the closer the predicted value is to the true value, the better the performance of the model. In crowd counting, commonly used loss functions include Euclidean distance, structural similarity and other losses. The network parameter value with the smallest loss function value is found by training neural network.

3.1 Euclidean Distance Loss

In the early days, most people counting work based on density map, such as MCNN^[20], CrowdNet^[18], Switch-CNN^[23], CSRNet^[36], etc., used pixel-level Euclidean distance as model loss function to measure the difference between estimated density map and real density map, as shown in formula (1):

$$L(i) = \frac{1}{2N} \sum_{i=1}^N \|D_i^{\text{EST}} - D_i^{\text{GT}}\|_2^2 \quad (1)$$

Where n is the number of images trained in a batch; D_i^{EST} is the estimated density value of the i th training sample, and the parameter is θ ; D_i^{GT} is the true density value.

Because Euclidean distance loss is simple, training speed is fast, and counting effect is good, it has been widely used in the early stage. However, the robustness of Euclidean distance loss is poor, and it is easy to affect the overall counting effect because of the extreme situation of individual pixels. In addition, Euclidean distance loss is the average of all pixels, and does not pay attention to the structured information of pictures. For the same picture, it is easy to have the problem that the predicted value of densely populated areas is too small and the predicted value of sparsely populated areas is too large, but the final average result does not reflect these problems, which leads to the fuzzy density map and unclear details.

3.2 Structural Similarity Loss

Because Euclidean distance loss is not enough to express the visual perception of pictures by human visual system, the generated density map is of low quality. In order to overcome the deficiency of Euclidean distance loss, SANet^[27] proposed structural similarity loss based on structural similarity index to measure the quality of density map. Structural similarity index is an image quality

evaluation standard proposed by Wang et al.^[66], which is named SSIM. Different from pixel-based error evaluation criteria, SSIM measures image similarity from three aspects: brightness, contrast and structure, and calculates the similarity between two images through three local statistics: mean, variance and covariance. The value range of SSIM is -1 to 1, The larger SSIM value is, the higher similarity is. The calculation method of structural similarity index SSIM is shown in formula (2):

$$SSIM = \frac{(2\mu_e\mu_g + c_1)(2\sigma_{eg} + c_2)}{(\mu_e^2 + \mu_g^2 + c_1)(\mu_e^2 + \mu_g^2 + c_2)} \quad (2)$$

In which e and g represent the generated density map and the real density map respectively, μ_e, σ_e^2 Represents the mean and variance of the true density map, μ_g, σ_g^2 Represents the covariance between the real density map and the generated density map. In order to avoid abnormality when denominator is 0, smoothing coefficients c1, c2 are set to very small constant values. L_{SSIM} is the structural similarity loss, and the calculation method is shown in formula (3):

$$L_{SSIM} = 1 - \frac{1}{N} \sum_x SSIM(x) \quad (3)$$

In formula (3), n represents the number of pixels in the density map, and x is the image block corresponding to the same pixel position of the generated density map and the real density map.

Experiments show that structural similarity loss can really improve the quality of density map. Compared with Euclidean distance loss, which focuses on the differences between pixels, structural similarity loss can pay more attention to the differences of corresponding local blocks between images, so as to generate density map better. In the subsequent research, the counting model SFCN^[41] also adopted a similar approach.

In order to further improve the counting accuracy, many scholars improved the structural similarity loss. DSSINet^[52] integrated the hole convolution into the structural similarity measurement, and built a hole convolution network DMS-SSIM to calculate the structural similarity loss SSIML. By expanding the receptive field of SSIM index, each pixel can fuse multi-scale information, so that at different scales, high-quality density map of local area can be output.

3.3 Generative Adversarial Loss

The crowd counting method based on density map usually takes a single static crowd image as input, and then outputs a crowd density map corresponding to the input image. This goal can be regarded as an image-to-image translation problem in essence. GAN^[22] provides a feasible idea for solving the image conversion problem, that is, by generating the network and judging the continuous game of the network, Then, the density distribution of the network learning crowd is generated, and the quality of the density map is gradually improved; Discriminating network also improves its discriminating ability through continuous training. As the key to generating countermeasure network, loss function is particularly important in the process of generating countermeasure network training and solving optimal value. In the field of crowd counting, we can use countermeasure loss function. The generated image is corrected by confrontation, so as to avoid the fuzzy problem of density map.

