ITC 3/52
Information Technology and Control
Vol. 52 / No. 3 / 2023
pp. 693-712
DOI 10.5755/j01.itc.52.3.33701

A Survey on Regression-Based Crowd Counting Techniques

Received 2023/03/23 | Accepted after revision 2023/05/05

HOW TO CITE: Hao, Y., Du, H., Mao, M., Liu, Y., Fan, J. (2023). A Survey on Regression-Based Crowd Counting Techniques. *Information Technology and Control*, 52(3), 693-712. https://doi.org/10.5755/j01.itc.52.3.33701

A Survey on Regression-Based Crowd Counting Techniques

Yu Hao, Huimin Du
School of Communications and Information Engineering, Xi'an University of Posts and Telecommunications, West Changan Street, Xi'an, China

Meiwen Mao
School of Artificial Intelligence, Xidian University, 266 Xinglong Section of Xifeng Road, Xi'an, China

Ying Liu, Jiulun Fan
School of Communications and Information Engineering, Xi'an University of Posts and Telecommunications, West Changan Street, Xi'an, China

Corresponding author: haoyu@xupt.edu.cn

Traditional detect and count strategy cannot well handle the extremely crowded footage in computer vision-based counting task. In recent years, deep learning approaches have been widely explored to tackle this challenge. By regressing visual features to density map, the total crowd number can be predicted while avoids the detection of their actual positions. Efforts of improving performance distribute at various phases of the detecting pipeline, such as optimizing feature extraction and eliminating deviation of regressed density map etc. In this article, we conduct a thorough review on the most representative and state-of-the-art techniques. They are systematically categorized into three topics: the evolving of front-end network, the handling of unbalanced density map prediction, and the selection of loss function. After evaluating most significant techniques, innovations of the state-of-the-art are inspected in detail to analyze specific reasons of achieving high performances. As conclusion, possible directions of enhancement are discussed to provide insights of future research.

KEYWORDS: crowd counting, feature extraction, density map, loss functions.

1. Introduction

Crowd counting techniques are essential to ensure the public safety, such as preventing stampede in a parade, or optimizing layout of the site. To estimate the number of people within a district, a common strategy is to count the number of mobile phones accessed to the base station [1]. This strategy is generally effective, but it cannot reveal the local crowd density in the certain area with high risk, such as crossroad and pla-

za. To address this issue, image/video captured with fixed cameras can be exploited by computer vision techniques to count the crowd in the footage.

In early stage, techniques of pedestrian detection are utilized for crowd counting. This strategy attempts to detect every pedestrian in the scene, and accumulate the result to obtain the final count. The Histogram of Gradient (HOG) feature with Support Vector Machine (SVM) is a common approach [54]. In this approach, a window slides through the entire footage to obtain image patches. For each patch, the HOG is extracted and feed to SVM to classify if current patch contains a pedestrian, as illustrated in Figure 1(a). Conventional approaches are further improved to deep learning-based, such as YOLO, for a higher detection accuracy [11]. However, as density of the crowd increases, heavy occlusion and insufficient information for single pedestrian will significantly impact the performance, as illustrated in Figure 1(b).

Figure 1

(a) Crowd counting with HOG and SVM approach in [9]. (b) Crowd with extremely high density which cannot be properly handled by the detection-based approaches

(a)

(b)

Therefore, an alternated strategy is required to handle the crowd counting with extreme high density.

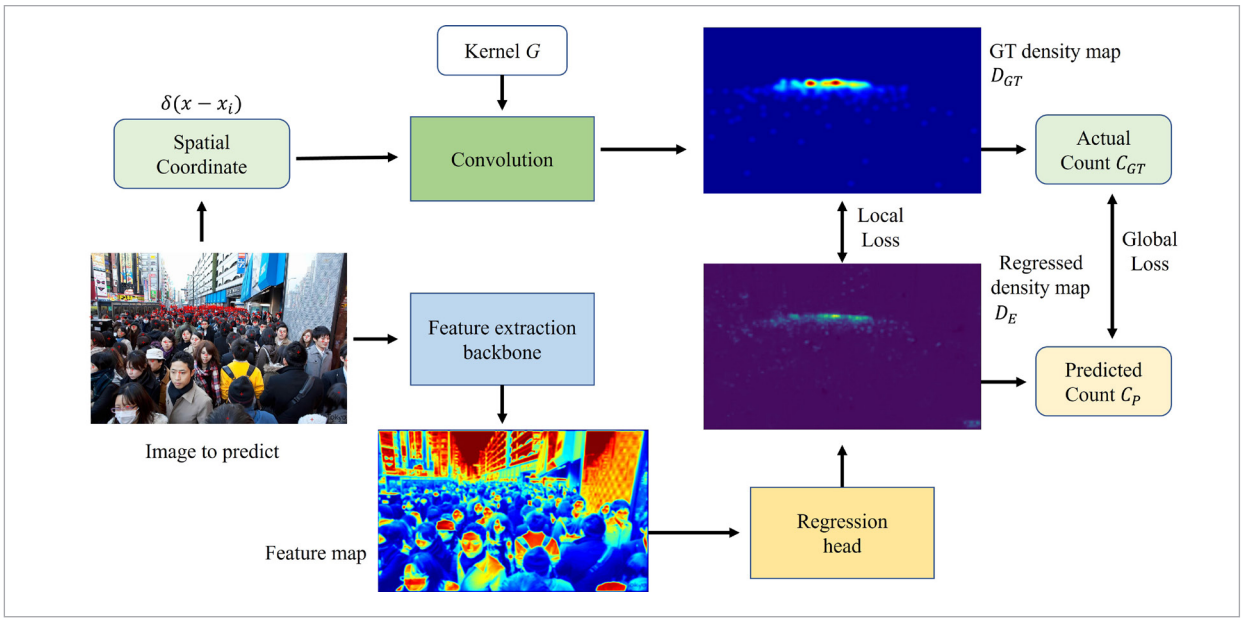
The regression-based approach addresses this issue via learning a mapping relationship between extracted visual features and estimated count. These approaches model density maps from ground-truth information as the regression target. In the training phase, extracted features are regressed towards the ground-truth/pseudo density map. In the detection phase, the trained model is exploited to predict the count with extracted features. The Figure 2 illustrates the methodology of regression-based approach. For each image, all pedestrians' heads are manually annotated with crosses as ground truth. A common approach to generate ground-truth density map D_{GT}, is to calculate the convolution of spatial information and a gaussian kernel G, which can be expressed as Equation (1).

$$D_{GT} = \sum_{i=1}^{N} \delta(x - x_i) * G, \tag{1}$$

where $\delta(x - x_i)$ represents the existence of pedestrian at spatial position x_i, the value is 1 if the pedestrian exists, otherwise it is set to 0. It can be observed that spatial position with higher crowd density corresponds with larger magnitude on the map. Simultaneously, feature map is extracted with front-end backbone network such as VGG16. Extracted features are fed to the back-end regression head to generate the regressed density map D_E. The D_E illustrated in Figure 2 is generally identical to D_{GT} except partial of background is unsuccessfully verified as the pedestrian. The predict count C_P can be simply obtained with a linear relation $C_P = k \cdot D_E$, where ratio k can be learnt from the ground-truth. The selection of loss is also crucial since it can significantly impact the prediction performance. The back-propagation process can adapt either Local or Global Loss in various techniques. The local loss measures the difference between the GT and estimated density maps. And the global loss measures the difference between actual and predict counts. Generally, the regression-based approach successfully overcomes the heavy occlusion issue which cannot be well handled by old school techniques, and makes itself a proper candidate on crowd counting.

However, the above-mentioned architecture is a primal procedure of regression-based approach, which

Figure 2
The architecture of regression-based crowd counting approach



is coarse and often inaccurate. Since the first work of regression-based approach Crowd-CNN [55] is proposed in 2015, researchers attempt to obtain better counting performance by improving this procedure on various phases. In this paper, we categorize these optimizing attempts into 3 general topics:

1 Strengthening the feature extraction network: Since the feature map is regressed to D_E, whose quality will directly impact the calculation of loss and counting result. Thus, research have been made to optimize the feature extraction process with various tactics. (1) The first encountered challenge is the pedestrians' heads often have different sizes within a footage since the existing of perspective. To extract the most appropriate feature with CNN-based network, multiple kernels with various scales are adapted. In this circumstance, approaches [56, 38, 15] attempt to devise novel networks to model better feature maps. (2) As a matter of fact, the 'ground-truth' density map in Figure 2 is not real ground-truth but generated. Thus, the generating process of D_{GT} is optimized in some approaches. Furthermore, to entirely bypass the impact of inaccurate D_{GT}, some approaches [41, 26, 46, 18] attempt to regress features into head positions instead of density map before calculating the loss. (3) The regression head requires tremendous amount of labelled data for training. However, the manual labeling of data for crowd counting is extremely exhausting. Therefore, the approach [29] attempts to achieve a satisfying performance with weak-supervised technique and limited labelled data. (4) The vision transformer has advantages such as global attention mechanism and weak supervised. Some approaches [16, 44, 18] exploit the transformer to encode feature map into sequences instead of density map, then predict the total count directly. (5) Features from sources other than image are expected to strengthen the counting result. For example, the approach [20] incorporates thermal information into the CSRNet [15] to improve the accuracy. ion into the CSRNet [15] to improve the accuracy.

2 The unbalanced prediction of density map: By analyzing the density map D_E predicted by the regression network, it is not hard to find the crowd area with ordinary density often get the most accurate prediction. The area with highest density will be over-estimated and the area with lowest density will be under-estimated. Efforts [53, 14, 42] have been made to address this issue. For example, the strategy of Learning to Scale (L2S) module [53]

is to segment and rescale the most crowded area. The rescaled area will have sparser density level and yield more accurate prediction. The Attention Scaling Network (ASNet) [14] segments D_E according to density levels and applies scaling factors on each area to adjust the unbalanced prediction within D^E. The L2S and ASNet attempt to adjust density values in certain areas, which is a coarse approach. The Scale-Adaptive Selection Network (SASNet) [42] applies the so-called Pyramid Region Awareness Loss, which refines the adjustment to pixel-level and yields better prediction.

