Neural Computing and Applications
https://doi.org/10.1007/s00521-022-07917-8

ORIGINAL ARTICLE

Exploring density rectification and domain adaption method for crowd counting

Sifan Peng[1] · Baoqun Yin[1] · Qianqian Yang[1] · Qing He[1] · Luyang Wang[1]

Received: 22 February 2022 / Accepted: 30 September 2022

Abstract

Crowd counting has received increasing attention due to its important roles in multiple fields, such as social security, commercial applications, epidemic prevention and control. To this end, we explore two critical issues that seriously affect the performance of crowd counting including nonuniform crowd density distribution and cross-domain problems. Aiming at the nonuniform crowd density distribution issue, we propose a density rectifying network (DRNet) that consists of several dual-layer pyramid fusion modules (DPFM) and a density rectification map (DRmap) auxiliary learning module. The proposed DPFM is embedded into DRNet to integrate multi-scale crowd density features through dual-layer pyramid fusion. The devised DRmap auxiliary learning module further rectifies the incorrect crowd density estimation by adaptively weighting the initial crowd density maps. With respect to the cross-domain issue, we develop a domain adaptation method of randomly cutting mixed dual-domain images, which learns domain-invariance features and decreases the domain gap between the source domain and the target domain from global and local perspectives. Experimental results indicate that the devised DRNet achieves the best mean absolute error (MAE) and competitive mean squared error (MSE) compared with other excellent methods on four benchmark datasets. Additionally, a series of cross-domain experiments are conducted to demonstrate the effectiveness of the proposed domain adaption method. Significantly, when the A and B parts of the Shanghaitech dataset are the source domain and target domain respectively, the proposed domain adaption method decreases the MAE of DRNet by 47.6%.

Keywords Crowd counting · DRmap auxiliary learning · Density rectifying · Domain adaption

1 Introduction

In the past few years, an increasing number of researchers have devoted themselves to crowd counting fields. Crowd counting aims to estimate the number of crowds and the crowd density value of arbitrary pixel position for the input images. Due to the higher adaptability and practicability of the crowd counting methods, they are widely applied in various scenarios including social security, commercial programme and so on. For instance, the spread of COVID-19 caused by large-scale crowd concentration can be avoided through real-time monitoring crowds. Furthermore, crowd counting methods can be used to manage the layout of shopping malls by analyzing the spatial distribution of crowds. The requirements in the above-mentioned application scenarios have promoted the further development of the crowd counting technologies.

✉ Baoqun Yin
bqyin@ustc.edu.cn

Sifan Peng
sifan@mail.ustc.edu.cn

Qianqian Yang
yanghcqq@mail.ustc.edu.cn

Qing He
heqing2020@mail.ustc.edu.cn

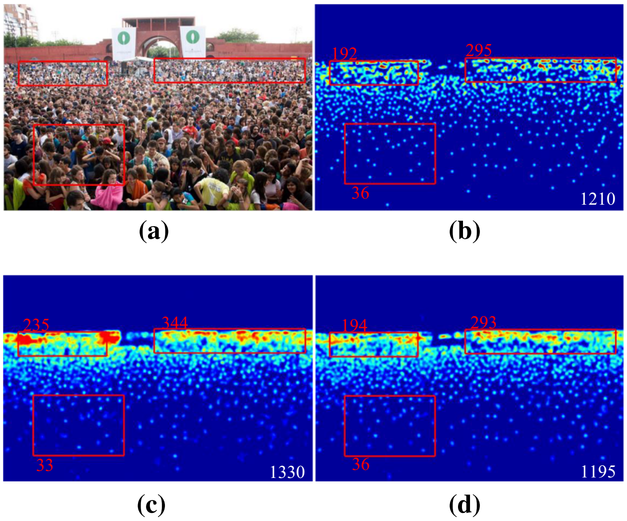
Luyang Wang
ly1105@mail.ustc.edu.cn

[1] Department of Automation, University of Science and Technology of China, Huangshan Road,
Hefei 230027, Anhui, China

Published online: 14 October 2022

Recently, with the prevalence of convolutional neural networks (CNNs) [1], large quantities of CNN-based approaches [2–9] have been proposed to address various problems in crowd counting fields. Though these methods improve the performance of crowd counting tasks to some extent, several inherent challenges remain, such as nonuniform density distribution and domain adaption. The former seriously affects the counting accuracy of the crowd counting networks in regions of various density levels, while the latter reflects the performance of the trained counting models in actual application scenarios. Therefore, we are deeply involved with the above two issues in this paper.

As for the nonuniform crowd density distribution problem, it is known that the larger the image depth is, the fewer pixels are occupied by the crowds, resulting in a high-density crowd distribution, and vice versa. Figure 1 shows the crowd density maps predicted by different networks in different density areas, which reveals that the baseline network used in this paper has relatively large estimation errors in both high-density and low-density areas. To solve nonuniform density distribution problem, Liu et al. [10] apply a FasterCNN-based detection network to detect crowds in sparse scenes while adopting a regression network to estimate crowd density in dense scenes, and adaptively weight the outputs of the detection network and regression network. Gao et al. [11] design a spatial-level attention module to perceive crowd density changes in input images by extracting global context information. These works alleviate the above problems to a certain extent from the perspective of network model design, but ignore the important fact that image depth is the direct factor resulting in nonuniform crowd density distribution problem.



(a)

(b)

(c)

(d)

Fig. 1 **a** and **b** represent the input crowd image and the ground truth density map. **c** and **d** indicate the density map output from the baseline(without DRmap and DPFM) and DRNet, respectively. The red boxes denote the crowd regions in different density levels. The number marked in red reveals the crowd number in the red box, and the white number presents the overall crowd count of the entire image

Domain adaption is the second issue that we concentrate on, which affects the generalization performance of the counting model in unseen scenarios. To our knowledge, there are many differences among different datasets including crowd density distributions, background types, and scene styles, which are collectively referred to as domain gaps. Due to the existence of the domain gap, the model well-trained in the source domain behaves badly in the target domain. Wang et al. [2] firstly propose a synthetic crowd dataset, and transform the synthetic images to real-world images by SE Cycle GAN. Gao et al. [12] employ adversarial learning to discriminate the origin (source domain or target domain) of the feature maps in the network, and reduce the domain gap of the feature space. These methods reduce the domain gap from a global perspective. However, they neglect to learn domain invariant features from a local perspective, such as learning different feature distributions in an image. To date, no one has addressed the cross-domain problem of crowd counting from both global and local perspectives.

To cope with the nonuniform crowd density distribution issue, we present a density rectification network (DRNet) as shown in Fig. 2. Different from previous methods [10, 11] that focus on the design of the network structure, we propose a density rectification map (DRmap) and an efficient algorithm for generating the ground truth DRmap. The devised DRmap is closely related to the image depth and head spacing, and the larger the pixel value in DRmap is, the greater the density of the crowd at the corresponding location. We introduce the density rectification auxiliary task into the network to generate a DRmap for weighting the predicted crowd density map, which aims to obtain more accurate crowd density estimations in different density areas. In addition to adding auxiliary tasks to correct the density estimation deviation, we also develop a dual-layer pyramid fusion module (DPFM) from the point of view of promoting feature fusion. We embed the designed DPFM module into the network as shown in Fig. 2, and carry out a dual-layer fusion of crowd density features of different scales, which promotes DRNet to generate high-quality crowd density maps. For the purpose of reducing the domain gap between different crowd scenes, we put forward a novel domain adaption method of randomly cutting mixed dual-domain images as shown in Fig. 4. We first leverage the model well-trained on the source domain to generate pseudo-labels for the target domain training set, and mix the source and target domain training data. In a training batch, we randomly cut a part of the target domain image and paste it into the source domain image, and the

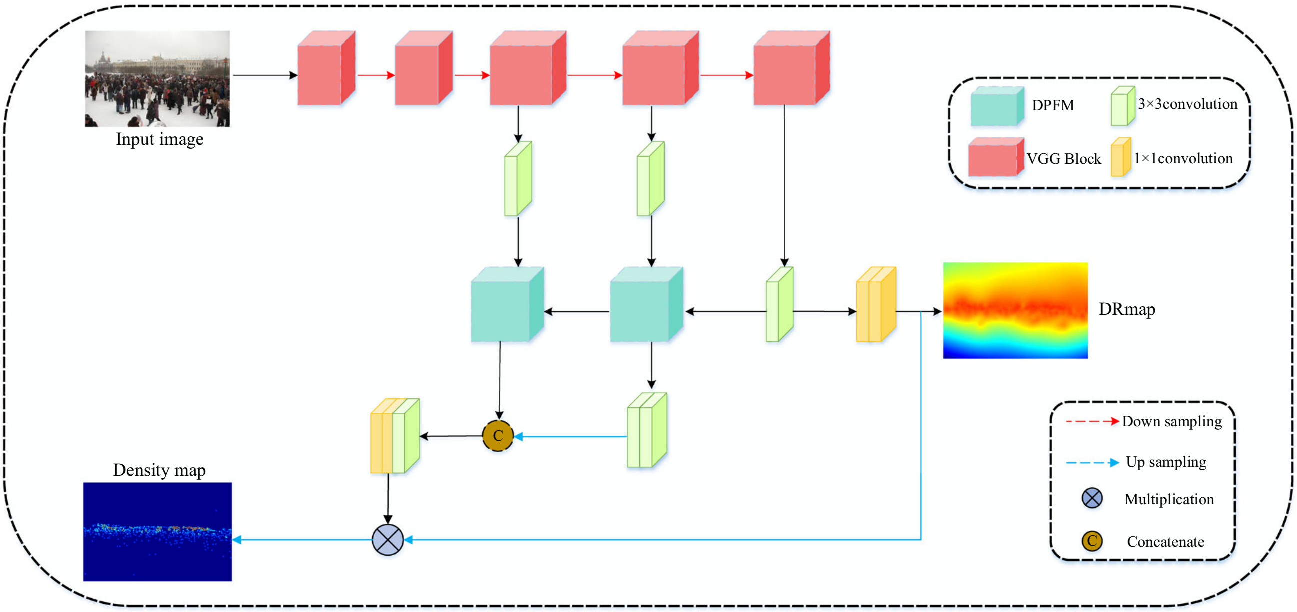


Fig. 2 The overall network architecture of the proposed DRNet. The Multiplication symbol specifically refers to pixel-by-pixel multiplication, and the Concatenate symbol denotes the fusion of feature maps in the channel dimension. For a given input image, the network outputs the corresponding DRmap and the corrected crowd density map

labels are also operated accordingly. Finally, we utilize the source domain and target domain data that are well-mixed in both a batch and a single image to train the network. The devised domain adaption method not only learns the domain invariant features between the source domain and the target domain from a global perspective, but also learns the different feature distributions of the source domain and the target domain on a single image from a local point of view. In general, the main contributions in this paper are summarized as follows:

1. We propose a density rectification map (DRmap) and an efficient algorithm to generate the ground truth DRmap. Based on the proposed DRmap, we design a DRmap auxiliary learning task to rectify the incorrect density estimations of the initial crowd density map in various density areas.
2. We develop a dual-layer pyramid fusion module (DPFM) to carry out a dual-layer fusion of crowd density features of different scales, which contributes to generating more accurate crowd density maps.
3. We put forward a domain adaption method of randomly cutting mixed dual-domain images to reduce the domain gap. Experimental results on different source domains and target domains demonstrate the effectiveness of the proposed domain adaptation method.