On the basis of Euclidean distance loss, CP-CNN^[21] network increases the generation countermeasure loss and improves the quality of predicted density map, and its loss function is as follows:

$$L_T = L_E + \lambda L_A \quad (2)$$

$$L_E = \frac{1}{WH} \sum_{\omega=1}^W \sum_{h=1}^H \|\phi(X^{\omega,h}) - (Y^{\omega,h})\|_2 \quad (3)$$

$$L_A = -\log(\phi_D(\phi(X))) \quad (4)$$

In which L_T is the total loss, L_E is the pixel-level Euclidean loss between the generated density map and the corresponding

$$L_A(G, D) = E_{x, y \sim P_{data(x, y)}} [\log D(x, y)] + E_{x \sim P_{data(x)}} [\log(1 - D(x, G(x)))] \quad (7)$$

x represents the training block and y represents the corresponding ground-truth. G is a generative network, D is a discriminant network, G tries to minimize this objective function, while D tries to maximize it. The final model is obtained through a joint training of discriminant network and generative network. RPNNet^[53] adopts a confrontation structure to extract the structural features of crowded areas.

Confrontation loss plays a significant role in improving the quality of density map, but it also has the disadvantage that it is difficult to train. In addition to the above three kinds of losses, there are many loss functions used in crowd counting tasks, such as crowd statistical loss, but each loss function has its own advantages and disadvantages, so in practical application, many kinds of losses are often combined. Build a comprehensive loss function together.

For crowd counting, the quality of density map will directly affect the counting performance. Although the existing loss function can generate density map, there are still many areas to be improved. In the future, how to define a new loss function to generate a high-quality density map is also a research focus in this field.

real density map, λ is weight factor, L_A is antagonism loss, is input image with size $W \times H$, y is ground truth density map, ϕ is a network composed of DME and F-CNN, ϕ_D is an identification sub-network used to calculate the confrontation loss, which is common in the subsequent research of crowd counting algorithm. ACSCP^[50] network uses U-Net as density map generator and uses countermeasure loss, which can be defined as:

IV. DATASETS AND EVALUATION CRITERIA

4.1 Dataset

With the rapid development of crowd counting, a large number of datasets are introduced, which can stimulate more algorithms to cope with various challenges, such as scale change, background clutter in surveillance video and variable environment, and outdoor illumination change.

In this section, we introduce some of the most commonly used demographic datasets, namely UCSD^[54], Mall^[55], UCF_CC_50^[56], WorldExpo'10^[57] and ShanghaiTech^[58], which are arranged in time sequence. The five datasets are sorted by age, and the specific statistics are listed in Table III.

UCSD^[54]. UCSD dataset is the first dataset for crowd counting, which contains 2000 pictures with resolution of 238*158. People are marked once every five pictures, and the marking information of the remaining pictures is obtained by interpolation. A total of 49,885 pedestrians are included in the dataset. 800 pictures from the 600th to the 1399th are selected as the training set, and the remaining 1,200 pictures are selected as the test set. This dataset records the flow of people in a fixed scene.

Mall^[55]. Mall dataset comes from the surveillance video of a public area of a

shopping mall, which contains 2000 pictures with resolution of 320*240, with a total of 62325 pedestrians. Compared with UCSD dataset, the scenes recorded by Mall dataset are relatively complicated: Illumination will change with the passage of time in a day, the density of crowd distribution is inconsistent, the static crowd and the active crowd exist at the same time, the perspective distortion is more serious, and the problem of object occlusion often exists.

UCF_CC_50^[56]. The first two datasets record data in fixed scenes, and UCF_CC_50 covers many different scenes, such as concerts, protest marches, stadiums, marathons and pilgrimages, which pose more challenges to the task of crowd counting. UCF_CC_50 contains only 50 pictures, and the number of people in each picture is at least 94 and at most 4543, with an average of 1280 people in each picture.

WorldExpo'10^[57]. WorldExpo'10 dataset originated from the 2010 Shanghai WorldExpo. It contains 1132 video sequences of 108 surveillance cameras with a resolution of 576*720, covering a large number of different scenes. The author's team marked 3,980 pictures, totaling 199,923 people, all of which were uniformly sampled from the video sequence.