3 Modeling the loss function: A proper loss function can update network parameters more effectively in back-propagation process and generate better model. (1) The most common way is to adapt L_1 or L^2 norm for either global or local loss. However, due to the uneven distribution of the crowd, these ordinary losses cannot well-handle the area with the highest density. Therefore, some approaches aim to implement pyramid strategy on loss calculation to improve performance. The ASNet iteratively divides D_E, and calculates loss on patch with total density lower than certain threshold. The SASNet applies similar strategy, but upgrades the approach to pixel-level. (2) The essential of typical density map-based approach is mapping density value to certain count. However, the pedestrian's head position can be predicted according to the probability of each point on D_E. Therefore, the Bayesian loss is adapted by various approaches [26, 46, 18] and achieves sound results. The probability-based approach directly predicts head positions, which is an advantage the typical approach does not have.

In this article, we will review and analyze the most representative approaches with their innovations based on the 3 above-mentioned topics. The following sections of this paper are organized as follows. Section 2 introduces the evolution of CNN-based feature extraction networks, the attempts to optimize the ground-truth density map, and the implementation of Transformer instead of CNN-based network. Section 3 introduces the attempts to eliminate the unbalanced density map prediction. Section 4 introduces the development of loss function, including approaches adapting pyramid loss and Bayesian loss. Section 5 introduces benchmarking datasets, evaluation metrics, performances of state-of-the-art crowd counting techniques, and discusses their innovations. Section 6 concludes this article.

2. Structures of Feature Extraction Networks

2.1. End-to-end Training

Despite using the regression-based structure, the Crowd-CNN [55] detects the Regions of Interest (ROI) within the footage, and extracts image patches in ROI for the training. Obviously, this strategy is not sound since it ignores regions with sparser crowd. To incorporate all information within the footage for training, the end-to-end training strategy is necessary and adapted by most CNN-based counting approaches. Besides, the Crowd-CNN must estimate the perspective of footage to adjust modeled features for regression. The incorrect estimation of perspective will lead to significant deviation of the prediction result. As the first end-to-end approach proposed in 2016, the Multi-Column Convolutional Neural Network (MCNN) [56] applies 3 columns of CNN with various filter scales on the entire input image to obtain visual features at different receptive fields. This multi-column feature extraction backbone network is expected to handle the irregular distribution of head's scale, which eliminates the requirement of perspective estimation. For the regression head, the fully connected layer is replaced with a 1×1 convolution layer [24] to avoid deformation of the feature map. A standard L_2 is adapted as Loss. Another milestone of MCNN is it proposed the ShanghaiTech dataset, which later becomes the benchmarking dataset for the evaluation of crowd counting techniques.

The ASNet pointed out different crowd density levels often lead to different predicting deviations, which is named as the problem of unbalanced density estimation and discussed later in Section 3 of this paper. Based on the principle of MCNN, the Context Pyramid CNN (CP-CNN) [38] integrates contextual information into the estimated density map D_E in year 2017. CP-CNN believes adjusting D_E according to different density levels will produce more accurate prediction. Thus, CP-CNN proposed a single-column network (Global Context Estimator) to classify the footage into different density levels. Similarly, the

Local Context Estimator is devised to classify patches within the footage. The results from Global/Local Estimator and MCNN backbone will be merged into the estimated density map with a Fusion-CNN. The 5 columns network structure of CP-CNN solved the unbalanced density map estimation problem in certain extent.

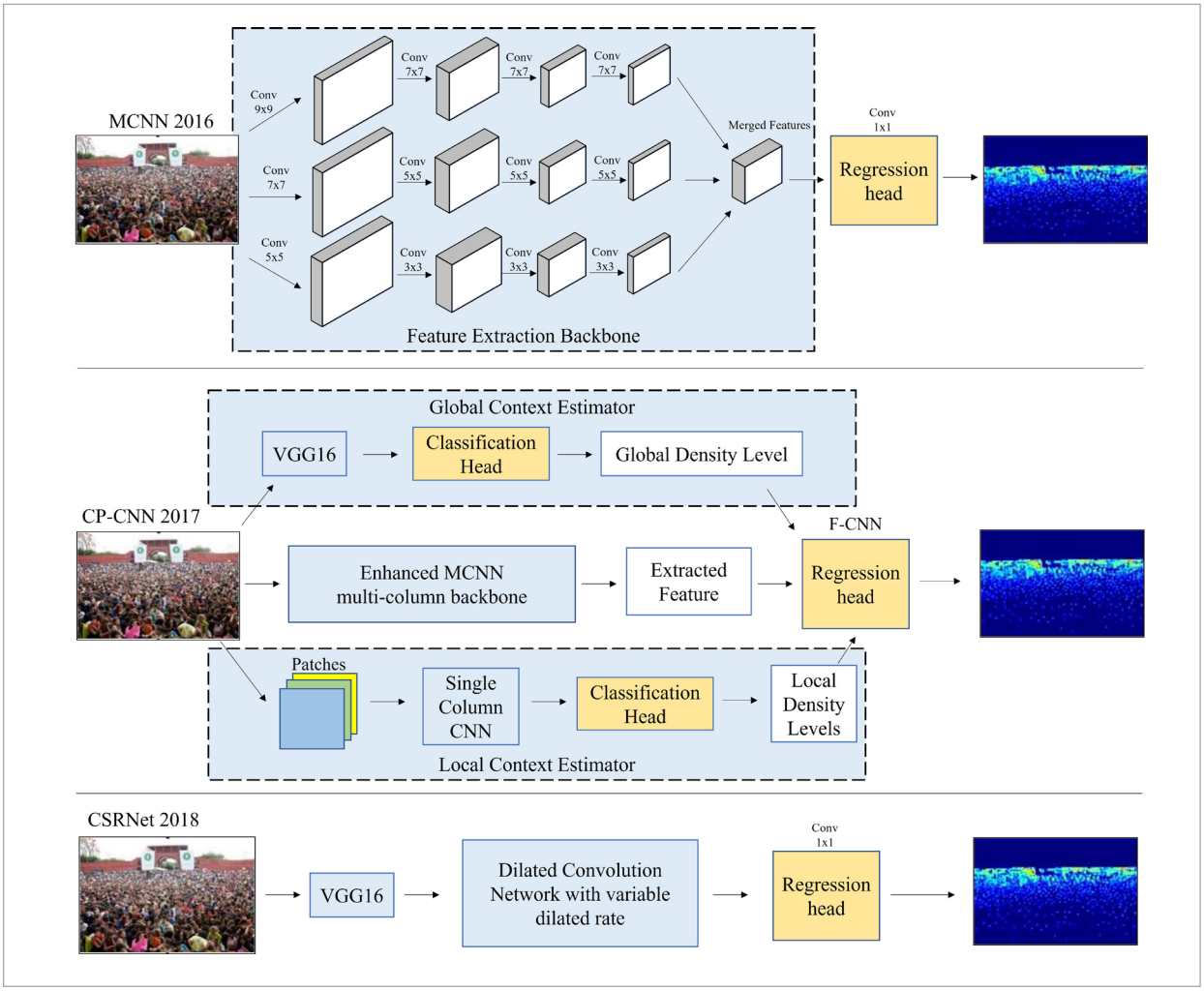
2.2. Single-Column Outperforms Multi-Column

The Switch-CNN [2] adapts the hybrid strategy of Crowd-CNN and MCNN. By setting a switch classifier before the multi-column backbone, divided image patches are fed to the single selected column according to the decision of the switch. Despite the Switch-CNN is not an end-to-end network, it still outperforms the MCNN. This observation indicates the multi-column structure does not always provide more 'valid' features than single-column.

Besides, the primary defect of multi-column network structure is the low efficiency of training process in each column. Moreover, experiment results indicate the actual performances of each column are identical, which cannot reveal the nature of different density levels as expected. One strategy to address this issue is replacing the MCNN's multi-columns with a single-column structure, to increase the efficiency of feature extraction. The perception process of various density

Figure 3

Structural evolution of front-end feature extraction networks at early stage (a) MCNN (b) CP-CNN (c) CSRNet



levels can be moved to the regression head. The CSRNet adapted the first 10 layers of a pre-trained VGG16. And a 4-way dilated convolution is attached ahead of the 1×1 regression head to expand the receptive field instead of max pooling. The CSRNET successfully reduces the MAE of ShanghaiTech-A from hundred level to 60 level, which is a tremendous boost on counting performance. As the result, the single-column feature extraction strategy became the main-stream and adapted by most CNN-based counting approaches.

2.3. The Inaccurate 'Ground Truth' Density Map

According to the architecture illustrated in Figure 3, the ground truth density map D_{GT} is obtained by convolving annotated head positions with kernel G. Since the D_{GT} will be used to calculate the Loss, it directly impacts the accuracy of the network. Ideally, the D_{GT} should objectively describes the actual distribution of the crowd. However, when a footage contains pedestrians both far and near toward the camera, occupied pixels of their heads will vary. If the scale of kernel G is fixed, the obtained D_{GT} will barely be optimized. Naturally, various scales of G can be applied to generate density map for different head sizes, which leads to the issue to be addressed: the strategy to implement kernels with various scales.

A common approach is to select the variance of G for pedestrian x_i according to the average distance of his/her m nearest neighbors. Assuming the m nearest distances to x_i is $\{d_1^i, d_2^i,...,d_m^i\}$, the average distance can be expressed as Equation (2).

$$\bar{d}^i = \frac{1}{m}\sum_{j=1}^{m} d_j^i \tag{2}$$

Therefore, the optimized D'_{GT} can be generated from the kernel G_σ with dynamic variance σ as Equation (3).

$$D'_{GT} = \sum_{i=1}^{N} \delta(x - x_i) * G_{\sigma_i} \quad with\ \sigma_i = \beta \bar{d}^i \tag{3}$$

Ablation experiments indicate using D'_{GT} will yield better prediction, and more research explored ways to generate D'_{GT} with higher quality. The Point to Point Network (P2PNet) [41] proposed a novel strategy to eliminate the impact of inaccurate ground truth density map. Instead of regressing features toward D_{GT}, the P2PNet directly regressed the VGG16 extracted feature map toward head points $\delta(x_i)$. A parallel classification network branch is then applied to provide confidence score of the predicted head points. Its Loss comprises with both regression and classification Losses. This strategy makes P2PNet to achieve the highest performance until 2022 to our best knowledge.