The remainder of this paper contains four sections. Firstly, we briefly introduce some classic counting networks and domain adaptation methods in the related work section. Then, we describe the technical details of the proposed method including the generation principle of DRmap, the structure of DPFM, and the implementation process of the proposed domain adaptation method. Moreover, we conduct several comparison experiments and ablation studies in the experiment section. Finally, we summarize this paper in the conclusion section.

2 Related work

Under the exploration of researchers, plenty of valuable crowd counting related works have emerged, which can be roughly divided into two categories, crowd counting networks and cross-domain methods. The former mainly includes network structure design, multi-task method research, multi-view fusion, drone counting and so on, which aim to improve the model performance in the test set. The latter introduces several representative domain adaptation approaches that make the network obtain better generalization ability in some unseen scenes.

2.1 Crowd counting networks

In the past five years, the design of network models has been a popular research direction in crowd counting fields, and a series of sophisticated network structures have been proposed to achieve state-of-the-art performance including single-column networks [5, 13, 14], multi-column networks [15–18] and multi-level fusion networks [19–21].

Amirgholipour et al. [13] first detect the head size and generate hyperparameters (HPs) about the head size through a fuzzy inference system. After that, they train a single-column network with HPs to adaptively generate a crowd density map. Liu et al. [22] put forward a self-supervised task through sub-image crowd ranking, and combine labeled and unlabeled data to train a VGG-based single-column counting network model. Liu et al. [14] incorporate a multi-scale contextual information extraction module into a single-column network to solve the perspective distortion problem. Wang et al. [5] present a novel single-column scale-invariance network that contains several scale-invariance transformation layers with dense connections to overcome large density shifts. Zhang et al. [15] are almost the first to construct a multi-column network to extract multi-scale crowd features. Cheng et al. [18] come up with a statistical network to minimize the mutual information of the multi-column network, which decreases the scale correlation of the acquired multi-scale crowd features. Sam et al. [19] nominate a bottom-up and top-down combination network, which adjusts the low-level features of the bottom-up network by the feedback of the top-down network to fix the dense crowd counting problem. Liu et al. [21] present a progressively refined density map network stacking multiple fully convolutional networks recursively, and leverage the output of the previous network as input. With the development of regularization technology, multi-task learning methods [23–26] are widely used in current crowd counting networks. To deal with high appearance similarity and perspective change issues, Gao et al. [23] construct a multi-task architecture named PCCNet that contains density classification, density estimation, foreground segmentation and perspective change perception. Zhao et al. [25] attempt to design a variety of auxiliary tasks to optimize the backbone network such as crowd segmentation, depth prediction and count regression, which indirectly solves scale variations, background clutter and crowd occlusions problems. Considering that the processed crowd images may come from multiple cameras with different angles, many multi-view-based networks [27, 28] have been put forward to fuse multiple input crowd images. For the input crowd images from multiple perspectives, Zhang et al. [27] propose two multi-view fusion schemes to output scene-level density maps. The first method is to extract features from multiple perspective images to generate density maps. After affine transformation, the density maps are projected to a horizontal surface, and then the transformed density maps are channel-cascaded to generate scene-level density maps. The second approach extracts features from the input multi-view images and performs affine transformation. Then the features are concatenated to generate a scene-level density map. Due to the massive deployment of vision applications on drone platforms, researchers begin to explore crowd counting networks [29, 30] based on drones. Wen et al. [30] firstly present a large-scale drone-captured dataset named DroneCrowd that contains 33,600 frames with 4.8 million annotated heads. Then, they design a space-time neighbor aware network to predict crowd density and localization.

2.2 Cross-domain approaches

The domain adaption methods are used to improve the performance of unfamiliar scenes that the model has not learned during the training process, which are widely applied in existing computer vision tasks [31–34]. Bai et al. [31] introduce an unsupervised multi-source domain adaption person re-identification method consisting of a rectification domain-specific batch normalization module and a multi-domain information fusion module, decreasing the domain gap between different source datasets. Faraki et al. [32] recommend a novel cross-domain triplet loss function to learn semantically meaningful representations to improve the performance of face recognition tasks in unknown scenes. He et al. [34] first offer an image style translation method to reduce the image gap in different domains, and then propose two collaborative learning strategies for learning domain-invariant features in semantic segmentation tasks.

Zhang et al. [35] first pay attention to cross-domain issues in the field of crowd counting, who try to fine-tune the network by selecting images of the target domain similar to the source domain. Based on the Grand Theft Auto V game, Wang et al. [2] generate a synthetic crowd dataset called GCC, and utilize SE CycleGAN to transform synthetic images into target domain style images to train the model. Taking into account that the converted images lack detailed texture in [2], Gao et al. [12] adopt adversarial learning to discriminate the origin (source domain or target domain) of the feature map of each layer in the network, and reduce the domain distance of the feature space. Hossain et al. [36] divide the counting network into an encoder and a decoder, and employ training set images to train the network. For specific application scenarios, the encoder parameters are fixed, and a target domain image is exploited to fine-tune the decoder parameters. Similar to [12], Li et al. [37] leverage the discriminator to distinguish whether the generated crowd density map comes from the source domain or the target domain to address domain adaption problem. Han et al. [38] propose a semantic consistency cross-domain method, which introduces adversarial learning to determine whether the extracted features are from the source domain or the target domain. Wang et al. [3] present a neuron linear transformation method to handle cross-domain crowd counting tasks.

Firstly, the traditional supervised learning method is used to learn the source domain model parameters, and then a small amount of labeled target data is used to learn the multiplication factor and bias of the source domain model. Finally, the parameters of these neurons in the target domain are updated through linear transformation. Based on the assumption that adjacent frames have the same crowd distribution, He et al. [39] construct a video-based unsupervised crowd counting cross-domain method by minimizing the density isomorphism reconstruction error and maximizing the estimation-reconstruction consistency between adjacent frames. As the backgrounds in different scenes vary significantly, Liu et al. [40] apply a point-derived crowd counting segmentation method to separate the crowd and the background, and design a Crowd Region Transfer module to extract domain-invariant features beyond background distractions. Inspired by the above excellent works, we devote ourselves to rectifying the incorrect density estimation and decreasing the domain gap between different crowd scenes for crowd counting.

3 Proposed method

In this paper, we focus on dealing with the nonuniform crowd density distribution problem and cross-domain crowd counting issue. In the case of nonuniform crowd density distribution problem, we bring forward a density rectifying network called DRNet as shown in Fig. 2. The first thirteen convolutional layers divided into five blocks by the pooling layer in VGG [41] serve as the backbone to extract abundant crowd features. Then, we employ the designed DPFM module as described in Fig. 3 to perform dual-layer pyramid fusion of the extracted multi-scale crowd density features, aiming to make full use of the crowd features extracted by the backbone network. Finally, the initial density map is weighted by the DRmap produced by the DRmap auxiliary learning task, which can rectify the incorrect density estimation due to nonuniform crowd distribution. Furthermore, we come up with a novel domain adaption approach to reduce the domain gap between different crowd scenes as depicted in Fig. 4, and more detailed descriptions are shown as follows.

3.1 DRmap auxiliary learning task

Nonuniform crowd density distribution is a difficult problem in the field of crowd counting. A test example of the baseline network (without DRmap and DPFM) in Fig. 1 presents that the test error in high-density and low-density areas reaches 22.4 and 8.3%. Therefore, it is necessary to correct the density map estimation of different density areas.

Inspired by the principle of camera imaging, we all know that the larger the image depth is, the smaller the head size and the closer the distance between the heads, resulting in a highly dense crowd distribution and vice versa. The image depth information needs to be annotated by a special camera, while the distance between the heads can be calculated from the existing head annotations. For the purpose of decreasing the cost of labeling, we propose DRmap based on the distance between the crowd heads to rectify the error density estimation, and the ground truth DRmap generation algorithm is as follows.

For a given image I containing N annotated heads, we calculate the distances D^i from an arbitrary head point (x_i, y_i) to the head points of all others where $D^i = \{D^i_1, D^i_2, \cdots D^i_{i-1}, D^i_{i+1}, \cdots D^i_N\}$, and sort the distance values in D^i in ascending order. The initial head spacing $P(x_i, y_i)$ in position (x_i, y_i) is defined as Eq. 1, where we set q as a constant 5 according to the ablation experiments results in Sect. 4.3.3.

$$P(x_i, y_i) = \frac{1}{q}\sum_{j=1}^{q} D^i_j \qquad (1)$$

Due to the randomness of the crowd distribution and the perspective phenomenon of the camera lens, the minority head spacing values obtained by Eq. 1 are too large or too

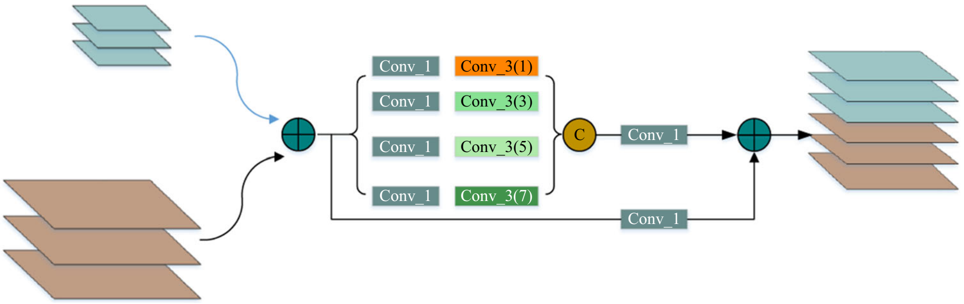


Fig. 3 The internal details presentation in the proposed DPFM module where the blue curve represents upsampling. Conv_1 means the 1×1 convolution, and Conv_3(3) denotes the 3×3 dilated convolution with a dilation rate of 3. ⊕ and © indicate pixel-wise addition and concatenation in channel dimension, respectively

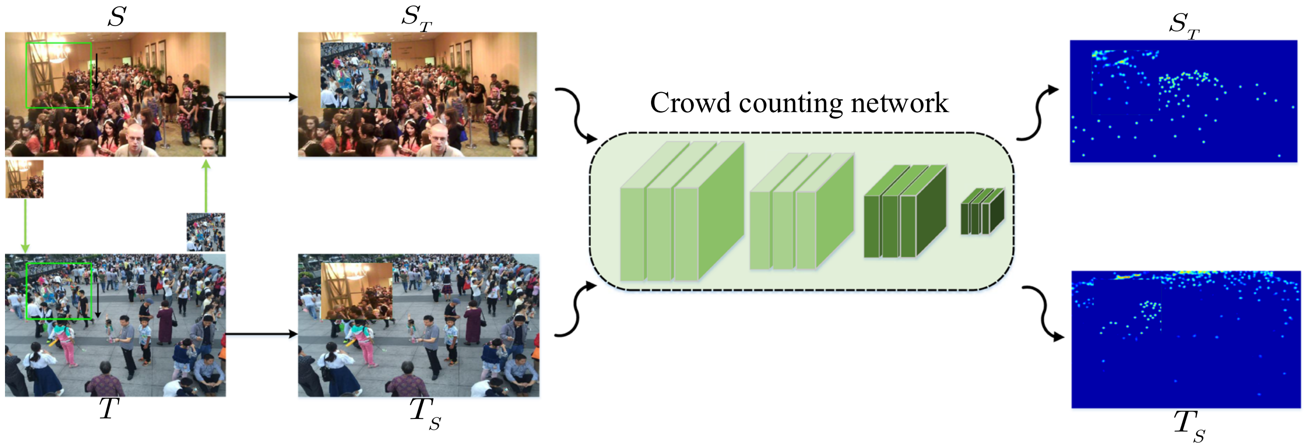


Fig. 4 The overall flowchart of the proposed cross-domain method. S and T denote source domain and target domain, respectively. The green box represents the selected mixed regions in source and target domain. S_T reveals the domain mixed by S and T as well as T_S

small. We use the 3σ criterion to eliminate outliers in P and get the normal head spacing $\hat{P}$ by Eq. 2, where Ave and σ denote the mean and variance in P.