ShanghaiTech^[58]. There are 1198 images in ShanghaiTech dataset, with a total of 330165 people. The whole dataset consists of two parts, Part A and Part B. PartA data is collected on the internet, including 482 pictures with different sizes, each picture has at least 33 people and at most 3139 people, including 24167 people in total. Part B captures the bustling street scene in the main urban area of Shanghai, including 716 pictures with resolution of 768*1024. Each picture has a minimum of 9 people and a maximum of 578 people, totaling 88,488 people.

Table III FIVE PUBLIC CROWD COUNTING DATASETS

Dataset	Resolution	Number	Maximum	Minimum	Average	Sum
UCSD	238*158	2000	46	11	25	49985
Mall	320*240	2000	53	13	31.2	62325
UCF_CC_50	different	50	4,543	94	1279.5	63,974
WorldExpo'10	576*720	3,980	253	1	50.2	199,923
ShanghaiTech	Part_A	different	482	33	501.4	241,677
	Part_B	768*1024	716	578	123.6	88,488

4.2 Evaluation Criteria

There are three indexes used to measure the performance of the model in crowd counting, MAE (Mean Absolute Error), MSE (Mean Square Error) and RMSE (Root Mean Square Error). Among them, the accuracy of MAE reaction counting and the robustness of MSE and RMSE reaction models. They can be expressed as follows:

$$MAE = \frac{1}{N} \sum_{i=1}^N |\hat{Y}_i - Y_i| \quad (8)$$

$$MSE = \frac{1}{N} \sum_{i=1}^N (\hat{Y}_i - Y_i)^2 \quad (9)$$

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (\hat{Y}_i - Y_i)^2} \quad (10)$$

Where N is the number of test set images; $\hat{Y}_i$ and Y_i are the predicted density map and the real density map of the crowd counting scene, respectively.

The original MAE, MSE and RMSE can only measure the global accuracy and robustness, but cannot evaluate the counting performance of local areas. Therefore, Tian et

al.^[59] extended MAE and RMSE into patch mean absolute error(PMAE) and patch mean squared error(PMSE) to evaluate the counting effect of local areas. In addition, for the crowd counting algorithm based on density map. The quality of density map plays a decisive role in the performance of the algorithm, so the existing image quality evaluation index can also be used to measure the performance of the counting model. PMAE and PRMSE are expressed as follows:

$$\text{PMAE} = \frac{1}{n * M} \sum_{i=1}^{n * M} |\hat{Y}_i - Y_i| \quad (11)$$

$$\text{PRMSE} = \sqrt{\frac{1}{n * M} \sum_{i=1}^{n * M} (\hat{Y}_i - Y_i)^2} \quad (12)$$

Specifically, each image is divided into n patches with the same size and no overlap, and MAE and RMSE of patches are calculated. That is to say, PMAE and PRMSE can fully reflect the local accuracy and robustness of the algorithm. When n is equal to 1, PMAE and PRMSE will degenerate into MAE and RMSE respectively. Table IV shows the MAE and MSE of crowd counting models.

TABLE IV COMPARISON OF CROWD COUNTING MODELS ON THE FIVE PUBLIC DATASETS.
SHT A and SHT B represent part A and part B of ShanghaiTech, respectively.