2.4. Weak/semi-supervised Solutions

The training of regression-based approaches usually requires large number of annotated data. However, the manual annotation of ground truth data for crowd counting is extremely exhausting - usually thousands of people need to be labelled for a single image. For example, the benchmarking ShanghaiTech A dataset [56] includes 482 images with average 1000+ pedestrians for each image. Therefore, weak/semi-supervised approaches [10, 29, 30, 59] are proposed to address this issue. The strategy of weak-supervised solution is adapting small-sample approach such as the Transformer [10]. The approach using transformer will be introduced in Section 2.5, and we will introduce semi-supervised solution here.

The Mean Teacher [43] is a semi-supervised network based on Temporal Ensembling and the ∏ model proposed in 2017. The mean teacher is composed with a double-routes teacher/student network as typical semi-supervised approaches. Unlike others, it updates the parameters of the teacher network with exponential moving average (EMA) to enhance performance, instead of replicating the student's parameters.

For the problem of crowd counting, labeled data are fed to student network and unlabeled data are fed to teacher network. Since the teacher's parameter is updated with the student, once trained, both networks can be used to predict the density map. Thus, the semi-supervised training is achieved. For further optimization, Semi [29] adapted the mean teacher as the baseline structure with the binary segmentation. The observation indicates that spatial information can be utilized to segment the crowd and background [57], which will improve the counting performance. Therefore, Semi exploits the binary segmentation to estimate the uncertain spatial regions from the regressed density map. The uncertainty map is obtained by calculating the entropy of density map and filtering with a threshold. With the uncertainty map, the density value in background area will be removed.

2.5. Replacing CNN with Transformer

The mechanism of CNN-based approaches for crowd counting determines its receptive filed is often local, since limited scales of the convolution. Despite initially utilized for natural language processing like LSTM, the Transformer [45] devised by Google is adapted by various computer vision tasks [7, 4, 58]. Due to its attention mechanism, the Transformer possess the global receptive field which can directly estimate the total crowd number, instead of accumulating the local predictions. To be specify, it is firstly adapted by Detection Transformer (DETR) [4], which utilizes CNN for feature extraction and Transformer for classification. The Vision Transformer [7] is the first to purely exploit the Transformer to address the task of image classification, by sequencing patches with the encoder instead of extracting features. The transformer is also used to generate the pre-trained image model in the Segmentation Transformers (SETR) [58]. The detailed introduction of exploiting the Transformer on image processing can be found at sources [10, 31, 19].

Another feature of transformer, as well as other language processing network such as RNN, is they can also be weak-supervised. Transformer can conduct a pre-training on the non-supervised large dataset, and then complete the training with an annotated small-sample dataset. Therefore, this weak-supervised feature makes transformer another possible candidate to handle the crowd counting task with limited annotated data.

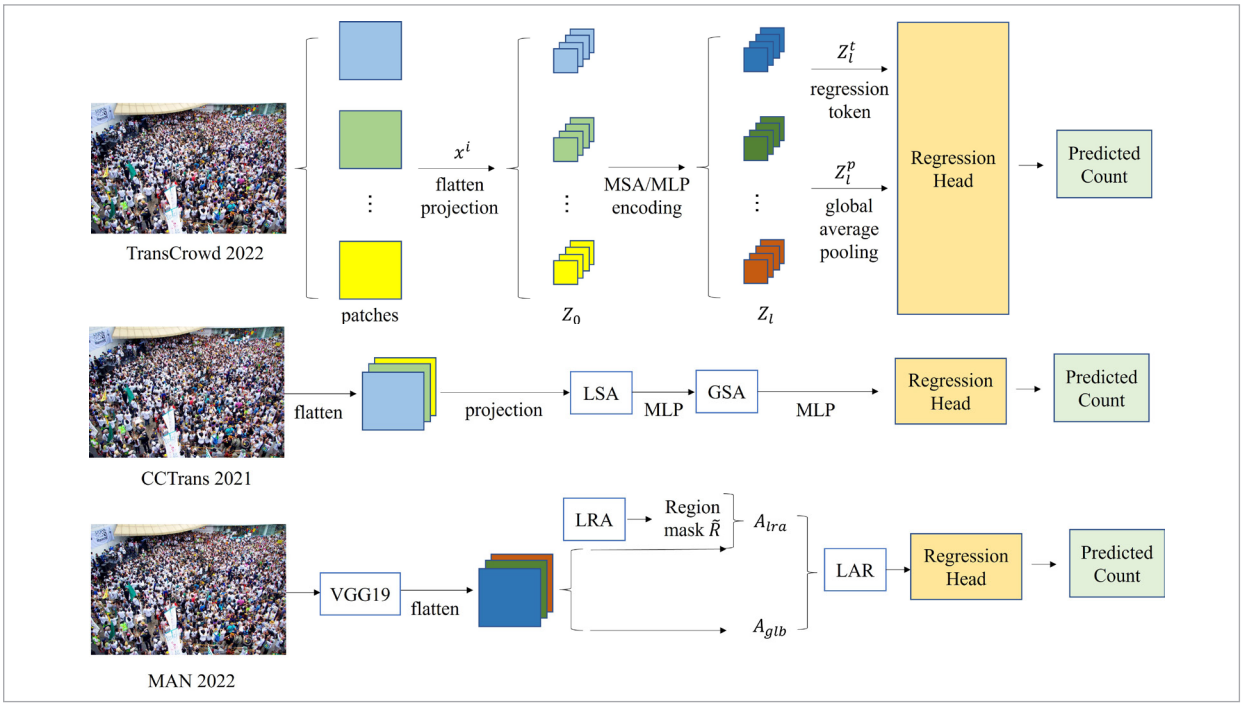
The TransCrowd [16] adapted the pure Transformer to achieve the crowd counting. As illustrated in Figure 4, its general strategy is encoding the divided patches into the vector sequences as the input, and feeding them to the encoder. Then, the encoded sequences processed with either regression token or global average pooling, will be used to predict the count with an ordinary regression head, instead of a decoder.

Specifically, the image patches are firstly flattened to N sequence, represented as $\{x^i \mid i = 1, ..., N\}$. Next, x^i is mapped with a learnable matrix E into a latent D-dimensional embedding feature. Furthermore, the spatial information $\{p^i \mid i = 1, ..., N\}$ is integrated as well to generate the input Z_0 for the encoding process.

$$Z_0 = [Z_0^1; Z_0^2; ...; Z_0^N] = \\ = [x^1E + p^1; x^2E + p^2; ...; x^NE + p^N]. \quad (4)$$

Figure 4

Structures of transformer-based approaches (a) TransCrowd (b) CCTrans (c) MAN



The encoder comprises with multiple layers of Multi-head Self-Attention (*MSA*) and Multilayer Perceptron (*MLP*). The output Z_l of the *l-th* layer can be expressed as Equation (5), where *LN* represents the layer normalization process.

$$Z'_l = MSA\big(LN(Z_{l-1})\big) + Z_{l-1} \tag{5.1}$$

$$Z_l = MLP\big(LN(Z'_l)\big) + Z'_l, \tag{5.2}$$

where the *MSA* calculates *m* Self Attention models SA_m with a reprojection matrix W_o, expressed as $MSA(Z_l) = [SA_1(Z_l); SA_1(Z_l); ...; SA_1(Z_l)]W_o$. The SA_m can be obtained with the typical Query(*Q*) /Key(*K*)/ Value(*V*) paradigm of classic transformer. The *MLP* uses 2 linear layers with the GELU [6] activation function to expand and shrink the embedding dimension of the feature.

$$SA(Z_l) = softmax\left(\frac{(QW^Q)(KW^K)^T}{\sqrt{D}}\right)VW^V \tag{6}$$

The obtained Z_l is further processed with either Regression Token or Global Average Pooling before sent to the regression head. The regression token procedure attaches an additional token Z^t_o to Z_0. After the *MSA* and *MLP*, the Z^t_l contains the global semantic crowd information. This strategy is adapted by Bert [5] as well. Applying the global average pooling to Z_0 will generated the pooled visual tokens Z^p_l. Since the Z^p_l has more discriminative semantic patterns, it obtained better performance than using Z^t_o.

Similarly, CCTrans [44] also adapted transformer as the feature extraction backbone. Unlike TransCrowd, patches x^i of the input image are flattened into a single sequence, then a learnable projection is applied to obtain the input sequence Z_{l-1} for the *l*-th layer. For the encoding backbone, the Twins network [18] is adapted. The Twins can perceive both local and global receptive fields via alternated local and global attentions, namely Spatially Separable Self-Attention (*SSSA*) module. Specifically, for the local attention, the Locally-grouped Self-Attention (*LSA*) and *MLP* are applied to $LN(Z_{l-1})$. For the global attention, Global Sub-sampled Attention (*GSA*) and *MLP* are further applied to obtain the feature sequence Z_l for regression head. Comparing with the TransCrowd, integrated local attention of the *SSSA* provides additional perceptions on local features. The CCTrans ranked first on the online dataset NWPU-Crowd in year 2021.

$$Z'_l = LSA\big(LN(Z_{i-1})\big) + Z_{i-1} \tag{7.1}$$

$$Z''_l = MLP\big(LN(Z'_l)\big) + Z'_l \tag{7.2}$$

$$Z'''_l = GSA\big(LN(Z''_l)\big) + Z''_l \tag{7.3}$$

$$Z_l = MLP\big(LN(Z'''_l)\big) + Z'''_l \tag{7.4}$$

The Multifaceted Attention Network (MAN) [18] also attempts to incorporate local attention into the transform. Unlike the TransCrowd and CCTrans, the MAN firstly used VGG19 to obtain the feature map. During the encoding phase, the MAN proposed a Learnable Region Attention (LRA) mechanism to optimize the local value within the final attention. After the region mask $\tilde{R}$ is obtained with the LRA, the regional attentions can be expressed as.

$$A_{lra} = softmax\left(\frac{\left(QW^Q_{loc}\right)(KW^K_{loc})^T \bullet \tilde{R}}{\sqrt{D}}\right)VW^V \tag{8}$$

where • is the Hadamard product. And the global attention A_{glb} can be expressed in an ordinary transform way. Note the A_{lra} and A_{glb} are sharing the same value vectors W^V.

$$A_{glb} = softmax\left(\frac{\left(QW^Q_{glb}\right)\left(KW^K_{glb}\right)^T}{\sqrt{D}}\right)VW^V \tag{9}$$

The final attention *A* can be obtained as:

$$A = A_{lra} + A_{glb}\ . \tag{10}$$

In summary, the core strategy of TransCrowd, CCTrans and MAN, is to strengthen the global attention via involving local attention during the encoding phase. The TransCrowd divides image into patches, and models self-attention *SA* for each patch as local attention. Then, the sequence of local attentions is fed to the regression head. The CCTrans encodes the image into a single sequence and perceives local attention before global attention. The MAN obtains the regional and global attentions separately, then integrates them into the final attention. De-

spite adapting different encoding procedures, all approaches receive sound results in the experiments.