$$\hat{P}(x_i, y_i) = \begin{cases} P(x_i, y_i) & if|P(x_i, y_i) - Ave| < 3\sigma, \\ 0 & otherwise \end{cases} \tag{2}$$

As the distribution of the labeled head points in the crowd image I is discrete, the head spacing matrix $\hat{P}$ obtained by Eqs. 1 and 2 is composed of a series of discrete points. Considering that discrete points are difficult to use for training the network, we need to make the discrete matrix $\hat{P}$ continuous. Inspired by the Biharmonic spline interpolation function introduced in [42], it can calculate the function value of the interpolation point at any position with high precision. We utilize the Biharmonic spline interpolation in two dimensions [42] to interpolate between the discrete points of the head spacing matrix $\hat{P}$, so a continuous head spacing matrix Q is acquired. The Biharmonic Green Function $\phi_2(x)$ used in the interpolation process of the Biharmonic spline interpolation function is defined as Eq. 3.

$$\phi_2(x) = |x|^2(\ln|x| - 1) \tag{3}$$

According to the above analysis that head spacing is approximately negatively correlated with crowd density, we get the DRmap R by formula (4), where K indicates the correction coefficient, and the selection of K is further discussed in the experiment section.

$$R(x_\mathrm{i}, y_\mathrm{i}) = K\left(1 - \frac{Q(x_i, y_i) - \mathrm{Min}Q}{\mathrm{Max}Q - \mathrm{Min}Q}\right) \tag{4}$$

The ground truth DRmap generated by the above algorithm is used to learn the density rectifying auxiliary task for rectifying the initial crowd density map. For the input crowd image I, the initial density map $F_1(\omega_1, I)$ and the density correction map $F_2(\omega_2, I)$ are generated through the density map generation network $F_1(\omega_1)$ and the density correction network $F_2(\omega_2)$, respectively. Then, we adopt the DRmap to rectify the initial density map and produce the refined density map by pixel-wise multiplication as shown in formula (5). $F(\omega)$ denotes the rectified density map output by DRNet, and × represents the multiplication of the corresponding elements.

$$F(\omega, I) = F_1(\omega_1, I) \times F_2(\omega_2, I) \tag{5}$$

Through pixel-by-pixel weighting, we assign greater weights to areas with higher crowd density, and lower weights to areas with lower crowd density. Therefore, the DRmap auxiliary learning task we design can rectify the wrong crowd density estimations in different crowd density areas.

3.2 Dual-layer pyramid fusion module

Density map regression is a pixel-level low-level visual task that is very sensitive to image resolution. Although the introduction of the pooling layer in the network will increase the nonlinearity of the model, it will inevitably lose a lot of detailed information. Furthermore, we know that a large number of dense crowds are concentrated in a small area of the image and contain relatively limited crowd information compared with the sparse crowds due to the nonuniform crowd density distribution. The pooling layer makes the detailed information in high-density areas further lost, resulting in erroneous crowd density estimation. Figure 1 reveals the test results output from the baseline network, where the estimation error in the high-density area is much larger than that in the low-density area. Therefore, we put forward the DPFM module depicted in Fig. 3 to remedy the impact of limited crowd information from the point of feature fusion, especially in high-density areas.

For the multi-scale crowd density feature f_1 and f_2 input into the DPFM module, we first fuse them by a feature pyramid fusion strategy. The small-size crowd density feature f_1 is up-sampled to the same size as the large-size feature map f_2 by the bilinear interpolation function B, and then the feature maps are directly fused by pixel-by-pixel addition. Different from the simple pyramid fusion in [43], for the purpose of further refining the fused multi-scale crowd density features, we design a multi-branch structure M where each branch contains dilated convolutions with different dilation rates. The receptive field structure composed of the multi-branch network is roughly distributed in a pyramid, which can finely refine the previous fused multi-scale crowd density features. Finally, we obtain the classy multi-scale crowd density features f through dual-layer pyramid fusion as shown in formula (6), which contributes to the network estimating accurate crowd density maps in different density regions.

$$f = M(f_2 + B(f_1)) \tag{6}$$

3.3 Domain adaption method

The methods introduced in Sects. 3.1 and 3.2 enable the crowd counting network to achieve lower counting errors on the training set and test set with similar feature distributions. However, the well-trained crowd counting model ultimately needs to be implemented and deployed to actual application scenarios. Due to the existence of the domain gap between the real application scenario (target domain) and the existing dataset (source domain), the generalization ability of the model is significantly reduced in unseen scenarios. To achieve better performance in the target domain, relabeling the target domain dataset to retrain the model is an intuitive method. Many practical factors make the above idea burdensome to realize, such as the polytropy of the application scenarios and the expensive data labeling costs. Inspired by the Cutmix algorithm [44] for object detection and classification tasks, which improves the performance of the model on the source domain through regional dropout operations, we propose a domain adaption method named randomly cutting mixed dual-domain images as shown in Fig. 4.

We define the source domain and the target domain as S and T, respectively. Training set images in S are defined as $S_I = \{S_i^I\}_{i=1}^N$ containing N annotated images, and the corresponding ground truth density maps set is defined as $S_D = \{S_i^D\}_{i=1}^N$. Training set images in T without labels are defined as $T_I = \{T_j^I\}_{j=1}^M$. For a given arbitrary initial counting network $\mathcal{F}$, we adopt source domain data S_I and S_D to optimize the model parameter ω and obtain an optimal model $\mathcal{F}(\omega)$. Based on the well-trained model on S, we generate the pseudo-truth density maps $T_D = \{T_j^D\}_{j=1}^M$ corresponding to the training set images T_I in T as shown in formula (7).

$$T_D = \mathcal{F}(\omega, T_I) \tag{7}$$

We mix all of source domain images S_I and target domain images T_I with pseudo-labels to train the network, which can extract domain invariant features from a global perspective. However, the domain invariant features learned from this global perspective are relatively rough, as there is a large domain gap in a single image between different domains. Therefore, we leverage the following method to further learn domain-invariance features from a local point of view. Assume that a training batch contains a source domain image S_i^I with ground truth S_i^D and a target domain image T_j^I with pseudo-truth label T_j^D. We randomly cut a subregion S_i^{I-C} and T_j^{I-C} in the upper left corner of S_i^I and T_j^I, the size of which is not more than half of the original image. The corresponding label is also cut at the same position in S_i^D and T_j^D to get subregion S_i^{D-C} and T_j^{D-C}. The sub-regions T_j^{I-C} and T_j^{D-C} cut from T are pasted to the corresponding positions of S_i^I and S_i^D. Then, we obtain S_T^I and S_T^D mixed with T in a single image as shown in Eq. 8.

$$\begin{cases} S_T^I = S_i^I - S_i^{I-C} + T_j^{I-C} \\ S_T^D = S_i^D - S_i^{D-C} + T_j^{D-C} \end{cases} \tag{8}$$

Similar to the above operation, we paste the sub-regions S_i^{I-C} and S_i^{D-C} obtained from S to the corresponding positions of T_j^I and T_j^D respectively. Then, we get the target domain data T_S^I and T_S^D mixed with S in an image as depicted in formula (9).

$$\begin{cases} T_S^I = T_j^I - T_j^{I-C} + S_i^{I-C} \\ T_S^D = T_j^D - T_j^{D-C} + S_i^{D-C} \end{cases} \tag{9}$$

The mixed S_T^I and T_S^I enable the network to learn different feature distributions from multiple domains in a single image, which make the model extract domain-invariance crowd features from a local perspective. In general, our cross-domain method trains the network to learn domain-invariant features by mixing source domain and target domain data at the dataset level and the single image level to reduce the domain gap.

3.4 Training details

We train the proposed DRNet with a multi-task learning strategy, and the detailed descriptions of training details such as ground truth density map generation, data augmentation methods, and loss functions are as follows.

3.4.1 Ground truth

As crowd head label is a series of isolated coordinates and contains limited information, the network trained with point coordinates is difficult to converge. Therefore, we generate a ground truth density map C^{GT} containing more crowd density information based on the coordinates of the head points $\{p_i\}_{i=1}^{n}$ as shown in formula (10).

$$C^{GT}(p) = \sum_{i=1}^{n} \delta(p - p_i) * G_{\sigma}(p) \tag{10}$$

The delta function $\delta(p - p_i)$ denotes a head in position p_i, while G_σ represents a Gaussian kernel function with variance σ. We obtain a continuous ground truth density map C^{GT} through the convolution operation $*$ between the Gaussian kernel function and the crowd head coordinate points.

3.4.2 Data augmentation

We firstly adopt a sliding window method to crop nine patches of 1/4 size of the original image to expand the dataset. In addition, we further increase the diversity of the image by adding Gaussian noise to the image, randomly adjusting the order of the three channels of the RGB image, gamma correction and randomly converting the RGB image to the gray image.

3.4.3 Loss functions

Both the density map regression task and the DRmap auxiliary regression task utilize the Euclidean distance loss function to train the network, which are respectively defined as L_{den} and L_{dep} as shown in the formula (11) and (12), where N represents the number of pixels of the input image I.

$$L_{den} = \frac{1}{N} \sum_{i=1}^{N} \left(F(\omega, I)_i - {C_i}^{GT}\right)^2 \tag{11}$$

$$L_{dep} = \frac{1}{N} \sum_{i=1}^{N} \left(F_2(\omega_2, I)_i - R_i\right)^2 \tag{12}$$

In addition, for the purpose of suppressing background interference, we use the head edge map proposed by Peng et al. [45] to supervise the network to generate discriminative crowd features. The head edge loss L_e is defined as (13),

$$\begin{aligned} L_e = & \frac{1}{N} \sum_{i=1}^{N} -(E_i \log(F_3(\omega_3, I)_i) \\ & + (1 - E_i) \log(1 - F_3(\omega_3, I)_i)) \end{aligned} \tag{13}$$

where E and $F_3(\omega_3, I)$ represent ground truth head edges and predicted head edges, respectively. Considering that Euclidean loss may cause the density map to be blurred, we also use Structural Similarity (SSIM) loss L_s to train the network from the perspective of luminance, contrast and structure, which is defined as Eq. 14.

$$L_s = 1 - \left(\frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \right) \tag{14}$$

μ_x and μ_y indicate the mean of the ground truth density map C^{GT} and predicted density map $F(\omega, I)$, while σ_x and σ_y denote the corresponding variance. C_1 and C_2 are both close to 0, preventing the denominator from being 0. Finally, the overall loss function L of the network is the linear weighted sum of the above loss as shown in (15).

$$L = L_{den} + \lambda_1 L_{dep} + \lambda_2 L_e + \lambda_3 L_s \tag{15}$$

Based on the principle that each loss value is in the same order of magnitude, we set λ_1, λ_2 and λ_3 to 1, 0.1 and 0.0001.