Type	Method	UCSD		Mall		UCF CC 5		WorldExpo'10	SHT A		SHT B	
		MAE	MSE	MAE	MSE	MAE	MSE	Avg.	MAE	MSE	MAE	MSE
Traditional approaches	Lemptsisky et al.	1.7	-	-	-	493.4	487.1	-	-	-	-	-
	RR	-	-	-	-	-	-	-	-	-	-	-
	CA-RR	2.07	6.86	3.43	17.7	-	-	-	-	-	-	-
	Count Forest	1.61	4.40	2.50	10.0	-	-	-	-	-	-	-
	MCNN	1.07	1.35	2.24	8.5	377.6	509.1	11.6	110.2	173.2	26.4	41.3
	Hydra-CNN	1.65	-	-	-	333.7	382.1	8.6	-	-	-	-
	MSCNN	-	-	-	-	363.7	468.4	11.7	83.8	127.4	17.7	30.2
	CP-CNN	-	-	-	-	295.8	320.9	8.86	73.6	106.4	20.1	30.1
	CMTL	-	-	-	-	322.8	397.9	-	101.3	152.4	20.0	31.1
	Switching-CNN	1.62	2.10	-	-	318.1	439.2	9.4	90.4	135	21.6	33.4
Modern approaches	DecideNet	-	-	1.52	1.9	-	-	9.23	-	-	21.53	31.98
	CSRNet	1.16	1.47	-	-	266.1	397.5	8.6	68.2	115	10.6	16.0
	SANet	1.02	1.29	-	-	258.4	334.9	8.2	67	104.5	8.4	13.6
	ACSCP	1.04	1.35	-	-	291.0	404.6	7.5	75.7	102.7	17.2	27.4
	ADCrowdNet	1.09	1.35	-	-	273.6	362.0	7.3	70.9	115.2	7.7	12.9
	HA-CNN	-	-	-	-	256.2	348.4	8.5	62.9	94.9	8.1	13.4
	S-DCNet	-	-	-	-	204.2	301.3	-	58.3	95.0	6.7	10.7
	DSSINet	-	-	-	-	216.9	302.4	6.67	60.63	96.4	6.85	10.34
	SAAN	-	-	1.28	1.68	271.6	391.0	-	-	-	16.86	28.41
	PGCNet	-	-	-	-	259.4	317.6	8.1	57.0	86.0	8.8	13.7
	DUBNet	1.03	1.24	-	-	235.2	332.7	-	66.4	111.1	9.4	15.1
	ADSCNet	-	-	-	-	-	-	-	55.4	97.7	6.4	11.3

V. DISCUSSION

As can be seen from the above, more and more researchers pay attention to crowd counting, however, the single scene leads to poor migration ability of the model, the low resolution of the image leads to unclear crowd characteristics, and the insufficient data scale

and sample type leads to crowd counting problems, which brings more new challenges to the crowd counting task.

In addition, the analysis of experimental data shows that attention mechanism, empty convolution or additional auxiliary information can improve network performance, and attention mechanism can help counting network focus on effective

information and eliminate noise interference; Hollow convolution can expand the receptive field, capture multi-scale information and retain more details of the image without increasing the model parameters and computational complexity; Additional auxiliary information, such as perspective, can help deal with multi-scale problems.

At present, although all kinds of crowd counting datasets have provided data support for verifying the effectiveness of counting algorithms, they still cannot meet the experimental requirements in the aspects of scene diversity, annotation accuracy and view diversity, which will be the key issues to be considered when constructing datasets in the future. For some scenarios. It is very difficult to collect images, and it is impossible to mark them accurately. At this time, artificial synthesis can be considered to generate images. For example, GCC^[60] artificially synthesizes a large number of images by generating countermeasure networks, which provides a new idea for building datasets.

VI. SUMMARY AND PROSPECT

With the development of deep learning, CNN-based crowd counting algorithm has made great progress, but it still faces many challenges to apply it to intelligent video surveillance system^[61]. In the future, we will work from the following aspects to further improve the robustness of the crowd counting algorithm:

1)Real-time. Most existing models are based on the image level, and little attention is paid to the real-time performance of the algorithm. It is a research focus of crowd counting to reduce the network complexity and improve the real-time performance of the algorithm while ensuring the counting accuracy.

2)Scene changes. Some datasets have a single scene, and the migration ability of the trained model is poor, so it is necessary to further expand the data scale and improve the adaptability of the model.

3)Block each other. With the increase of crowd density, the occlusion between people is inevitable. The next step is to study how to count people under the occlusion condition, and at the same time, obtain the detailed information such as the spatial distribution of people.

4)Illumination changes. When the illumination changes obviously, the pictures taken by the camera are often blurred, and the head cannot be clearly identified. The next step is to study how to deal with the crowd counting problem under the illumination changes.

In this paper, the related papers in the field of crowd counting are investigated. After a brief review of the traditional crowd counting algorithms, the CNN-based crowd counting methods are systematically summarized and comprehensively compared, and the future research trends in this direction are given, hoping to provide some reference for related researchers.

ACKNOWLEDGMENT:

The authors would like to thank the anonymous reviewers for their insightful suggestions. This work was supported by National Key R&D Program of China(No.2018*****02).

* is corresponding author.