2.6. Features from Other Sources

All above-mentioned approaches only exploit visual features extracted from images or video frames. Research [20, 23, 52] attempts to improve the counting performance via utilizing features from other source. The multi-model techniques include multi-model representation, translation, alignment, multi-model fusion and co-learning. The multi-model fusion integrates various types of information and gives prediction. A typical application of multi-model fusion in image processing is the visual-audio recognition, which extract visual and audio features to perform personal identification.

The Information Aggregation Distribution Module (IADM) [20] devised a multi-model approach, which incorporates thermal information with visual features. In the Information Aggregation Transfer phase, 3 branches of CSRNet are used to extract visual, thermal, and modality-shared features. The modality-shared feature describes the complementary information between visual and thermal features. In the information distribution transfer phase, the contextual information obtained from modality-shared feature is used to refine the thermal and visual features for the regression of density map.

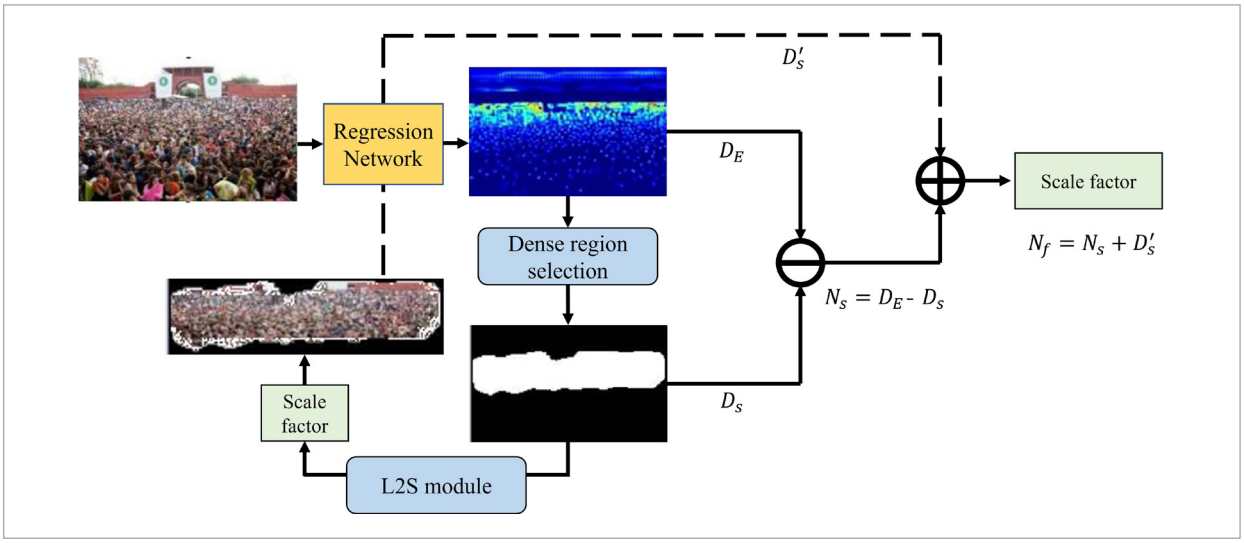
3. The Problem of Unbalanced Density Estimation

The ASNet [14] observed a phenomenon that the sparse area in the regressed density map D_E often yield smaller predict count than ground truth, and dense area in D_E often yield larger count. Therefore, the performance of existing regression-based approaches is significantly impacted on datasets with wide density range, such as UCF-QNRF [13] and UCF_CCF_50 [12]. The strategy to address this issue is rescaling the regions within the image to same density level before/after the prediction. Two approaches are proposed to handle this issue.

3.1. Rescaling Regions of the Image into Identical Density Level Before Prediction

The L2S [53] attempts to locate regions with high density, and rescale these regions until the density is identical to the sparser region. The partially rescaled image has more evenly distributed density, and the prediction is expected to be more accurate. As illustrated in Figure 5, an initial density map is firstly predicted with the regression-based network. Next, the dense region is selected by a threshold, and the L2S module is exploited to generate a Scale factor. The dense region is then rescaled with the factor, and fed to the network again to obtain the optimized local

Figure 5
Density map re-scale process proposed in L2S [53]



prediction. This approach achieved the best performance in 2019 crowd counting game CV101.

3.2. Predicting the Density Map and Applying Factors to Regions with Different Density Level for the Magnitude Adjustment

The ASNet [14] devised a post-processing mechanism to optimize D_E. As illustrated in Figure 6, image is segmented into multiple density regions with the Density Attention Network (DANet) to generate the attention masks $[M_1, M_2, ..., M_n]$. Once D_E is generated with the feature extraction network, the Attention Scaling Network (ASNet) will map regions with scaling factors $[s_1, s_2, ..., s_n]$. With the attention mask, region with high density level is set lower to give more accurate prediction, and vice versa.

The above-mentioned approaches attempt to align density of regions within the image, which boost the performance on certain dataset. However, if the approach is practically applied, data for prediction can be obtained from various sources and datasets. The experiment result of Scale Distribution Alignment (SDA) [25] indicates the performance of a state-of-the-art [26] suffered a substantially decrease by simply rescaling testing images from the same dataset for training. It is expected images with different scales from other dataset will case a greater impact on prediction.

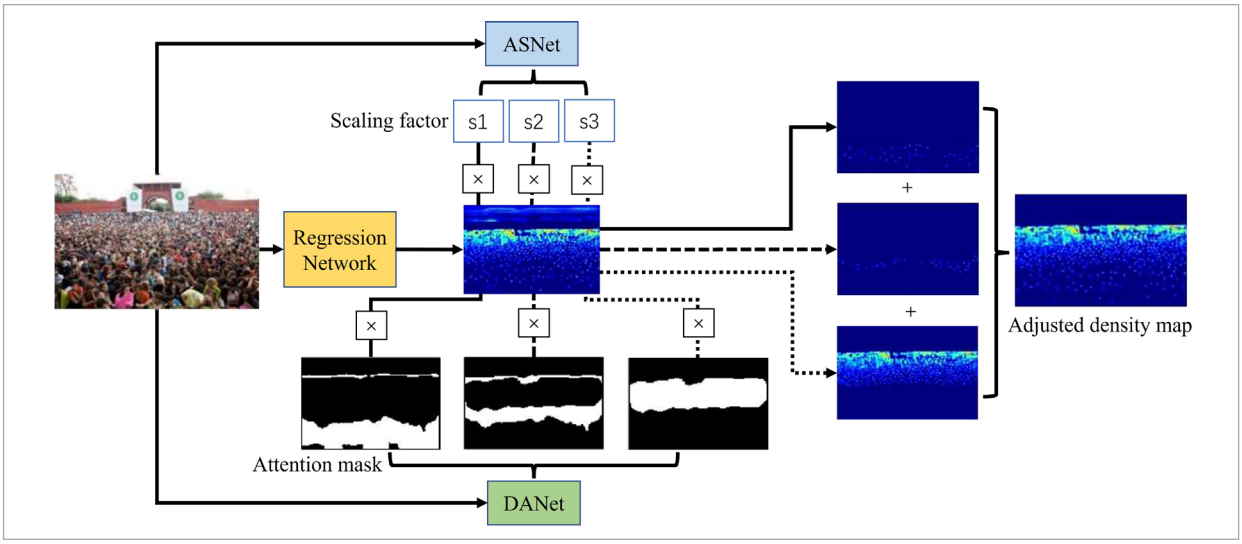
The strategy of SDA is to align the image scales of multiple datasets with learnt rescaling factors. In this approach, the Scale Distribution Network (SD-Net) is devised to estimate the scale distribution of each image. Next, images are divided into patches for the alignment by re-scaling them with the optimal translation factor. The factor for each patch is calculated with its actual distribution and Wasserstein barycenter of the estimated scale distribution. With this process, the scale distributions for 4 benchmarking datasets are aligned to the same level. The ablation experiment shows the aligned database outperforms the original on main-stream approaches such as CSRNet and BL [26].

4. The Selection of Loss Function

4.1. Basic Loss

The loss function evaluates the difference between the predicted and ground-truth count, which is exploited to adjust the weight of network in the back-propagation process. Therefore, the appropriate selection of Loss function can directly boost the accuracy of prediction. The straight-forward way is calculating the L_1 norm between the estimated density map D_E and GT density map D_{GT}, which can be expressed as $||D_E - D_{GT}||^1$ or the following equation, where M is the total amount of images for training.

Figure 6
Rescaling process of Density map in ASNet [14]



$$L_1 = \frac{1}{M}\sum_{i=1}^{M}\left|D_E^i - D_{GT}^i\right| \tag{11}$$

Some approaches such as SASNet adapt L_2 norm $||D_E - D_{GT}||^2$ as the loss.

$$L_2 = \frac{1}{M}\sum_{i=1}^{M}\sqrt{\left(D_E^i - D_{GT}^i\right)^2} \tag{12}$$

Since transformer-related approaches encode features into sequences instead of density maps, D_E and D_{GT} are replaced with the predicted count C_P and actual count C_{GT}. As claimed in Section 1, we note the density map generated loss as Local Loss, and the count generated loss as Global Loss.