4 Experiments

In the experiment section, we first train and test the proposed DRNet on multiple datasets such as Shanghaitech [15], UCF-QNRF [46], JHU-CROWD++ [47] and NWPU-Crowd [48], and compare with other state-of-the-art algorithms to present the superiority of the designed DRNet. Then, we verify the effectiveness of each component in DRNet in the ablation experiment section, and select the most appropriate density correction coefficient K and hyperparameter q. Finally, we demonstrate the effectiveness and universality of our domain adaption method on different source and target domains.

4.1 Evaluation metric

All of the experiments are based on the PyTorch deep learning framework, and NVIDIA GeForce RTX 2080 Ti is used for model training acceleration. We adopt the mean absolute error (MAE) and mean squared error (MSE) to evaluate the performance of the proposed algorithm, which are defined as formula (16) and (17).

$$MAE = \frac{1}{N_I}\sum_{i=1}^{N_I}|Y_i - \hat{Y}_i| \quad (16)$$

$$MSE = \sqrt{\frac{1}{N_I}\sum_{i=1}^{N_I}|Y_i - \hat{Y}_i|^{2}} \quad (17)$$

N_I represents the number of the test images. Y_i and $\hat{Y}_i$ indicate the ground truth crowd numbers and the corresponding estimated crowd numbers, respectively. MAE measures the accuracy of the algorithm, while MSE reveals the robustness of the algorithm. The smaller the MAE value and the MSE value, the better the algorithm.

4.2 Experimental results on different datasets

We compare the performance of the proposed DRNet with other advanced algorithms in Shanghaitech, UCF-QNRF, JHU-CROWD++ and NWPU-Crowd datasets. Furthermore, we analyze the reasons that our method achieves superior results.

4.2.1 Results on Shanghaitech dataset

The Shanghaitech dataset is proposed by Zhang et al. [15], which contains 1198 images and 330,165 labeled crowd heads. The dataset is divided into two parts named Part_A (SHA) and Part_B (SHB). All of the images on SHA are randomly picked from the Internet, of which 300 are utilized for training and 182 are used for testing. The image resolution on SHA is diverse and the crowd count in a single image ranges from 482 to 3139. The average crowd number on SHA is about 501. Compared with other excellent algorithms, the proposed DRNet achieves the best MAE and competitive MSE on the SHA dataset as shown in Table 1. Figure 5 shows some estimated crowd density maps processed by our method in SHA test set. The images on the SHB dataset are taken with a camera in the bustling streets of Shanghai, among which 400 images are adopted for training and 316 are used for testing. The resolution of images are fixed at 768 × 1024 pixels with an average crowd number of 123. The crowd count in a single image changes between 9 and 578. Table 1 indicates that our algorithm achieves the lowest MAE and MSE compared with the state-of-the-art methods on the SHB dataset. Figure 7 presents some estimated crowd density maps output from our method on the SHB test set.

To verify the performance of our algorithm at different density levels, we divide the SHA and SHB datasets into 5 different density levels according to the number of people. Figure 6 reveals that the proposed DRNet performs well on all density levels. The crowd density in SHA is relatively large and the difference in density distribution is large, while the crowd density in SHB is small and the difference in density distribution is little. The proposed DRmap enables the network to adapt to different density changes by rectifying the density map. Therefore, the proposed DRNet achieves great performance on both SHA and SHB.

Table 1 Performance comparisons of different methods on Shanghaitech dataset

Method	SHA		SHB	
	MAE	MSE	MAE	MSE
MCNN [15]	110.2	173.2	26.4	41.3
IG-CNN [49]	72.5	118.2	13.6	21.1
ADCrowdNet [50]	63.2	98.9	8.2	15.7
BAYESIAN+ [51]	62.8	101.8	7.7	12.7
S-DCNet [52]	58.3	95.0	6.7	10.7
SPN+L2SM [53]	64.2	98.4	7.2	11.1
PGCNet [54]	57.0	**86.0**	8.8	13.7
AMSNet [9]	56.7	93.4	6.7	10.2
Liu et al. [55]	61.59	98.36	7.02	11.00
Yang et al. [4]	61.2	96.9	8.1	11.6
Jiang et al. [56]	57.78	90.13	–	–
SDANet [57]	63.6	101.8	7.8	10.2
DUBNet [58]	64.6	106.8	7.7	12.5
Wan et al. [59]	61.3	95.4	7.3	11.7
HPANet [60]	60.7	92.8	7.9	13.5
DMDCNet [61]	63.6	106.2	8.0	12.4
DRNet(Ours)	**54.9**	97.3	**6.5**	**10.2**

Bold values represent the best results

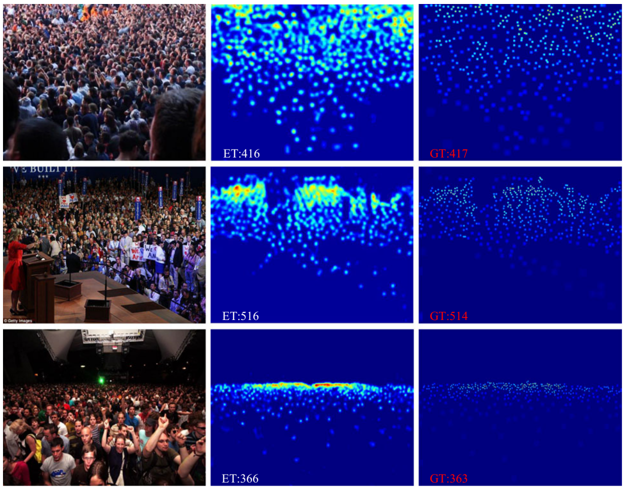


Fig. 5 The exhibition of the ground truth density maps and estimated density maps generated by the devised DRNet on the SHA dataset. The first column denotes input crowd images. The second column and the third column present estimated density maps and ground truth density maps, respectively. GT and ET represent ground truth crowd numbers and estimated crowd numbers

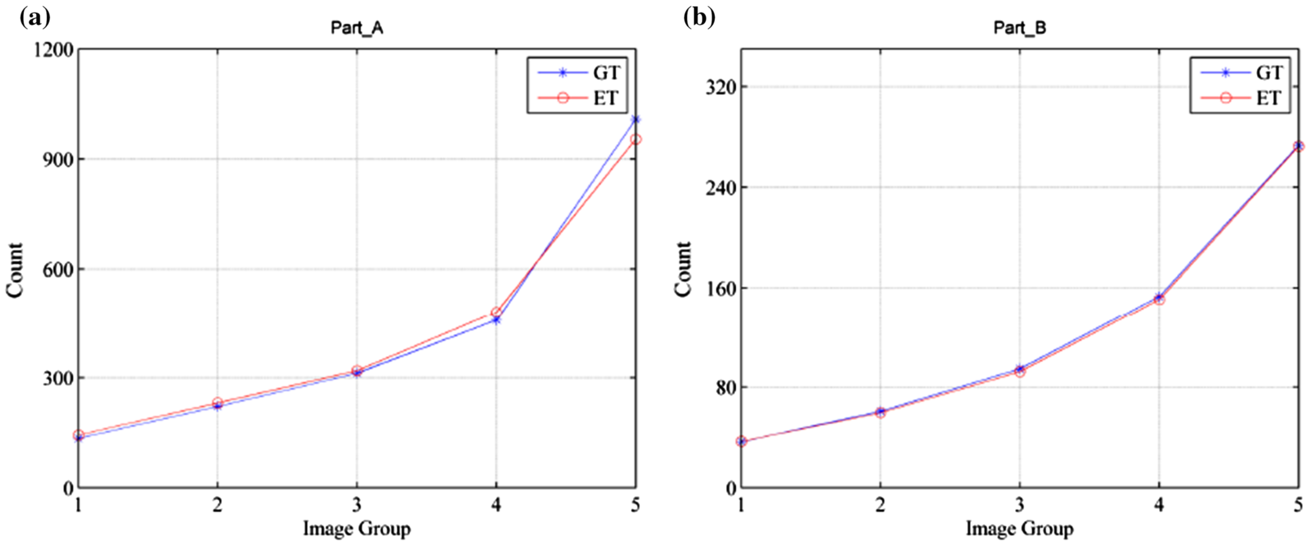


Fig. 6 The ground truth crowd counts and the estimated crowd counts output from DRNet on SHA and SHB. GT represents the ground truth, while ET indicates the estimated value. Image Group denotes different crowd density levels

4.2.2 Results on UCF-QNRF dataset

Idress et al. [46] come up with the UCF-QNRF (QNRF) dataset where the images are captured from Flickr, Web Search and Hajj footage. The images from Hajj contain various positions, perspectives, angles and times, while the search keywords leveraged in Flickr and the Web include crowd, hajj, spectator crowd, pilgrimage, protest and so on. The unqualified images will be discarded, such as low resolution, low-density crowd, blurred images and images with watermark. 14 annotators and 4 verifiers spend a total of 2,000 human-hours annotating these images. There are 1535 images with 1, 251, 462 labeled heads on the QNRF dataset, and the average size of the images are 2013 × 2902 pixels. The maximum number of people and the minimum number of people on QNRF are 12,865 and 49, and the median and average are 425 and 815.4, respectively. The experimental results in Table 2 denote that our algorithm achieves the best MAE and MSE on QNRF compared with other superior approaches.

We believe that the following reasons contribute to the above excellent results. The image resolution on QNRF is high, and a single image contains more marked head points, which is conducive to generate a more accurate ground truth DRmap for training DRNet. In addition, the designed DPFM module enhances the crowd features in different density areas through dual-level feature fusion, and promotes the network to generate more accurate density maps. Figure 8 depicts some estimated crowd density maps output from our method on the QNRF test set.