REFERENCE :

- [1] Y. Zhang, D. Zhou, S. Chen, S. Gao, and Y. Ma, "Single-image crowd counting via multi-column convolutional neural network," in CVPR, 2016, pp. 589–597.

-
- [2] H. Zhang, Z. Kyaw, S.-F. Chang, and T.-S. Chua, "Visual translation embedding network for visual relation detection," in CVPR, 2017, pp. 5532–5540.
 - [3] E. Walach and L. Wolf, "Learning to count with cnn boosting," in ECCV. Springer, 2016, pp. 660–676.
 - [4] B. Leibe, E. Seemann, and B. Schiele, "Pedestrian detection in crowded scenes," in CVPR, vol. 1. IEEE, 2005, pp. 878–885.
 - [5] M. Li, Z. Zhang, K. Huang, and T. Tan, "Estimating the number of people in crowded scenes by mid based foreground segmentation and head-shoulder detection," in ICPR. IEEE, 2008, pp. 1–4. 1.
 - [6] M. Enzweiler and D. M. Gavrila, "Monocular pedestrian detection: Survey and experiments," TPAMI, vol. 31, no. 12, pp. 2179–2195, 2009.
 - [7] I. S. Topkaya, H. Erdogan, and F. Porikli, "Counting people by clustering person detector outputs," in AVSS. IEEE, 2014, pp. 313–318. 1
 - [8] R. Girshick, "Fast r-cnn," in ICCV, 2015, pp. 1440–1448. 2, 16
 - [9] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards realtime object detection with region proposal networks," in NIPS, 2015, pp. 91–99. 2, 14.
 - [10] K. He, G. Gkioxari, P. Dollar, and R. Girshick, "Mask r-cnn," in ICCV, 2017, pp. 2961–2969. 2.
 - [11] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in CVPR, 2016, pp. 779–788.
 - [12] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in ECCV. Springer, 2016, pp. 21–37.
 - [13] A. B. Chan, Z.-S. J. Liang, and N. Vasconcelos, "Privacy preserving crowd monitoring: Counting people without people models or tracking," in CVPR. IEEE, 2008, pp. 1–7.
 - [14] H. Idrees, I. Saleemi, C. Seibert, and M. Shah, "Multi-source multiscale counting in extremely dense crowd images," in CVPR, 2013, pp. 2547–2554.
 - [15] V. Lempitsky and A. Zisserman, "Learning to count objects in images," in NIPS, 2010, pp. 1324–1332.
 - [16] V.-Q. Pham, T. Kozakaya, O. Yamaguchi, and R. Okada, "Count forest: Co-voting uncertain number of targets using random forest for crowd density estimation," in ICCV, 2015, pp. 3253–3261.
 - [17] Wang Zhou, Bovik A C, Sheikh H R, et al. Image quality assessment: From error visibility to structural similarity[J]. IEEE Transactions on Image Processing, 2004, 13(4) 600–612.
 - [18] Boominathan L, Kruthiventi S S S, Babu R V, et al. CrowdNet: A deep convolutional network for dense crowd counting [C] // Proc of the 24th ACM Int Conf on Multimedia. New York: ACM, 2016: 640-644.
 - [19] Zhang Yingying, Zhou Desen, Chen Siqin, et al. Single-image crowd counting via multi-column convolutional neural network [C] // Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2016: 589-597
 - [20] Ciregan D, Meier U, Schmidhuber J. Multi-column deep neural networks for image classification [C] // Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2012: 3642-3649.
 - [21] Sindagi V A, Patel V M. Generating high-quality crowd density maps using contextual pyramid CNNs [C] // Proc of