To improve the prediction performance, research [44, 14, 42, 26, 46, 18] are conducted to further optimize the loss function. The Smooth L_1 is an alternate version of L_1 norm, and often adapted to handle the exploding gradient problem. It is a hybrid of L_1 and L_2, and can be expressed as Equation (13). CCTrans [44] uses the Smooth L_1 as the Loss and claims to obtain better performance than the ordinary L_1.

$$Smooth\ L_1 = \begin{cases} |x| - 0.5, |x| > 1 \\ 0.5x^2, |x| < 1 \end{cases}. \tag{13}$$

4.2. Pyramid Loss

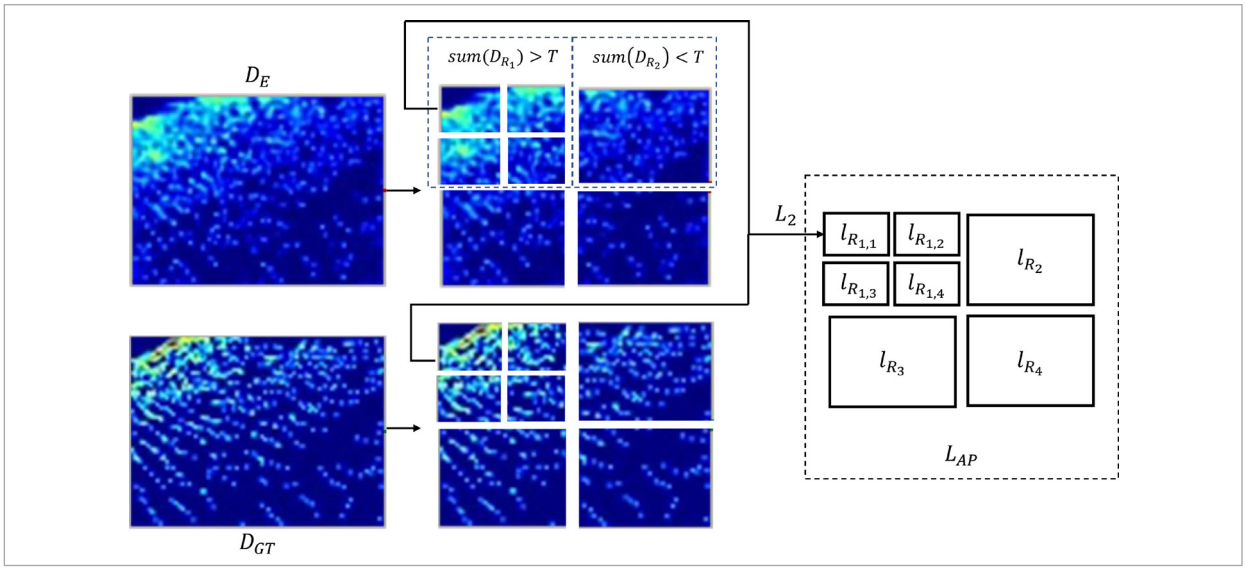
As mentioned in Section 3, the unbalanced density will impact the accuracy of count prediction. Regions with identical density level still have different density values. This issue also affects the loss calculation while training. Therefore, the ASNet [14] devised the Adaptive Pyramid Loss (APL) to handle the unbalanced density within the predicted density map. By iteratively dividing the map with count larger than threshold T into 4 subregions, the APL L_{AP} can measure the difference more accurately in extremely crowded regions. As Figure 7 illustrated, the L_{AP} can be defined as

$$L_{AP} = \frac{1}{M}\sum_{k=1}^{M}\sum_{i_1=1}^{4} l_{R_{i_1}}^k, \tag{14}$$

where the density map k is divided into 4 subregions R_1~R_4 in the first iteration i_1, and the Loss $l_{R_{i_1}}^k$ for each subregion can be defined as

$$l_{R_{i_1,\ldots,i_{n-1}}}^k = \begin{cases} \frac{\left\|D_{R_{i_1,\ldots,i_{n-1}}}^E - D_{R_{i_1,\ldots,i_{n-1}}}^{GT}\right\|^2}{sum\left(D_{R_{i_1,\ldots,i_{n-1}}}^k\right) + \sigma}, sum\left(D_{R_{i_1,\ldots,i_{n-1}}}^k\right) < T \\ \sum_{i_n=1}^{4} l_{R_{i_1,\ldots,i_n}}^k, \ otherwise \end{cases} \tag{15}$$

Figure 7
Adaptive Pyramid Loss



Similarly, SASNet devised the Pyramid Region Awareness Loss (PRAL) to handle the over-estimated value of density map. This approach divides the predicted density map into 4 subregions, and locates the most over-estimated subregion. Next, the located subregion will be iteratively divided to pixel-level. All selected pixels will be collected as a hard pixel set H. Then the PRAL L_{PRAL} can be modelled as the following Equation (16), where γ is a weight term.

$$L_{PRAL} = L_2 + \gamma \| D_E^H - D_{GT}^H \|^2. \quad (16)$$

4.3. Bayesian Loss

The ground-truth density map is generated from the head annotations with fixed/dynamic Guassian kernel. In either way, the density map is not real 'ground-truth'. In Section 2.3, approaches attempt to model the most accurate ground-truth density map. On the other hand, some approaches aim to tackle this issue by skipping the density map. The Bayesian Loss (BL) [26] exploits the probability of every spatial location belongs to the head annotation, to calculate the Bayesian Loss L_{Bayes} of the entire image. In this case, each pixel on the feature map will be mapped to all head annotations with probabilities. L_{Bayes} can be expressed as Equation (17).

$$L_{Bayes} = \sum_{n=1}^{N} \| 1 - E[c_n] \|^1 \quad (17)$$

Where N is the total head annotation count within the image, $E[\bullet]$ is the expectation, c_n denotes the count that pixels x_i belongs to the $n-th$ annotation. Since the head location is annotated with single pixel, the $E[\bullet]$ of ground-truth count will be 1.

Considering background pixels can be far away from any pedestrian's head, they should not be mapped to any annotation. Therefore, BL improved the Bayesian Loss into L_{Bayes+} by adapting the expectation of background count $E[c_0]$.

$$L_{Bayes+} = \sum_{n=1}^{N} \| 1 - E[c_n] \|^1 + \| 0 - E[c_0] \|^1 \quad (18)$$

The BL did not consider the cost of mapping pixels to annotations, which is referred as the Transport Cost in Generalized Loss (GL) [46]. For example, for the crowd far from the camera, heads are more compact. The transport cost should be higher to produce higher Loss value. The GL introduced a generalized Loss based on the hybrid of multi-Loss functions and the transport cost.

$$L_C = min\langle C, c_n \rangle - \varepsilon H(c_n) + \tau L_1 + \tau L_{Bayes} \quad (19)$$

here C is the transport cost, by minimizing the $\langle C, c_n \rangle$, the predicted density is pushed toward the annotation. The entropic regularization term $H(\bullet)$ can make the distribution of density sparser. GL later proved the L_1 and L_{Baye} are special and suboptimal cases of L_c.

The MAN attempts to address the over-estimated issue as [14, 42] using Bayesian Loss. To suppress the false positive prediction, MAN proposed Instance Attention Loss L_{IA} by adapting the instance attention mask to prune the L_{Bayes} value larger than threshold δ. The L_{IA} can be expressed as.

$$L_{IA} = \sum_{n=1}^{N} m_n \bullet \| 1 - E[c_n] \|^1, \quad (20)$$

where the instance attention mask m can be expressed as follows. This means if δ is larger than the calculated Bayesian loss of 80% annotated points, 20% of points will be pruned in back-propagation process. Therefore, the issue of over-estimated prediction can be handled when the δ is properly set.

$$m_n = \begin{cases} 1, & if \| 1 - E[c_n] \|^1 < \delta \\ 0, & otherwise \end{cases} \quad (21)$$

5. Experiments and Discussions

5.1. Benchmarking Datasets

Standard datasets are used to measure the performance between different approaches. The following section gives introductions to mainstream datasets.

The Shanghai Tech (SHT) dataset [56] is proposed with the MCNN. It soon became the first benchmarking dataset and is adapted by most of the regression-based approaches. The SHT comprises with subsets A and B, pedestrians are manually annotated in all image samples. The subset A contains 300 training samples and 182 testing samples collected from the Internet. The subset B contains 400 training samples and 316 testing samples collected from cameras installed at streets. The crowd density in subset A is

significantly higher than subset B. Therefore, some approaches did not choose subset B while using SHT.

The UCF-QNRF [13] is a challenging dataset to evaluate adaptiveness of the approach, since it has a wide range of crowd number. The minimum count is 49, and the maximum count is 12,865. The scale of UCF-QNRF is also very large, it contains 1,535 image samples and 1.25 million annotations, where 1,201 samples belong to training set and 334 samples belong to testing set. The large dataset NWPU-Crowd [50] has 5,109 images and 2.13 million annotations. Similarly, it has a wide count ranges from 0 to 20,033.

The UCF_CCF_50 [12] is a small but frequently referenced dataset with 50 gray-scale images. Like the UCF-QNRF, its range of count varies from 94 to 4,543, and the average count for each image is 1,280. The total number of annotations is 63,974. Note that the UCF_CCF_50 is not divided as training and testing data. Usually, researcher defines 50% of the set as the training data [26].

The JHU-Crowd++ [40] is another large-scale dataset, which contains 4,372 images and 1.51 million annotations. This dataset is divided into 2,772 training and 1,600 testing images. Images of JHU-Crowd++ are carefully collected according to adverse weather conditions.

The WorldExpo'10 [55] is a video dataset, where partial frames are annotated with 199,923 pedestrians. Its training set contains 3380 frames from 103 scenes, and its testing set contains 600 frames from 5 scenes.

The following table 1 lists the scales and annotation statistics of above-mentioned datasets to provide a straight-forward comparison.

Aiming to achieve crowd counting with multi-modal approach, datasets with information from additional sources are also proposed. The RGBT-CC [20] dataset contains 2030 RGB images with their corresponding thermal versions collected from optical-thermal camera. It has total 138,389 annotations. One unique feature of RGBT-CC is that partial thermal/RGB images are captured in darkness, which makes it a fine choice to evaluate performance under limited brightness circumstance.