Table 2 Performance comparisons of different methods on UCF-QNRF dataset

Method	MAE	MSE
MCNN [15]	315	508
Idrees [46]	132	191
BAYESIAN+ [51]	88.7	154.8
S-DCNet [52]	104.4	176.1
SPN+L2SM [53]	104.7	173.6
AMSNet [9]	101.8	163.2
CAN [14]	107	183
Liu et al. [55]	86.6	152.2
Jiang et al. [56]	91.59	159.71
DUBNet [58]	105.6	180.5
Wan et al. [59]	84.3	147.5
HPANet [60]	107.7	188.5
DMDCNet [61]	108	189
DRNet(Ours)	**82.1**	**140.3**

Bold values represent the best results

4.2.3 Results on JHU-CROWD++ dataset

The JHU-CROWD++ (JHU) dataset is put forward by Sindagi et al. [47], and the images are downloaded on the Internet through searching different keywords such as crowd, crowd+outdoor, crowd+conference and crowd+station. There are 4372 images on JHU with a total of 1,515,005 marked heads. The average crowd number in each iamge is about 346. The JHU dataset provides both image-level and head-level annotations. The head label includes the position of the head point, the occlusion level

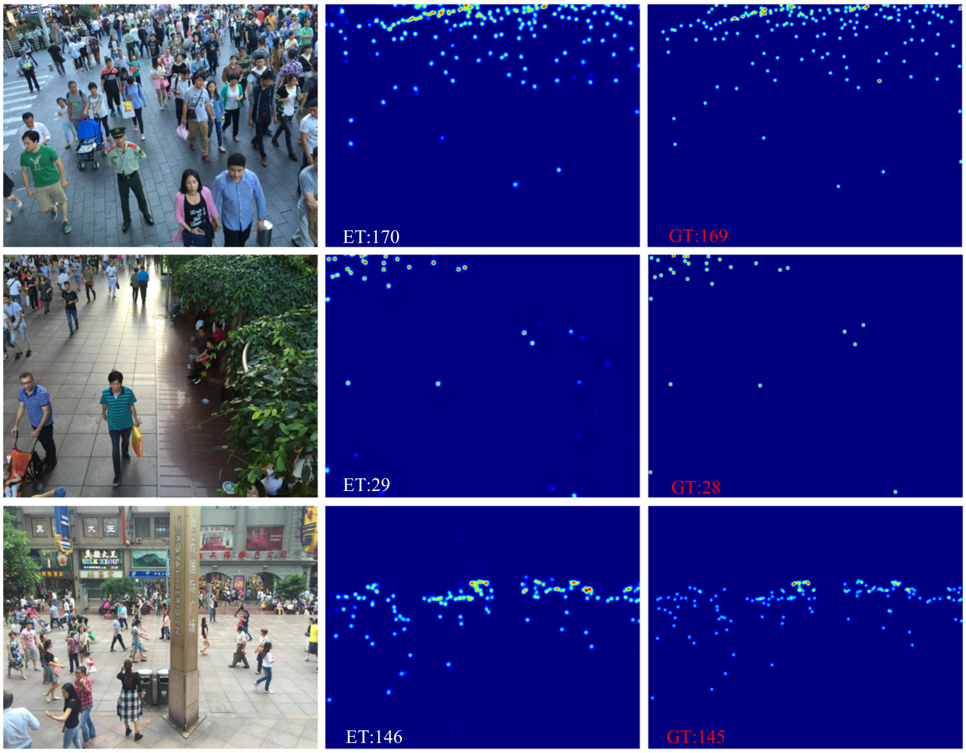


Fig. 7 The exhibition of the ground truth density maps and estimated density maps generated by the proposed DRNet on the SHB dataset. The first column denotes input crowd images. The second column and the third column present estimated density maps and ground truth density maps, respectively. GT and ET represent ground truth crowd numbers and estimated crowd numbers

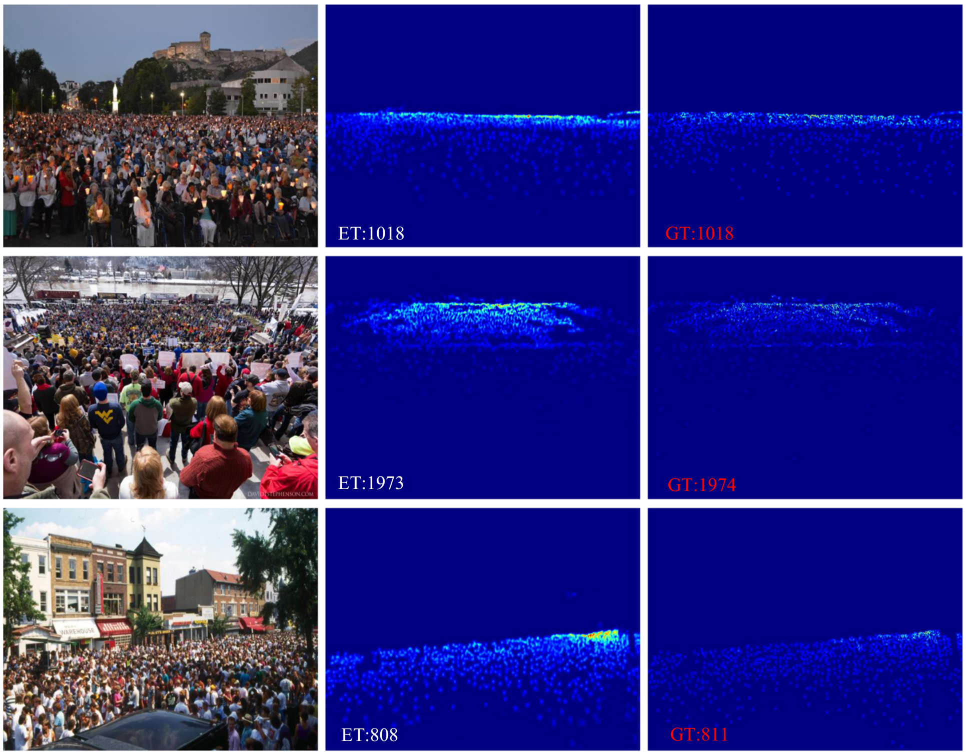


Fig. 8 The exhibition of the ground truth density maps and estimated density maps generated by DRNet on UCF-QNRF dataset. The first column denotes input crowd images. The second column and the third column present estimated density maps and ground truth density maps, respectively. GT and ET represent ground truth crowd numbers and estimated crowd numbers

of the head point (no occlusion, partial occlusion, full occlusion), blur level (blur, no blur) and head size, while the image-level annotation contains scene tags (mall, gathering, street, stadium, rally, protest, railway station) and weather tags (rain, snow, fog). The 4372 images on JHU are divided into a training set, validation set and test

Table 3 Performance comparisons of different methods on the JHU-CROWD++ (val set) dataset

Method	Overall		Low		Medium		High		Weather	
	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE
MCNN [15]	160.6	377.7	90.6	202.9	125.3	259.5	494.9	856.0	241.1	532.2
CMTL [62]	138.1	379.5	50.2	129.2	88.1	170.7	583.1	986.5	165.0	312.9
CSRNet [63]	72.2	249.9	22.2	40.0	49.0	99.5	302.5	669.5	83.0	168.7
SANet [64]	82.1	272.6	13.6	26.8	50.4	78.0	397.8	749.2	72.2	126.7
CAN [14]	89.5	239.3	34.2	69.5	65.6	115.3	336.4	619.7	101.8	179.3
SFCN [2]	62.9	247.5	11.8	19.8	39.3	73.4	297.4	679.4	52.3	93.6
DSSI-Net [65]	116.6	317.4	50.3	85.9	82.4	164.5	436.6	814.0	155.7	314.8
MBTTBF [66]	73.8	256.8	23.3	48.5	53.2	119.9	294.5	674.5	88.2	200.8
CG-DRCN-CC-VGG16 [47]	67.9	262.1	17.1	44.7	40.8	71.2	317.4	719.8	63.5	116.6
CG-DRCN-CC-Res101 [47]	57.6	244.4	11.7	24.8	35.2	57.5	273.9	676.8	54.0	106.8
DRNet(Ours)	**50.5**	**203.7**	**8.5**	**14.9**	**31.0**	**50.7**	**243.9**	**563.6**	**39.4**	**78.6**

Bold values represent the best results

Overall represents all of the images in the val set that contains four sub-categories such as low, medium, high and weather

Table 4 Performance comparisons of different methods on the JHU-CROWD++ (test set) dataset

Method	Overall		Low		Medium		High		Weather	
	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE
MCNN [15]	188.9	483.4	97.1	192.3	121.4	191.3	618.6	1166.7	330.6	852.1
CMTL [62]	157.8	490.4	58.5	136.4	81.7	144.7	635.3	1225.3	261.6	816.0
CSRNet [63]	85.9	309.2	27.1	64.9	43.9	71.2	356.2	784.4	141.4	640.1
SANet [64]	91.1	320.4	17.3	37.9	46.8	69.1	397.9	817.7	154.2	685.7
CAN [14]	100.1	314.0	37.6	78.8	56.4	86.2	384.2	789.0	155.4	617.0
SFCN [2]	77.5	297.6	16.5	55.7	38.1	59.8	341.8	758.8	122.8	606.3
DSSI-Net [65]	133.5	416.5	53.6	112.8	70.3	108.6	525.5	1047.4	229.1	760.3
MBTTBF [66]	81.8	299.1	19.2	58.5	41.6	66.0	352.2	760.4	138.7	631.6
CG-DRCN-CC-VGG16 [47]	82.3	328.0	19.5	58.7	38.4	62.7	367.3	837.5	138.6	654.0
CG-DRCN-CC-Res101 [47]	71.0	278.6	14.0	42.8	35.0	**53.7**	314.7	712.3	120.0	580.8
DRNet(Ours)	**61.7**	**260.1**	**11.3**	**32.5**	**31.2**	55.0	**272.5**	**664.1**	**96.1**	**579.9**

Bold values represent the best results

Overall denotes the sum total of the images in the test set that contains four sub-categories such as low, medium, high, and weather

set. The training set has 2272 images, including 636 low-density images (0–50), 1307 medium-density images (51–500) and 329 high-density images (500+). There are 76 rain images, 102 snow images and 81 fog images in the training set. The validation set has a total of 500 images, separated into 163 low-density images, 274 medium-density images and 64 high-density images. Moreover, there are 20 rain images, 21 snow images and 23 fog images in the validation set. The test set has 1600 images, containing 429 low density images, 931 medium density images and 240 high density images. The weather category includes 49 rain images, 78 snow images and 64 fog images.

Tables 3 and 4 reveal that our method achieves the best MAE and MSE compared with other state-of-the-art methods on the JHU validation set and test set. In particular, the proposed DRNet achieves the best performance in multiple subcategories, such as low, medium, high and weather. In general, our network has achieved excellent performance in different density levels and various outdoor weather based on the proposed DPFM feature fusion

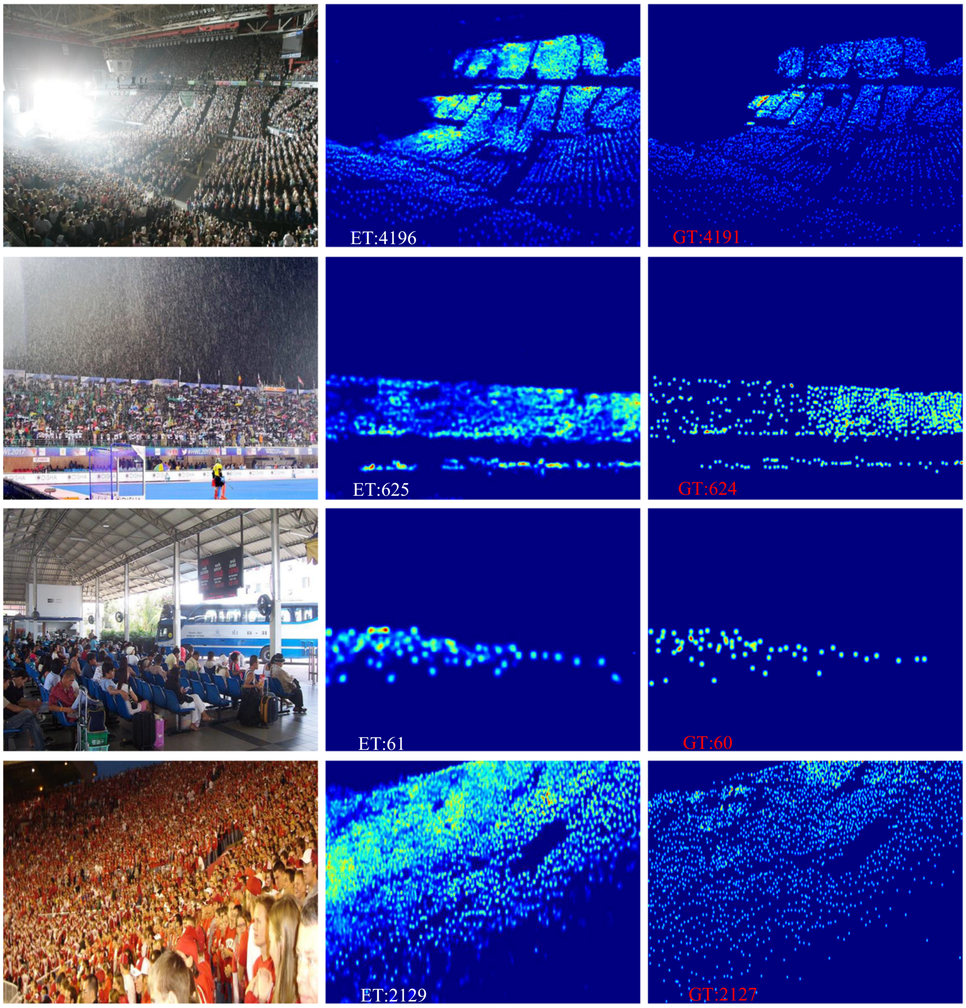


Fig. 9 The exhibition of the ground truth density maps and estimated density maps generated by DRNet on JHU-CROWD++ dataset. The first column denotes input crowd images. The first two rows are from the validation set, and the rest are from the test set. The second column and the third column present estimated density maps and ground truth density maps, respectively. GT and ET represent ground truth crowd numbers and estimated crowd numbers