-
- the IEEE Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2017: 1861-1870.
- [22] Goodfellow I J, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets [C] // Proc of the 27th Int Conf on Neural Information Processing Systems. Cambridge, MA: MIT Press, 2014: 2672-2680.
- [23] Sam D B, Surya S, Babu R V, et al. Switching convolutional neural network for crowd counting [C] // Proc of the IEEE Conf on Computer Vision and Pattern Recognition Honolulu. Piscataway, NJ: IEEE, 2017:4031-4039.
- [24] Sam D B, Sajjan N N, Babu R V, et al. Divide and grow: Capturing huge diversity in crowd images with incrementally growing CNN [C] //Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2018: 3618-3626.
- [25] Dollar P, Wojek C, Schiele B, et al. Pedestrian detection: An evaluation of the state of the art [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2011, 34(4): 743-761.
- [26] Liu Jiang, Gao Chenqiang, Meng Deyu, et al. Decidenet: Counting varying density crowds through attention guided detection and density estimation [C] // Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2018: 5197-5206.
- [27] Cao Xinkun, Wang Zhipeng, Zhao Yanyun, et al. Scale aggregation network for accurate and efficient crowd counting [G] // LNCS 11209:Proc of the 15th European Conf on Computer Vision. Berlin: Springer, 2018: 734-750.
- [28] Szegedy C, Liu Wei, Jia Yangqing, et al. Going deeper with convolutions [C] // Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2015: 1-9.
- [29] Hossain M, Hosseinzadeh M, Chanda O, et al. Crowd counting using scale-aware attention networks [C] // Proc of the IEEE Winter Conf on Applications of Computer Vision. Piscataway, NJ: IEEE, 2019:1280-1288.
- [30] Wang Chuan, Zhang Hua, Yang Liang, et al. Deep people counting in extremely dense crowds [C] // Proc of the 23rd ACM Int Conf on Multimedia. New York: ACM, 2015: 1299-1302.
- [31] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks [C] // Proc of the 25th Int Conf on Neural Information Processing systems. Cambridge, MA: MIT Press, 2012: 1097-1105.
- [32] Zhang Cong, Li Hongsheng, Wang Xiaogang, et al. Cross-scene crowd counting via deep convolutional neural networks [C] //Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2015: 833-841.
- [33] Yu F, Koltun V. Multi-scale context aggregation by dilated convolutions [EB/OL].(2016-04-30)[2020-12-23]. <https://arxiv.org/abs/1511.07122>.
- [34] Dai Jifeng, Qi Haozhi, Xiong Yuwen, et al. Deformable convolutional networks [C] // Proc of the IEEE Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2017: 764-773.
- [35] Goodfellow I J, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets [C] // Proc of the 27th Int Conf on Neural Information Processing Systems. Cambridge, MA: MIT Press, 2014: 2672-2680.
- [36] Li Yuhong, Zhang Xiaofan, Chen Deming. CSRNET: Dilated convolutional neural networks for understanding the highly congested scenes [C] // Proc of the IEEE Conf on Computer Vision and

-
- Pattern Recognition. Piscataway, NJ: IEEE, 2018: 1091-1100.
- [37] Liu Weizhe, Salzmann M, Fua P. Context-aware crowd counting [C] //Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2019: 5099-5108.
- [38] Isola P, Zhu Junyan, Zhou Tinghui, et al. Image-to-image translation with conditional adversarial networks [C] //Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2017:1125-1134.
- [39] Shen Zan, Xu Yi, Ni Bingbing, et al. Crowd counting via adversarial cross-scale consistency pursuit [C] // Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2018: 5245-5254.
- [40] Ronneberger O, Fischer P, Brox T. U-Net: Convolutional networks for biomedical image segmentation [C] // Proc of Int Conf on Medical Image Computing and Computer-Assisted Intervention. Berlin: Springer, 2015: 234-241.
- [41] Wang Qi, Gao Junyu, Lin Wei, et al. Learning from synthetic data for crowd counting in the wild [C] // Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2019: 8198-8207.
- [42] Zhu Junyan, Park T, Isola P, et al. Unpaired image-to-image translation using cycle-consistent adversarial networks [C] // Proc of the IEEE Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2017: 2242-2251.
- [43] Gao Junyu, Han Tao, Wang Qi, et al. Domain-adaptive crowd counting via inter-domain features segregation and gaussian-prior reconstruction [EB/OL]. (2019-12-13)[2020-12-23]. <https://arxiv.org/abs/1912.03677>.
- [44] Liu Ning, Long Yongchao, Zou Changqing, et al. ADCrowdNet: An attention-injective deformable convolutional network for crowd understanding [C] // Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2019: 3225-3234.
- [45] Wei Yunchao, Xiao Huaxin, Shi Honghui, et al. Revisiting dilated convolution: A simple approach for weakly-and semi-supervised semantic segmentation [C] // Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2018:7268-7277.
- [46] Guo Dan, Li Kun, Zha Zhengjun, et al. DADNet: Dilated-Attention-Deformable ConvNet for crowd counting [C] // Proc of the 27th ACM Int Conf on Multimedia. New York: ACM, 2019:1823-1832.
- [47] He Kaiming, Gkioxari G, Dollár P, et al. Mask R-CNN [C] // Proc of IEEE Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2017:2980-2988.
- [48] Zhu Liang, Zhao Zhijian, Lu Chao, et al. Dual path multi-scale fusion networks with attention for crowd counting [EB/OL]. (2019-02-04)[2020-12-23]. <https://arxiv.org/abs/1911.07990>.
- [49] Jiang Shenqin, Lu Xiaobo, Lei Yinjie, et al. Mask-aware networks for crowd counting [J]. IEEE Transactions on Circuits and Systems for Video Technology, 2020, 30(9): 3119-3129.
- [50] Valloli V K, Mehta K. W-Net: Reinforced U-Net for density map estimation [EB/OL]. (2019-05-29)[2020-12-23]. <https://arxiv.org/abs/1903.11249>.
- [51] Wang Qian, Breckon T P. Crowd counting via segmentation guided attention network and curriculum loss [EB/OL]. (2020-08-03)[2020-12-23]. <https://arxiv.org/abs/1911.07990>.
- [52] Liu Lingbo, Qiu Zhilin, Li Guanbin, et al.