5.2. Evaluation Metrics

The evaluation of the performance of the approach is straight-forward: if the predicted count matches the actual crowd number. Without considering the computational complexity, Mean Absolute Error(MAE) $/L_1$ Loss and Mean Squared Error (MSE)$/L_2$ Loss are often adapted as evaluation metrics by regression-based approaches.

$$MAE = \frac{1}{M}\sum_{i=1}^{M}\left|C_P^i - C_{GT}^i\right| \tag{22.1}$$

$$MSE = \sqrt{\frac{1}{M}\sum_{i=1}^{M}\left|C_P^i - C_{GT}^i\right|^2} \tag{22.2}$$

$$NAE = \frac{1}{M}\sum_{i=1}^{M}\frac{\left|C_P^i - C_{GT}^i\right|}{C_{GT}^i} \tag{22.3}$$

Where C_P^i is the predicted count, C_{GT}^i is the ground-truth, M is the total number of images in the dataset. Furthermore, the mean Normalized Absolute Error (NAE) is a recently proposed metric, which is adapted by some research [49]. Additionally, the Grid Average Mean Absolute Error (GAME) [8] is proposed to measure the MAE within different regions. For any level l, the image is divided into 4^l non-overlapping regions. Then the GAME at level i can be expressed as Equation (23).

$$GAME(l) = \frac{1}{M}\sum_{i=1}^{M}\sum_{j=1}^{4^l}\left|C_P^{i,j} - C_{GT}^{i,j}\right|. \tag{23}$$

Some metrics are devised to provide a more general presentation of basic metrics. For example, MAE and GAME can be special cases of Patch Mean Absolute Error (PMAE), and MSE is a special case of Patch

Table 1
Patterns of mainstream datasets for crowd counting

Dataset	SHT A	UCF-QNRF	UCF_CCF_50	NWPU	JHU++	WorldExpo'10
Images	482	1535	50	5109	4372	3980
Annotations	244167	1.25million	63974	2.13million	1.51million	199923

Mean Squared Error (PMSE). Moreover, conventional metrics for signal processing are also adapted in some cases, such as Peak Signal to Noise Ratio (PSNR) and Structural Similarity Index (SSIM). However, MAE and MSE are still at dominate position and adapted by nearly all approaches.

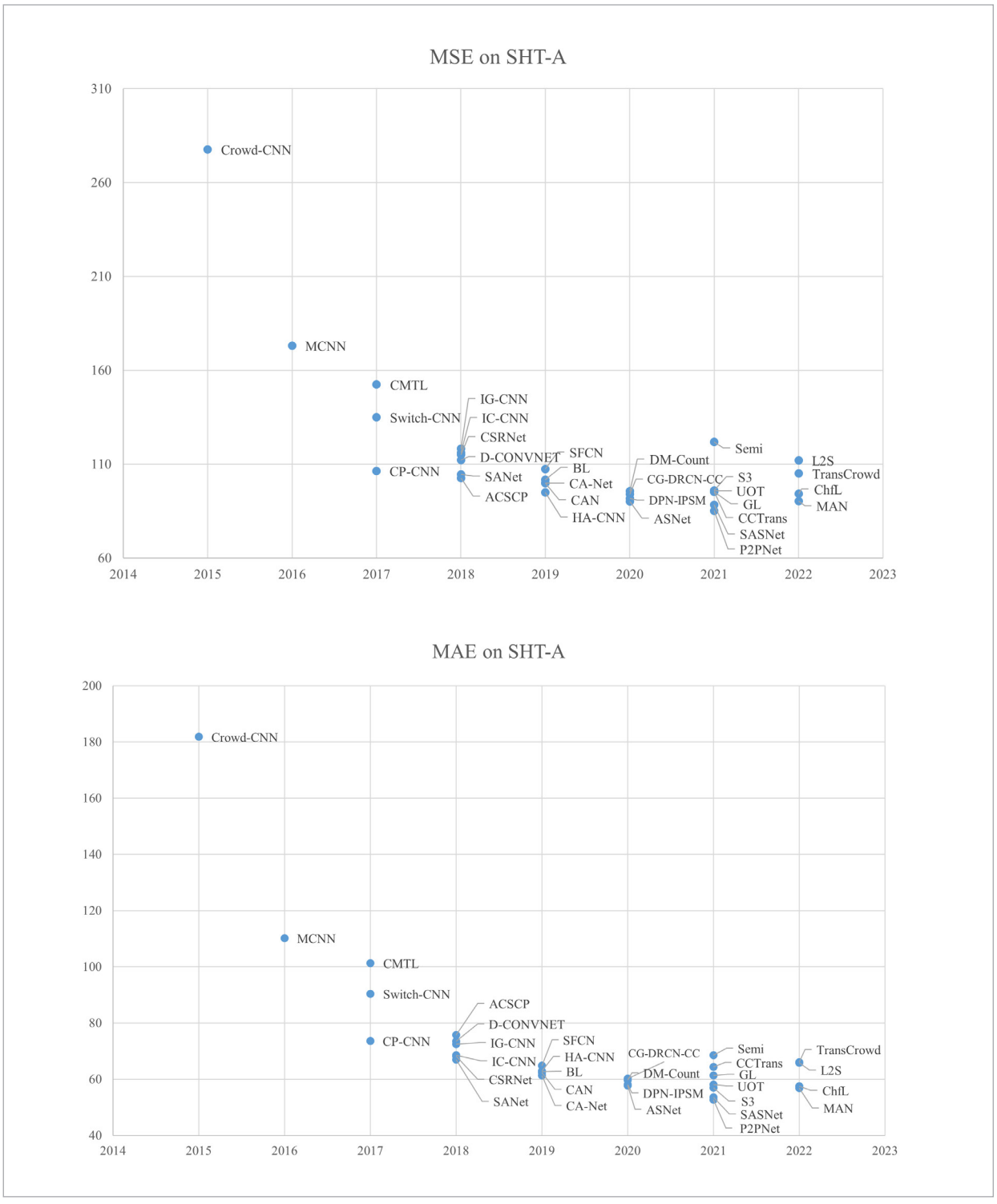
5.3. Performances and Discussions

Performances of the frequently cited approaches on 5 benchmarking datasets are collected and listed in the following Table 2. All listed approaches are implemented on SHT-A, and performances on other main-

Table 2
Performance list on benchmarking datasets

Dataset		SHT A		SHT B		UCF-QNRF		JHU+		NWPU		UCF_CC_50	
Approach	Year	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE
Crowd-CNN [55]	2015	181.8	277.7	32	49.8	181.8	277.7	/	/	/	/	467	498.5
MCNN [56]	2016	110.2	173.2	26.4	41.3	227	426	188.9	483.4	232.5	714.6	377.6	509.1
CP-CNN [38]	2017	73.6	106.4	20.1	30.1	/	/	/	/	/	/	295.8	320.9
CMTL2017 [37]	2017	101.3	152.4	20	31.1	252	514	/	/	/	/	322.8	341.4
Switch-CNN [2]	2017	90.4	135	21.6	33.4	228	445	/	/	/	/	318.1	439.2
CSRNet [15]	2018	68.2	115	10.6	16	120.3	208.5	85.9	309.2	121.3	387.8	266.1	397.5
SANet [3]	2018	67	104.5	8.4	13.6	/	/	91.1	320.4	190.6	491.4	258.4	334.9
ACSCP [34]	2018	75.7	102.7	17.2	27.4	/	/	/	/	/	/	291	404.6
D-CONVNET [35]	2018	73.5	112.3	18.7	26	/	/	/	/	/	/	288.4	404.7
IG-CNN [33]	2018	72.5	118.2	13.6	21.1	/	/	/	/	/	/	291.4	349.4
IC-CNN [32]	2018	68.5	116.2	10.7	16	/	/	/	/	/	/	260.9	365.5
CA-Net [21]	2019	61.3	100	/	/	107	183	100.1	314	/	/	/	/
CAN [21]	2019	62.3	100	7.8	12.2	107	183	100.1	314	106.3	386.5	212.2	243.7
HA-CNN [39]	2019	62.9	94.9	/	/	118.1	180.4	/	/	/	/	256.2	348.4
SFCN [22]	2019	64.8	107.5	7.4	11.8	102	171.4	77.5	297.6	105.7	424.1	214.2	318.2
BL [26]	2019	62.8	101.8	7.7	12.7	88.7	154.8	75	299.9	105.4	454.2	229.3	308.2
CG-DRCN-CC [40]	2020	60.2	94	/	/	95.5	164.3	71	278.6	/	/	/	/
DPN-IPSM [27]	2020	58.1	91.7	/	/	84.7	147.2	/	/	/	/	/	/
DM-Count [47]	2020	59.7	95.7	7.4	11.8	85.6	148.3	68.4	283.3	88.4	388.6	211	291.5
ASNet [14]	2020	57.78	90.13	/	/	91.59	159.71	/	/	/	/	174.84	251.63
UOT [28]	2021	58.1	95.9	6.5	10.2	83.3	142.3	60.5	252.7	87.8	387.5	/	/
L2S [53]	2022	65.8	112.1	8.6	13.9	104.4	174.2	85.6	356.1	97.3	571.2	/	/
S3 [17]	2021	57	96	6.3	10.6	80.6	139.8	59.4	244	83.5	346.9	/	/
Semi [29]	2021	68.5	121.9	14.1	20.6	130.3	226.3	80.7	290.8	111.7	443.2	/	/
P2Pnet [41]	2021	52.7	85.1	6.3	9.9	85.3	154.5	/	/	77.4	362	172.7	256.2
GL [46]	2021	61.3	95.4	7.3	11.7	84.3	147.5	59.9	259.5	79.3	346.1	211	291.5
CCTrans [44]	2021	64.4	95.4	7	11.5	92.1	158.9	/	/	/	/	245	343.6
SASNet [42]	2021	53.59	88.38	6.35	9.9	85.2	147.3	/	/	/	/	161.4	234.46
MAN [18]	2022	56.8	90.3	/	/	77.3	131.5	53.4	209.9	76.5	323	/	/
ChfL [36]	2022	57.5	94.3	6.9	11	80.3	137.6	57	235.7	76.8	343	/	/
TransCrowd [16]	2022	66.1	105.1	9.3	16.1	97.2	168.5	56.8	193.6	88.4	400.5	272.2	395.3

Figure 8
Performance trend on selected approach



stream datasets are partially missed. Thus, MAE and MSE of different approaches on SHT-A will be referred as primary factor for analysis. Furthermore, Figure 8 gives a perceptual illustration to exhibit the trending of performance development.

For most approaches proposed in recent years, values of MAE and MSE on SHT-A are limited under 60 and 100. The state-of-the-art approaches are the P2Pnet and SASNet proposed in 2021. The Characteristic function Loss (ChfL) [36] proposed in 2022 has the best performance among Bayesian loss related approaches, which regresses features into points instead of density map. Generally, approaches proposed at 2021 outperform those proposed at 2022. The reason is newest approaches focused more on the exploration of novel network structure, such as adapting transformer and Bayesian loss. Despite current performance is lower, they have exhibited significant advantages and potential which could outperform the main-stream density map-based approaches in the future.