Table 5 Performance comparisons of different methods on NWPU-Crowd dataset

Method	Validation set		Test set					
	Overall		Overall		Scene level (only MAE)		Luminance (only MAE)	
	MAE	MSE	MAE	MSE	Avg	S0 ∼ S4	Avg	L0 ∼ L2
MCNN [15]	218.53	700.61	232.5	714.6	1171.9	356.0/72.1/103.5/509.5/4818.2	220.9	472.9/230.1/181.6
SANet [64]	171.16	471.51	190.6	491.4	716.3	432.0/65.0/104.2/385.1/2595.4	153.8	254.2/192.3/169.7
Reg+det net [10]	245.8	700.3	264.9	759.0	1242.5	443.0/125.5/140.5/461.5/5036.6	313.6	464.2/267.4/209.1
PCC-net-light [23]	141.37	630.72	167.4	566.2	944.9	85.3/25.6/80.4/424.2/4108.9	141.2	253.1/167.9/144.9
C3F-VGG [67]	105.79	504.39	127.0	439.6	666.9	140.9/26.5/58.0/307.1/2801.8	127.9	296.1/125.3/91.3
CSRNet [63]	104.89	433.48	121.2	387.8	522.7	176.0/35.8/59.8/285.8/2055.8	112.0	232.4/121.0/95.5
PCC-Net-VGG [23]	100.77	573.19	112.3	457.0	777.6	103.9/13.7/42.0/259.5/3469.1	111.0	251.3/111.0/82.6
CANet [14]	93.58	489.90	106.3	386.5	612.2	82.6/14.7/46.6/269.7/2647.0	102.1	222.1/104.9/82.3
SCAR [11]	81.57	**397.92**	110.0	495.3	718.3	122.9/16.7/46.0/241.7/3164.3	102.3	223.7/112.7/73.9
BL [51]	93.64	470.38	105.4	454.2	750.5	66.5/8.7/41.2/249.9/3386.4	115.8	293.4/102.7/68.0
SFCN [2]	95.46	608.32	105.7	424.1	712.7	54.2/14.8/44.4/249.6/3200.5	106.8	245.9/103.4/78.8
DRNet(Ours)	**81.4**	512.0	**86.8**	**351.2**	**513.6**	80.4/10.1/37.0/217.5/2223.2	**85.0**	187.1/84.5/69.5

Bold values represent the best results

Overall represents the entire validation set or test set. Avg. indicates the average of different sub-category levels

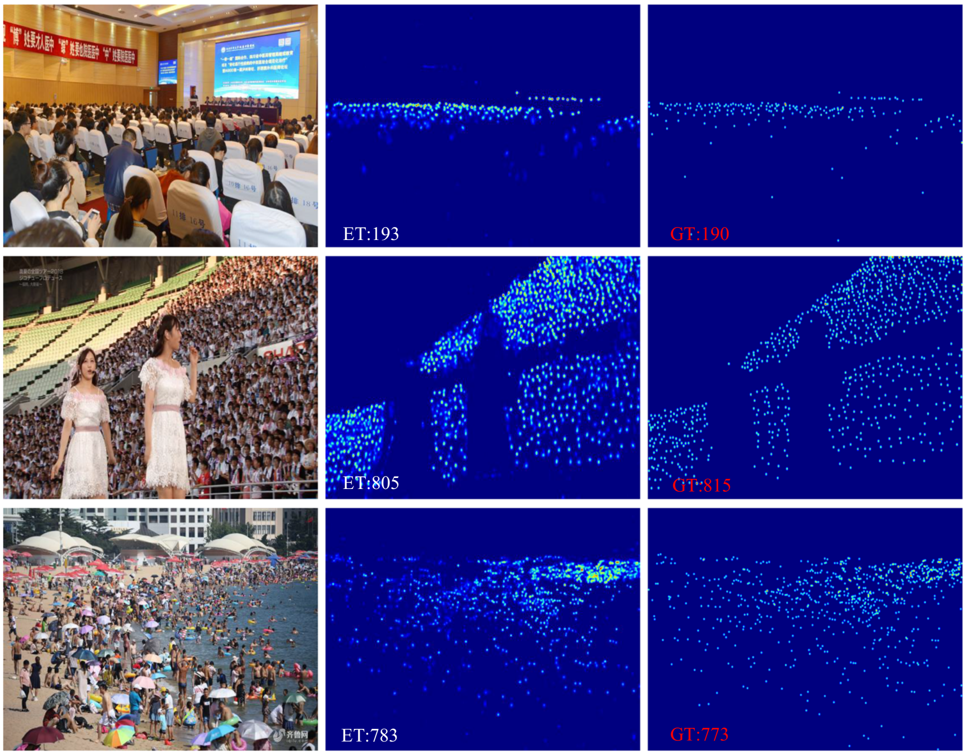


Fig. 10 The exhibition of the ground truth density maps and estimated density maps generated by the proposed DRNet on NWPU-Crowd dataset. The first column denotes input crowd images. The second column and the third column present estimated density maps and ground truth density maps, respectively. GT and ET represent ground truth crowd numbers and estimated crowd numbers

module and the DRmap auxiliary learning task. Several density maps output by DRNet on the JHU-CROWD++ validation set and test set are shown in Fig. 9. The proposed DRNet outputs accurate density estimation in crowd scenes of different density levels.

4.2.4 Results on NWPU-Crowd dataset

The NWPU-Crowd dataset is constructed and annotated by Wang et al. [48], where crowd images are sourced from camera shots and Internet downloads. For the former, more than 2,000 images are taken in some typical crowd scenes including tourist places, pedestrian streets, campuses, shopping malls, squares, museums and platforms. To collect images with denser crowds, Wang et al. [48] search the Internet for different keywords such as spring festival travel, crowded seas, job fairs and crowding through the Baidu, Bing and Sogou search engines. The NWPU-Crowd dataset contains 5,109 labeled images with a total of 2,133,375 labeled people heads. The average resolution of the images is 2191 × 3209 pixels, and the number of people in a single image is in [0, 20033]. Different from other datasets, the NWPU-Crowd dataset contains 351 negative samples with texture features similar to crowds in congested scenes, such as migrating animal communities, sculptures and terracotta warriors, which improves the adaptability of the model in practical application scenarios. The NWPU-Crowd dataset is divided into three parts: 3109 for training, 500 for validation and 1500 for testing. The test set images do not contain annotations, but researchers can obtain the test results through the online evaluation benchmark website.

The experimental results in Table 5 indicate that our method achieves close counting errors on the NWPU-Crowd validation set and test set, which confirm the good generalization ability of the proposed DRNet. Compared with other algorithms in Table 5, the proposed DRNet achieves competitive performance at the entire dataset level and multiple sub-category levels such as different density scene levels (S0 ~ S4) and different luminance levels (L0 ~ L2), further revealing the robustness of the proposed DRNet in different crowd scenes. In addition, Fig. 10 shows some examples of density map output by the proposed DRNet aimed at crowd images occluded by

Table 6 Performance comparisons of different network architectures on SHA dataset

Method	MAE	MSE
W/O DRmap+DPFM	66.7	105.1
W/O DRmap	60.8	105.9
W/O DPFM	60.1	100.5
DRNet(Ours)	**54.9**	**97.3**

Bold values represent the best results

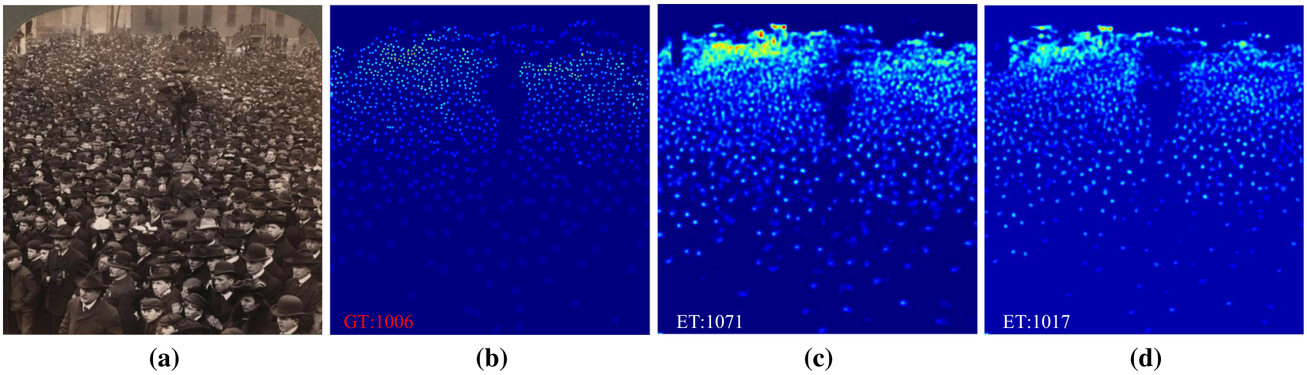


Fig. 11 Validation of DRmap auxiliary learning in occlusion scenes

desks, chairs and umbrellas, which demonstrates that the DRmap auxiliary learning contributes to generating accurate density maps in heavily occluded scenes.

4.3 Ablation experiments

In the ablation experiments section, we first verify the effectiveness of the proposed DRmap and DPFM. Then, we select the optimal correction coefficient K and hyperparameter q in DRmap through comparative experiments. Finally, we compare the effects of the feature pyramid fusion module (FPN) [43] and the proposed DPFM module on model performance.

4.3.1 Network architecture

We remove each component in DRNet consecutively to demonstrate the effectiveness of the proposed DRmap and DPFM. As shown in Table 6, "W/O DRmap+DPFM" represents the baseline network that removes DRmap and DPFM. "W/O DRmap" means that DRNet removes the DRmap density correction auxiliary task, while "W/O DPFM " denotes that DPFM module is removed from DRNet.