-
- Crowd counting with deep structured scale integration network [C]// Proc of the IEEE Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2019: 1774-1783.
- [53] Yang Yifan, Li Guorong, Wu Zhe, et al. Reverse perspective network for perspective-aware object counting [C] // Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2020:4374-4383.
- [54] Shi Zenglin, Mettes P, Snoek C. Counting with focus for free [C] //Proc of the IEEE Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2019:4199-4208.
- [55] K. Chen, C.C.Loy, S.Gong, and T. Xiang, "Feature mining for localised crowd counting." in BMVC, vol. 1, no.2, 2012, p. 3.
- [56] H. Idrees, I. Saleemi, C. Seibert, and M. Shah, "Multi-source multi-scale counting in extremely dense crowd images," in CVPR, 2013, pp.2547–2554.
- [57] C. Zhang, H. Li, X. Wang, and X. Yang, "Cross-scene crowd counting via deep convolutional neural networks," in CVPR, 2015, pp. 833–841.
- [58] Y. Zhang, D. Zhou, S. Chen, S. Gao, and Y. Ma, "Single-image crowd counting via multi-column convolutional neural network," in CVPR, 2016, pp. 589–597.
- [59] Tian Yukun, Lei Yiming, Zhang Junping, et al. PaDNet: Pan-density crowd counting [J]. IEEE Transactions on Image Processing, 2020, 29: 2714-2727.
- [60] Wang Qi, Gao Junyu, Lin Wei, et al. Learning from synthetic data for crowd counting in the wild [C] // Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2019:8198-8207.
- [61] Gao Guangshuai, Gao Junyu, Liu Qingjie, et al. CNN-based density estimation and crowd counting: a Survey [EB/OL].(2020-03-28)[2020-12-23]. <https://arxiv.org/abs/2003.12783>.
- [62] Zeng Lingke, Xu Xiangmin, Cai Bolun, et al. Multi-scale convolutional neural networks for crowd counting [C] // Proc of the IEEE Int Conf on Image Processing. Piscataway, NJ: IEEE, 2017: 465-469.
- [63] Yan Zhaoyi, Yuan Yuchen, Zuo Wangmeng, et al. Perspective-guided convolution networks for crowd counting [C] // Proc of the IEEE Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2019: 952-961.
- [64] Oh M, Olsen P, Ramamurthy K N. Crowd counting with decomposed uncertainty [C] // Proc of the 34th AAAI Conf on Artificial Intelligence. Menlo Park, CA: AAAI Press 2020: 11799-11806.
- [65] Wang, Jiawei, Hongyun Xiong, Haibo Wang and X. Nian. "ADSCNet: asymmetric depthwise separable convolution for semantic segmentation in real-time." Applied Intelligence 50 (2019): 1045-1056.
- [66] Wang Zhou, Bovik A C, Sheikh H R, et al. Image quality assessment: From error visibility to structural similarity [J]. IEEE Transactions on Image Processing, 2004, 13(4) 600–612.