Since the P2Pnet and SASNet achieved the highest performance so far, it is necessary to take a closer inspection of their innovations..

P2Pnet: The strategy of P2Pnet is a pseudo version of point-regression approaches. Instead of calculating the confidence score of every pixel from the feature map as pedestrian's head, the P2Pnet divides the extracted feature map with a grid of the stride length *s*. Each cell of the grid is assigned as a potential head candidate/proposal with the confidence score, and this strategy is referred as "one to one match".

The network of P2PNet has an ordinary front/back-end structure. Its front-end part is a VGG16 for feature extraction. The extracted feature map will be divided by *s* to generate proposals who will be fed to the back-end. The back-end network is comprised with a regression head and a classification head. The regression head calculates the confidence score of each proposal. The classification head determines if proposal belongs to pedestrian's head or background according to the confidence score. Finally, the Loss will be calculated with predicted pedestrians and the ground truth.

The P2PNet mentions the "one to one match" can effectively address the unbalance estimation problem who hampers the prediction accuracy. However, the adaptiveness of P2Pnet is limited in practice since the *s* must be manually set. The inaccurate selection of *s* will generate direct impact on the prediction of the densest crowd within the footage. Thus, it seems the reason of P2Pnet achieved such outstanding performance in experimental environment, is heavily related to the selection of hyperparameter *s*. If the selection of *s* is made self-adaptive, the performance of P2Pnet could be more persuasive.

SASNet: As the approach with the second-best MAE/MSE in the Table 2, the SASNet also has a standard front/back-end network structure. The VGG16 is adapted as the front-end network, whose features from strides with {1,2,4,8,16} are selected as feature maps in 5 scales. Like P2Pnet, the back-end network of SASNet is also dual-heads. The map from each scale level is fed to confidence head and regression head respectively. The regression head of SASNet predicts the density map. The confidence head generates confidence map, which indicates if the current scale level can most properly describe the actual density. The confidence and density maps in each scale can be further modeled into the final density map. Furthermore, the Pyramid Region Awareness Loss (PRAL) is devised to handle the unbalance density prediction, which is explained in previous Loss section. The final Loss is obtained with the summation of the density Loss, confidence Loss and PRAL. Overall, the SASNet applies a conventional density map regression-based strategy with a composite loss.

The ablation experiment reveals 2 crucial factors for SASNet to achieve such high performance. (1) As introduced in the paragraph above, the front-end network extracts feature maps in 5 scales. If only the average feature map is adapted for regression, the MAE on SHT-A is 57.48. This MAE is at same level as S3 [17] (57) and MAN (56.8). If the feature map with highest confidence score is adapted, the MAE will be improved to 55.71. By applying weight average on all 5 maps with confidence scores, the final feature map is generated and adapted for regression. The result shows the MAE is boosted to 54.75. This experiment proves optimizing the scale selection of feature map is still a feasible path to further improve the performance, and this optimizing path can be backtracked to the MCNN in 2017. Besides, by selecting pixels from different feature levels with minimum error,

and aggregating into a "ground truth" feature map, the ideal MAE can be reduced to 46.19, which can be considered as a short-term goal for the optimization of scale selection. (2) The SASNet adapts the Pyramid Region Awareness Loss (PRAL) to handle the unbalanced prediction by pruning the most over/under-estimated pixels out, and integrating their Euclidean Distance Loss into the final Loss. This process lifts the precision of Loss calculation for the pooled feature map to pixel level. As the ablation experiment shows, by involving PRAL into the SASNet, the MAE is further boosted from 54.75 to 53.59. Comparing to the PRAL, the unbalance density handling approaches of L2S and ASNet are coarser since they are not at pixel level.

6. Conclusion

This paper reviews most representative regression-based crowd counting techniques and inspects their innovations in network architecture, the handling of unbalanced prediction and the devising of loss function. As conclusion, some observations and possible future trends is proposed.

Regression to points instead of density map. Point-regression approaches can provide position information which conventional density map based techniques cannot. This strategy is continually explored by Bayesian loss related approaches since 2019. Despite still lessor than the state-of-the-art [41, 42], the performance of Point-regression approaches such as BL, GL and Chfl is steadily increasing. Considering the high performance of the state-of-the-art heavily relies on the installation of super parameters in feature extraction process, their practical feasibility still needs to be further evaluated. But point-regression approaches minimize this impact by focusing on the optimization of regression-head. Together with the capability of providing position information, point-regression approaches can be promising candidates for practical application.

Replacing CNN with transformer: As a former natural language processing network, transformer shows magnificent capability in computer vision. In recent years, multiple transformer-related approaches such as MAN and TransCrowd are proposed and achieved sound performance. By integrating global attention, the transformer-based approaches achieve generally high performance. However, it still cannot match the state-of-the-art. Another defect is the deep ViT network for encoding, which makes the transformer more difficult to handle the real-time task. However, the semi-supervised version of transformer, such as Semiformer [51], can help the approach to maintain acceptable accuracy with the limited training data.

Enhancements of CNN-based approach: As conventional approaches, CNN-based techniques still have performance advantage on others. To sustain the leading position, some possible optimizations are worth to be further investigated. Despite typical CNN-based approaches do not regress to points, the success of P2Pnet proved higher performance can be achieved by involving point regression strategy. Secondly, as a primitive problem, the selection of feature in proper scale can be further explored to address the perspective in footage. The SASNet exploits the weight-averaged feature map to lift the performance. Theoretically, the scale selection of feature map can be modelled at pixel-level to obtain the best result. Finally, as the issue of unbalanced prediction is handled at pixel-level in the most recent work, techniques can be continuously probed to further improve accuracy.

Funding

This paper is funded by Key Research and Development Program of Shaanxi, grant number 2023-YBGY-026; National Science Foundation of China, grant number 62071378; Key Projects of Postgraduate Joint Cultivation Workstation of Xi'an University of Posts and Telecommunications, grant number YJGJ201902.

Data sharing agreement

The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Conflict of Interest

The authors have no conflicts of interest to declare.

References

1. Aggrey, O., Pius, A., Evarist, N. Towards People Crowd Detection using Wireless Sensor Networks. European Journal of Technology, 2022, 6, 32-48. https://doi.org/10.47672/ejt.1071
2. Babu Sam, D., Surya, S., Venkatesh Babu, R. Switching Convolutional Neural Network for Crowd Counting. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Hawaii, USA, July 21-26, 2017, 5744-5752.
3. Cao, X., Wang, Z., Zhao, Y., et al. Scale Aggregation Network for Accurate and Efficient Crowd Counting. Proceedings of the European Conference on Computer Vision (ECCV), 2018, 734-750. https://doi.org/10.1007/978-3-030-01228-1_45
4. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S. End-to-end Object Detection with Transformers. Proceedings of the 16th European Conference on Computer Vision (ECCV), August 23-28, 2020, 213-229. https://doi.org/10.1007/978-3-030-58452-8_13
5. Chu, X., Tian, Z., Wang, Y., Zhang, B., Ren, H., Wei, X. et al. Twins: Revisiting the Design of Spatial Attention in Vision Transformers. Advances in Neural Information Processing Systems, 2021, 34, 9355-9366.
6. Devlin, J., Chang, M. W., Lee, K., Toutanova, K. Bert: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of NAACL-HLT, Minneapolis, June 2-7, 2019, 4171-4186
7. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, 2021.
8. Guerrero-Gómez-Olmedo, R., Torre-Jiménez, B., López-Sastre, R., Maldonado-Bascón, S., Onoro-Rubio, D. Extremely Overlapping Vehicle Counting. Pattern Recognition and Image Analysis: 7th Iberian Conference, (IbPRIA), Santiago de Compostela, Spain, June 17-19, 2015, 423-431. https://doi.org/10.1007/978-3-319-19390-8_48
9. Hao, Y., Marples, D., Liu, Y., Xu, Z. Unsupervised Pedestrian Sample Extraction for Model Training. Proceedings of the 13th Multi-Conference on Computer Graphics, Visualization, Computer Vision, and Image Processing, Porto, Portugal, July 16-18, 2019, 213-229.
10. Hendrycks, D., Gimpel, K. Bridging Nonlinearities and Stochastic Regularizers with Gaussian Error Linear Units, 2016.
11. Hsu, W. Y., Lin, W. Y. Ratio-and-scale-aware YOLO for Pedestrian Detection. IEEE Transactions on Image Processing, 2020, 30, 934-947. https://doi.org/10.1109/TIP.2020.3039574
12. Idrees, H., Saleemi, I., Seibert, C., Shah, M. Multi-source Multi-scale Counting in Extremely Dense Crowd Images. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Portland, USA, June 23-27, 2013, 2547-2554. https://doi.org/10.1109/CVPR.2013.329
13. Idrees, H., Tayyab, M., Athrey, K., Zhang, D., Al-Maadeed, S., Rajpoot, N., Shah, M. Composition Loss for Counting, Density Map Estimation and Localization in Dense Crowds. Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, September 8-14, 2018, 532-546. https://doi.org/10.1007/978-3-030-01216-8_33
14. Jiang, X., Zhang, L., Xu, M., et al. Attention Scaling for Crowd Counting. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, USA, June 14-19, 2020, 4706-4715. https://doi.org/10.1109/CVPR42600.2020.00476
15. Li, Y., Zhang, X., Chen, D. CSRNet: Dilated Convolutional Neural Networks for Understanding the Highly Congested Scenes. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, USA, June 18-22, 2018, 1091-1100. https://doi.org/10.1109/CVPR.2018.00120
16. Liang, D., Chen, X., Xu, W., Zhou, Y., Bai, X. Transcrowd: Weakly-Supervised Crowd Counting with Transformers. Science China Information Sciences, 2022, 65(6), 1-14. https://doi.org/10.1007/s11432-021-3445-y
17. Lin, H., Hong, X., Ma, Z., Wei, X., Qiu, Y., Wang, Y., Gong, Y. Direct Measure Matching for Crowd Counting. In IJCAI, 2021. https://doi.org/10.24963/ijcai.2021/116
18. Lin, H., Ma, Z., Ji, R., Wang, Y., Hong, X. Boosting Crowd Counting via Multifaceted Attention. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, USA, June 19-24, 2022, 19628-19637. https://doi.org/10.1109/CVPR52688.2022.01901
19. Liu, J., Li, H., Kong, W. Multi-level Learning Counting via Pyramid Vision Transformer and CNN. Engineering Applications of Artificial Intelligence, 2023, 123, 106184. https://doi.org/10.1016/j.engappai.2023.106184