Table 7 Performance comparisons of DRNet with different correction coefficient on SHA dataset

Correction coefficient	MAE	MSE
$K=1$	56.9	102.9
$K=2$	**54.9**	**97.3**
$K=3$	56.0	98.4
$K=4$	58.2	103.9
$K=5$	58.6	102.0

Bold values represent the best results

Table 8 Performance comparisons of DRNet with different hyperparameters q on SHA dataset

Hyperparameter	MAE	MSE
$q=3$	57.3	101.2
$q=5$	**54.9**	**97.3**
$q=7$	55.9	100.2
$q=9$	56.1	100.5

Bold values represent the best results

The experimental results in Table 6 indicate that the designed DRmap and DPFM decreases the counting errors of the baseline network by 9.8% and 8.8%, respectively. The proposed DRNet includes both DRmap and DPFM, which improves the baseline network by 17.6%. The above experimental results fully prove that the proposed DRmap and DPFM are effective in improving the counting performance of the model.

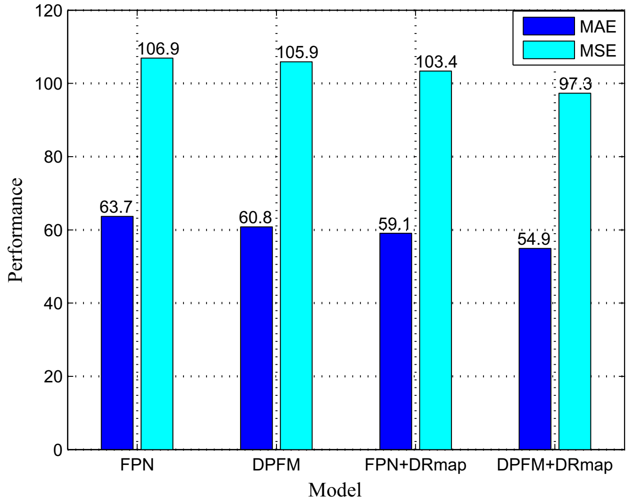


Fig. 12 Effects of DPFM module and FPN module on model performance

Table 9 Cross-domain experiment comparisons among SHA, SHB, and QNRF

Method	DA	SHA → SHB		SHA → QNRF		SHB → SHA		SHB → QNRF	
		MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE
D-ConvNet-v1 [68]	✗	49.1	99.2	–	–	140.4	226.1	–	–
CACC [40]	✔	–	–	–	–	115.6	199.5		–
MCNN [15]	✗	130.7	161.9	396.9	584.1	213.5	315.5	415.5	676.6
MCNN+Our	✔	53.5	86.4	263.6	406.3	154.8	230.1	299.9	488.3
DRNet	✗	27.9	42.7	132.8	244.6	111.5	202.4	203.9	**355.5**
DRNet+Our	✔	**14.6**	**32.9**	**124.1**	**219.8**	**101.6**	**179.6**	**199.9**	371.6

Method	DA	QNRF → SHA		QNRF → SHB	
		MAE	MSE	MAE	MSE
MCNN [15]	✗	149.6	221.9	72.6	102.4
MCNN+Our	✔	136.6	217	42.7	65.7
DRNet	✗	65.8	98.9	10.7	18.6
DRNet+Our	✔	**58.9**	**94.6**	**9.3**	**15.1**

Bold values represent the best results

DA represents the domain adaptation method. The left-end of the arrow denotes the source domain, and the right-end reveals the target domain

To verify the effectiveness of DRmap auxiliary learning in severely occluded scenes, we adopt the trained DRNet model and "W/O DRmap" model to estimate the crowd density maps of occluded scene images as shown in Fig. 11. Figure 11 a is the input image, where crowd heads are heavily occluded by hats, and heads in distant dense crowd areas occlude each other. Figure 11 b denotes the ground-truth density map, while Fig. 11c and d are estimated density maps generated by the "W/O DRmap" model and DRNet model, respectively. Compared with Fig. 11c, the counting error of Fig. 11d is reduced by 54, and the density map distribution is closer to the ground-truth density map. The density map comparison in Fig. 11 reveals that the proposed DRmap auxiliary learning enables the model to accurately estimate crowd density maps in images of heavily occluded scenes.

4.3.2 The correction coefficient of DRmap

The correction coefficient in DRmap is a key parameter that affects the counting performance of the network. If the correction value is too large, it is easy to overestimate the crowd density, and vice versa. We leverage the DRmaps generated by different correction coefficients to rectify the initial density map generated by the network. As shown in Table 7, the experimental results reveal that the network obtains the lowest MAE as the correction coefficient K is set to 2. Therefore, we finally select $K = 2$ as the correction coefficient in the proposed DRmap.

4.3.3 The hyperparameter q in DRmap

As introduced in Sect. 3.1, we propose a DRmap based on the distance between the crowd heads to rectify counting errors in different density regions. In Sect. 3.1, the initial crowd head distances are calculated by Eq. 1. To analyze the effects of different hyperparameters q in Eq. 1, we conduct several ablation experiments on SHA dataset as shown in Table 8. The hyperparameter q in formula (1) is set to 3, 5, 7, and 9 for generating the corresponding ground truth DRmap. Then, the proposed density rectification network (DRNet) is trained separately with different ground truth DRmaps. The experimental results in Table 8 show that DRNet achieves the lowest MAE and MSE when the q is set to 5. To obtain better counting performance, hyperparameter q is selected as 5 in formula (1).

4.3.4 Comparison of FPN and DPFM module

To further demonstrate the effectiveness of the DPFM module, we conduct several ablation experiments as shown in Fig. 12. The FPN model and DPFM model represent deploying the FPN module [43] and DPFM module on the VGG baseline network, respectively. The FPN+DRmap model denotes that the FPN module [43] and DRmap auxiliary learning task are deployed on the VGG baseline network, while the DPFM+DRmap model means that the proposed DPFM module and DRmap auxiliary learning task are applied on the VGG baseline network.

The experimental results in Fig. 12 reveal that the DPFM model decreases the MAE and MSE by 2.9 and 1.0 based on the FPN model. Compared with the FPN+DRmap model, the MAE and MSE of DPFM+DRmap model are reduced by 4.2 and 6.1. The comparison of the above experimental results confirms that the proposed DPFM module can reduce the counting error more effectively than the FPN module [43].

4.4 Cross-domain research

The study of cross-domain issue is a significant work, which can relieve the domain gap between different crowd scenarios and speed up the landing process of crowd counting. To evaluate our proposed domain adaption method more objectively, we choose three datasets with large differences in scenarios and small gaps in image numbers to conduct cross-domain experiments including SHA, SHB, and QNRF. We first conduct cross-domain comparison experiments based on DRNet proposed in this paper. As shown in Table 9, "DRNet " denotes that DRNet is trained on the source domain and directly tested on the target domain, while "DRNet+Our" means that we train DRNet on both the source and target domains using our domain adaption approach. When SHA is selected as the source domain and the target domain is SHB or QNRF, the proposed domain adaption method reduces the MAE of DRNet on SHB and QNRF by 13.3 and 8.7. Then, SHB is served as the source domain, the domain adaption method decreases the MAE of DRNet on SHA and QNRF by 9.9 and 4.0. After that, we choose QNRF as the source domain, the MAE of DRNet on SHA and SHB is reduced by 6.9 and 1.4 when leveraging our domain adaptation method. Moreover, we compare our method with other superior cross-domain methods as shown in Table 9. The experimental results show that our domain adaption approach obtains the best cross-domain performance compared with others.

Considering that the designed DRNet contains a VGG pre-trained model, we also verify the effectiveness of the proposed domain adaption method on networks that do not include any pre-training models, such as MCNN [15]. As depicted in Table 9, "MCNN" reveals that the well-trained MCNN model on the source domain is directly tested on the target domain, while "MCNN+Our" represents that we combine the source domain and the target domain images to train MCNN by adopting our domain adaption algorithm. Firstly, SHA is selected as the source domain and we accept SHB and QNRF as the target domain, the proposed domain adaption algorithm decreases the MAE of the MCNN in SHB and QNRF by 77.2 and 133.6. Furthermore, SHB is chosen as the source domain, and the domain adaption method reduces the MAE of MCNN on SHA and QNRF by 58.7 and 115.6. Moreover, QNRF is selected as the source domain, the MAE of MCNN on SHA and SHB are reduced by 13 and 29.9 by utilizing our domain adaption method, respectively. The above experimental results indicate that the cross-domain method we put forward can effectively relieve the domain gap between different source domains and target domains, and improve the performance of the model in unknown scenarios. In addition, the experimental results further reveal that our domain adaption is more effective in networks that do not include any pre-training model such as MCNN. The main reason for our analysis is that the VGG pre-training model has learned much knowledge from the object detection field, while MCNN has not previously learned information other than the source domain. The proposed domain adaption method encourages the network to learn abundant domain invariant features, improving the cross-domain performance.

5 Conclusion

In this paper, we propose a density rectifying network (DRNet) and a domain adaption method to address nonuniform density distribution and cross-domain issues. The proposed DRNet contains several DPFM modules that carry out a dual-layer fusion of crowd density features of different scales for generating high quality density maps. The devised DRmap auxiliary learning task further rectifies the incorrect density estimation by adaptively weighting the initial crowd density maps pixel-by-pixel. To deal with the cross-domain problem, the proposed domain adaption method learns domain invariant features between the source domain and the target domain by randomly cutting mixed dual-domain images from global and local perspectives. Experimental results prove that the devised DRNet achieves the lowest MAE and superior MSE compared with other excellent algorithms on multiple mainstream datasets including Shanghaitech, UCF-QNRF, JHU-CROWD++ and NWPU-Crowd. In addition, we conduct several cross-domain experiments on different source domains and target domains. Experimental results demonstrate that the proposed domain adaption method is effective in improving the cross-domain performance of the models and obtains the best MAE and MSE on the target domain compared with other approaches.

Acknowledgements This work is supported by the National Natural Science Foundation of China under grant No.62133013 and sponsored by the CAAI-Huawei MindSpore Open Fund.

Availability of data and materials The datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.

Declarations

Conflict of interest The authors declared that they have no conflicts of interest in this article.

References

1. LeCun Y, Bottou L, Bengio Y, Haffner P (1998) Gradient-based learning applied to document recognition. Proc IEEE 86(11):2278–2324
2. Wang Q, Gao J, Lin W, Yuan Y (2019) Learning from synthetic data for crowd counting in the wild. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 8198–8207
3. Wang Q, Han T, Gao J, Yuan Y (2021) Neuron linear transformation: modeling the domain shift for crowd counting. IEEE Trans Neural Netw Learn Syst. https://doi.org/10.1109/TNNLS.2021.3051
4. Yang Y, Li G, Wu Z, Su L, Huang Q, Sebe N (2020) Reverse perspective network for perspective-aware object counting. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4374–4383 (2020)
5. Wang M, Cai H, Zhou J, Gong M (2021) Interlayer and intralayer scale aggregation for scale-invariant crowd counting. Neurocomputing 441:128–137
6. Peng S, Wang L, Yin B, Li Y, Xia Y, Hao X (2021) Adaptive weighted crowd receptive field network for crowd counting. Pattern Anal Appl 24(2):805–817
7. Sam DB, Sajjan NN, Maurya H, Babu RV (2019) Almost unsupervised learning for dense crowd counting. In: Proceedings of the AAAI conference on artificial intelligence, pp 8868–8875
8. Sindagi VA, Yasarla R, Babu DS, Babu RV, Patel VM (2020) Learning to count in the crowd from limited labeled data. In: Proceedings of the european conference on computer vision, pp 212–229
9. Hu Y, Jiang X, Liu X, Zhang B, Han J, Cao X, Doermann D (2020) Nas-count: counting-by-density with neural architecture search. In: Proceedings of the european conference on computer vision, pp 747–766
10. Liu J, Gao C, Meng D, Hauptmann AG (2018) Decidenet: counting varying density crowds through attention guided detection and density estimation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 5197–5206 (2018)
11. Gao J, Wang Q, Yuan Y (2019) Scar:spatial-/channel-wise attention regression networks for crowd counting. Neurocomputing 363:1–8
12. Gao J, Yuan Y, Wang Q (2021) Feature-aware adaptation and density alignment for crowd counting in video surveillance. IEEE Trans Cybernetics 51(10):4822–4833
13. Amirgholipour, S., He, X., Jia, W., Wang, D., Zeibots M (2018) A-CCNN: adaptive CCNN for density estimation and crowd counting. In: Proceedings of the IEEE international conference on image processing, pp 948–952. IEEE
14. Liu W, Salzmann M, Fua P (2019) Context-aware crowd counting. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 5099–5108
15. Zhang Y, Zhou D, Chen S, Gao S, Ma Y (2016) Single-image crowd counting via multi-column convolutional neural network. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 589–597
16. Babu Sam D, Surya S, Venkatesh Babu R (2017) Switching convolutional neural network for crowd counting. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 5744–5752
17. Sindagi VA, Patel VM (2017) Generating high-quality crowd density maps using contextual pyramid cnns. In: Proceedings of the IEEE international conference on computer vision, pp 1861–1870
18. Cheng Z-Q, Li J-X, Dai Q, Wu X, He J-Y, Hauptmann AG (2019) Improving the learning of multi-column convolutional neural network for crowd counting. In: Proceedings of the 27th ACM international conference on multimedia, pp 1897–1906
19. Sam DB, Babu RV (2018) Top-down feedback for crowd counting convolutional neural network. In: Proceedings of the AAAI conference on artificial intelligence, pp 7323–7330
20. Jiang X, Xiao Z, Zhang B, Zhen X, Cao X, Doermann D, Shao L (2019) Crowd counting and density estimation by trellis encoder-decoder networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 6133–6142
21. Liu Y, Wen Q, Chen H, Liu W, Qin J, Han G, He S (2020) Crowd counting via cross-stage refinement networks. IEEE Trans Image Process 29:6800–6812
22. Liu X, Van De Weijer J, Bagdanov AD (2019) Exploiting unlabeled data in cnns by self-supervised learning to rank. IEEE Trans Pattern Anal Machine Intell 41(8):1862–1878
23. Gao J, Wang Q, Li X (2019) Pcc net: perspective crowd counting via spatial convolutional network. IEEE Trans Circuits Syst Video Technol 30(10):3486–3498
24. Shi Z, Zhang L, Sun Y, Ye Y (2018) Multiscale multitask deep netvlad for crowd counting. IEEE Trans Industrial Inform 14(11):4953–4962
25. Zhao M, Zhang J, Zhang C, Zhang W (2019) Leveraging heterogeneous auxiliary tasks to assist crowd counting. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 12736–12745
26. Jiang X, Zhang L, Zhang T, Lv P, Zhou B, Pang Y, Xu M, Xu C (2020) Density-aware multi-task learning for crowd counting. IEEE Trans Multimed 23:443–453
27. Zhang Q, Chan AB (2019) Wide-area crowd counting via ground-plane density maps and multi-view fusion cnns. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 8297–8306
28. Zhang Q, Lin W, Chan AB (2021) Cross-view cross-scene multi-view crowd counting. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 557–567
29. Peng T, Li Q, Zhu P (2020) Rgb-t crowd counting from drone: a benchmark and mmccn network. In: Proceedings of the Asian conference on computer vision, pp 497–513
30. Wen L, Du D, Zhu P, Hu Q, Wang Q, Bo L, Lyu S (2021) Detection, tracking, and counting meets drones in crowds: a benchmark. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 7812–7821
31. Bai Z, Wang Z, Wang J, Hu D, Ding E (2021) Unsupervised multi-source domain adaptation for person re-identification. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 12914–12923
32. Faraki M, Yu X, Tsai Y-H, Suh Y, Chandraker M (2021) Cross-domain similarity learning for face recognition in unseen domains. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 15292–15301
33. Fu Y, Zhang M, Xu X, Cao Z, Ma C, Ji Y, Zuo K, Lu H (2021) Partial feature selection and alignment for multi-source domain adaptation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 16654–16663
34. He J, Jia X, Chen S, Liu J (2021) Multi-source domain adaptation with collaborative learning for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 11008–11017

35. Zhang C, Li H, Wang X, Yang X (2015) Cross-scene crowd counting via deep convolutional neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 833–841
36. Hossain MA, Kumar M, Hosseinzadeh M, Chanda O, Wang Y (2019) One-shot scene-specific crowd counting. In: Proceedings of the British machine vision conference, pp 1–11
37. Li W, Yongbo L, Xiangyang X (2019) Coda: Counting objects via scale-aware adversarial density adaption. In: Proceedings of the International conference on multimedia and expo, pp 193–198
38. Han T, Gao J, Yuan Y, Wang Q (2020) Focus on semantic consistency for cross-domain crowd understanding. In: ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp 1848–1852 . IEEE
39. He Y, Ma Z, Wei X, Hong X, Ke W, Gong Y (2021) Error-aware density isomorphism reconstruction for unsupervised cross-domain crowd counting. In: Proceedings of the AAAI conference on artificial intelligence, pp 1540–1548
40. Liu Y, Xu D, Ren S, Wu H, Cai H, He S (2021) Fine-grained domain adaptive crowd counting via point-derived segmentation. arXiv preprint arXiv:2108.02980
41. Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556
42. Sandwell DT (1987) Biharmonic spline interpolation of geos-3 and seasat altimeter data. Geophys Res Lett 14(2):139–142
43. Lin T-Y, Dollár P, Girshick R, He K, Hariharan B, Belongie S (2017) Feature pyramid networks for object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2117–2125
44. Yun S, Han D, Oh SJ, Chun S, Choe J, Yoo Y (2019) Cutmix: Regularization strategy to train strong classifiers with localizable features. In: Proceedings of the IEEE international conference on computer vision, pp 6023–6032
45. Peng S, Yin B, Hao X, Yang Q, Kumar A, Wang L (2021) Depth and edge auxiliary learning for still image crowd density estimation. Pattern Anal Appl 24(4):1777–1792
46. Idrees H, Tayyab M, Athrey K, Zhang D, Al-Maadeed S, Rajpoot N, Shah M (2018) Composition loss for counting, density map estimation and localization in dense crowds. In: Proceedings of the European conference on computer vision, pp 532–546
47. Sindagi V, Yasarla R, Patel VM (2022) Jhu-crowd++: Large-scale crowd counting dataset and a benchmark method. IEEE Trans Pattern Anal Machine Intell 44(5):2594–2609
48. Wang Q, Gao J, Lin W, Li X (2020) Nwpu-crowd: a large-scale benchmark for crowd counting and localization. IEEE Trans Pattern Anal Machine intell 43(6):2141–2149
49. Sam DB, Sajjan NN, Babu RV, Srinivasan M (2018) Divide and grow: capturing huge diversity in crowd images with incrementally growing cnn. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3618–3626
50. Liu N, Long Y, Zou C, Niu Q, Pan L, Wu H (2019) Adcrowdnet: an attention-injective deformable convolutional network for crowd understanding. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3225–3234
51. Ma Z, Wei X, Hong X, Gong Y (2019) Bayesian loss for crowd count estimation with point supervision. In: Proceedings of the IEEE International conference on computer vision, pp 6142–6151
52. Xiong H, Lu H, Liu C, Liu L, Cao Z, Shen C (2019) From open set to closed set: counting objects by spatial divide-and-conquer. In: Proceedings of the IEEE international conference on computer vision, pp 8362–8371
53. Xu C, Qiu K, Fu J, Bai S, Xu Y, Bai X (2019) Learn to scale: generating multipolar normalized density maps for crowd counting. In: Proceedings of the IEEE international conference on computer vision, pp 8382–8390
54. Yan Z, Yuan Y, Zuo W, Tan X, Wang Y, Wen S, Ding E (2019) Perspective-guided convolution networks for crowd counting. In: Proceedings of the IEEE international conference on computer vision, pp 952–961
55. Liu X, Yang J, Ding W, Wang T, Wang Z, Xiong J (2020) Adaptive mixture regression network with local counting map for crowd counting. In: Proceedings of the European conference on computer vision, pp 241–257
56. Jiang X, Zhang L, Xu M, Zhang T, Lv P, Zhou B, Yang X, Pang Y (2020) Attention scaling for crowd counting. In: Proceedings of the IEEE Conference on computer vision and pattern recognition, pp 4706–4715
57. Miao Y, Lin Z, Ding G, Han J (2020) Shallow feature based dense attention network for crowd counting. In: Proceedings of the AAAI conference on artificial intelligence, pp 11765–11772
58. Oh M-h, Olsen P, Ramamurthy KN (2020) Crowd counting with decomposed uncertainty. In: Proceedings of the AAAI conference on artificial intelligence, pp 11799–11806
59. Wan J, Liu Z, Chan AB (2021) A generalized loss function for crowd counting and localization. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1974–1983
60. Zhang S, Zhang X, Li H, He H, Song D, Wang L (2022) Hierarchical pyramid attentive network with spatial separable convolution for crowd counting. Eng Appl Artif Intell 108:1–10
61. Yan L, Zhang L, Zheng X, Li F (2022) Deeper multi-column dilated convolutional network for congested crowd understanding. Neural Comput Appl 34(2):1407–1422
62. Sindagi VA, Patel VM (2017) Cnn-based cascaded multi-task learning of high-level prior and density estimation for crowd counting. In: Proceedings of the IEEE international conference on advanced video and signal based surveillance, pp 1–6
63. Li Y, Zhang X, Chen D (2018) Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1091–1100
64. Cao X, Wang Z, Zhao Y, Su F (2018) Scale aggregation network for accurate and efficient crowd counting. In: Proceedings of the European conference on computer vision, pp 734–750
65. Liu L, Qiu Z, Li G, Liu S, Ouyang W, Lin L (2019) Crowd counting with deep structured scale integration network. In: Proceedings of the IEEE international conference on computer vision, pp 1774–1783
66. Sindagi VA, Patel VM (2019) Multi-level bottom-top and top-bottom feature fusion for crowd counting. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 1002–1012
67. Gao J, Lin W, Zhao B, Wang D, Gao C, Wen J (2019) C^3 framework: An open-source pytorch code for crowd counting. arXiv preprint arXiv:1907.02724
68. Shi Z, Zhang L, Liu Y, Cao X, Ye Y, Cheng MM, Zheng G (2018) Crowd counting with deep negative correlation learning. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 5382–5390

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.