20. Liu, L., Chen, J., Wu, H., Li, G., Li, C., Lin, L. Cross-modal Collaborative Representation Learning and a Large-scale RGBT Benchmark for Crowd Counting. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 19-25, 2021, 4823-4833. https://doi.org/10.1109/CVPR46437.2021.00479

21. Liu, W., Salzmann, M., Fua, P. Context-aware Crowd Counting. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, USA, June 16-20, 2019, 5099-5108. https://doi.org/10.1109/CVPR.2019.00524

22. Liu, W., Salzmann, M., Fua, P. Context-aware Crowd Counting. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, USA, June 16-20, 2019, 5099-5108. https://doi.org/10.1109/CVPR.2019.00524

23. Liu, Y., Cao, G., Shi, B., Hu, Y. CCANet: A Collaborative Cross-modal Attention Network for RGB-D Crowd Counting. IEEE Transactions on Multimedia, 2023. https://doi.org/10.1109/TMM.2023.3262978

24. Long, J., Shelhamer, E., Darrell, T. Fully Convolutional Networks for Semantic Segmentation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, USA, June 7-12, 2015, 3431-3440. https://doi.org/10.1109/CVPR.2015.7298965

25. Ma, Z., Hong, X., Wei, X., Qiu, Y., Gong, Y. Towards a Universal Model for Cross-dataset Crowd Counting. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, Canada, October 10-17, 2021, 3205-3214. https://doi.org/10.1109/ICCV48922.2021.00319

26. Ma, Z., Wei, X., Hong, X., Gong, Y. Bayesian Loss for Crowd Count Estimation with Point Supervision. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea (South), October 27-November 2, 2019, 6142-6151. https://doi.org/10.1109/ICCV.2019.00624

27. Ma, Z., Wei, X., Hong, X., Gong, Y. Learning Scales from Points: A Scale-aware Probabilistic Model for Crowd Counting. Proceedings of the 28th ACM International Conference on Multimedia, 2023, 220-228.

28. Ma, Z., Wei, X., Hong, X., Lin, H., Qiu, Y., Gong, Y. Learning to Count via Unbalanced Optimal Transport. Proceedings of the AAAI Conference on Artificial Intelligence, Vancouver, Canada, February 2-9, 2023, 2319-2327. https://doi.org/10.1609/aaai.v35i3.16332

29. Meng, Y., Zhang, H., Zhao, Y., Yang, X., Qian, X., Huang, X., Zheng, Y. Spatial Uncertainty-aware Semi-supervised Crowd Counting. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, Canada, October 10-17, 2021, 15549-15559. https://doi.org/10.1109/ICCV48922.2021.01526

30. Miao, Z., Zhang, Y., Peng, Y., Peng, H., Yin, B. DTCC: Multi-level Dilated Convolution with Transformer for Weakly-Supervised Crowd Counting. Computational Visual Media, 2023, 1-15. https://doi.org/10.1007/s41095-022-0313-5

31. Parmar, N., Vaswani, A., Uszkoreit, J., et al. Image Transformer. Proceedings on International Conference on Machine Learning (PMLR), Stockholm, Sweden, July 10-15, 2018, 4055-4064.

32. Ranjan, V., Le, H., Hoai, M. Iterative Crowd Counting. Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, September 8-14, 2018, 270-285.

33. Sam, D. B., Sajjan, N. N., Babu, R. V., Srinivasan, M. Divide and Grow: Capturing Huge Diversity in Crowd Images with Incrementally Growing CNN. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, USA, June 18-22, 2018, 3618-3626. https://doi.org/10.1109/CVPR.2018.00381

34. Shen, Z., Xu, Y., Ni, B., Wang, M., Hu, J., Yang, X. Crowd Counting via Adversarial Cross-scale Consistency Pursuit. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, USA, June 18-22, 2018, 5245-5254. https://doi.org/10.1109/CVPR.2018.00550

35. Shi, Z., Zhang, L., Liu, Y., Cao, X., Ye, Y., Cheng, M. M., Zheng, G. Crowd Counting with Deep Negative Correlation Learning. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, USA, June 18-22, 2018, 5382-539. https://doi.org/10.1109/CVPR.2018.00564

36. Shu, W., Wan, J., Tan, K. C., Kwong, S., Chan, A. B. Crowd Counting in the Frequency Domain. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, USA, June 19-24, 2022, 19618-19627. https://doi.org/10.1109/CVPR52688.2022.01900

37. Sindagi, V. A., Patel, V. M. CNN-based Cascaded Multi-task Learning of High-level Prior and Density Estimation for Crowd Counting. Proceedings 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), Lecce, Italy, Aug 29, 2017 - Sep 1, 2017, 1-6. https://doi.org/10.1109/AVSS.2017.8078491

38. Sindagi, V. A., Patel, V. M. Generating High-quality Crowd Density Maps using Contextual Pyramid CNNs. Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, October 22-29, 2017, 1861-1870. https://doi.org/10.1109/ICCV.2017.206

39. Sindagi, V. A., Patel, V. M. Ha-ccn: Hierarchical Attention-based Crowd Counting Network. IEEE Transactions on Image Processing, 2019, 29, 323-335. https://doi.org/10.1109/TIP.2019.2928634

40. Sindagi, V. A., Yasarla, R., Patel, V. M. Jhu-crowd++: Large-scale Crowd Counting Dataset and a Benchmark Method. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 44(5), 2594-2609. https://doi.org/10.1109/TPAMI.2020.3035969

41. Song, Q., Wang, C., Jiang, Z., et al. Rethinking Counting and Localization in Crowds: A Purely Point-based Framework. Proceedings of the International Conference on Computer Vision (ICCV), Montreal, Canada, October 10-17, 2021, 3365-3374. https://doi.org/10.1109/ICCV48922.2021.00335

42. Song, Q., Wang, C., Wang, Y., et al. To Choose or to Fuse? Scale Selection for Crowd Counting. Proceedings of the AAAI Conference on Artificial Intelligence, Vancouver, Canada, February 2-9, 2021, 2576-2583. https://doi.org/10.1609/aaai.v35i3.16360

43. Tarvainen, A., Valpola, H. Mean Teachers are Better Role Models: Weight-averaged Consistency Targets Improve Semi-supervised Deep Learning Results. Processing in Neural Information Processing Systems (NIPS), 2017.

44. Tian, Y., Chu, X., Wang, H. Cctrans: Simplifying and Improving Crowd Counting with Transformer. 2021.

45. Vaswani, A., Shazeer, N., Parmar, N., et al. Attention is All You Need. Proceedings in Neural Information Processing Systems (NIPS), 2017.

46. Wan, J., Liu, Z., Chan, A. B. A Generalized Loss Function for Crowd Counting and Localization. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 19-25, 2021, 1974-1983. https://doi.org/10.1109/CVPR46437.2021.00201

47. Wang, B., Liu, H., Samaras, D., Nguyen, M. H. Distribution Matching for Crowd Counting. Advances in Neural Information Processing Systems, 2020, 33, 1595-1607.

48. Wang, Q., Gao, J., Lin, W., Yuan, Y. Learning From Synthetic Data for Crowd Counting in the Wild. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, USA, June 16-20, 2023, 8198-8207.

49. Wang, Q., Gao, J., Lin, W., Li, X. NWPU-crowd: A large-scale Benchmark for Crowd Counting and Localization. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 43(6), 2141-2149. https://doi.org/10.1109/TPAMI.2020.3013269

50. Wang, Q., Gao, J., Lin, W., Li, X. NWPU-crowd: A Large-scale Benchmark for Crowd Counting and Localization. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 43(6), 2141-2149. https://doi.org/10.1109/TPAMI.2020.3013269

51. Weng, Z., Yang, X., Li, A., Wu, Z., Jiang, Y. G. Semi-supervised Vision Transformers. Proceedings 17th European Conference on Computer Vision (ECCV), Tel-Aviv, Israel, October 23-27, 2022, 605-620. https://doi.org/10.1007/978-3-031-20056-4_35

52. Xie, J., Xu, W., Liang, D., et al. Super-Resolution Information Enhancement for Crowd Counting. arXiv preprint arXiv:2303.06925, 2023. https://doi.org/10.1109/ICASSP49357.2023.10097102

53. Xu, C., Liang, D., Xu, Y., Bai, S., Zhan, W., Bai, X., Tomizuka, M. Autoscale: Learning to Scale for Crowd Counting. International Journal of Computer Vision, 2022, 130(2), 405-434. https://doi.org/10.1007/s11263-021-01542-z

54. Xu, T., Chen, X., Wei, G., Wang, W. Crowd Counting using Accumulated HOG. Proceedings of the 12th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD), Changsha China, August 13-15, 2016, 1877-1881. https://doi.org/10.1109/FSKD.2016.7603465

55. Zhang, C., Li, H., Wang, X., Yang, X. Cross-scene Crowd Counting via Deep Convolutional Neural Networks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, USA, June 7-12, 2015, 833-841.

56. Zhang, Y., Zhou, D., Chen, S., Gao, S., Ma, Y. Single-image Crowd Counting via Multi-column Convolutional Neural Network. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, USA, June 1-26, 2016, 589-597. https://doi.org/10.1109/CVPR.2016.70

57. Zhao, M., Zhang, J., Zhang, C., Zhang, W. Leveraging Heterogeneous Auxiliary Tasks to Assist Crowd Counting. Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, USA, June 16-20, 2019, 12736-12745. https://doi.org/10.1109/CVPR.2019.01302

58. Zheng, S., Lu, J., Zhao, H., et al. Rethinking Semantic Segmentation from a Sequence-to-sequence Perspective with Transformers. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 19-25, 2021, 6881-6890. https://doi.org/10.1109/CVPR46437.2021.00681

59. Zhu, P., Li, J., Cao, B., Hu, Q. Multi-Task Credible Pseudo-Label Learning for Semi-Supervised Crowd Counting. IEEE Transactions on Neural Networks and Learning Systems, 2023. https://doi.org/10.1109/TNNLS.2023.3241211