Front. Comput. Sci., 2023, 17(1): 171304
https://doi.org/10.1007/s11704-021-1207-x

RESEARCH ARTICLE

Scene-adaptive crowd counting method based on meta learning with dual-input network DMNet

Haoyu ZHAO[1], Weidong MIN (✉)[2,3], Jianqiang XU[1], Qi WANG[1], Yi ZOU[1], Qiyan FU[1]

1 School of Information Engineering, Nanchang University, Nanchang 330031, China
2 School of Software, Nanchang University, Nanchang 330047, China
3 Jiangxi Key Laboratory of Smart City, Nanchang 330047, China

Abstract Crowd counting is recently becoming a hot research topic, which aims to count the number of the people in different crowded scenes. Existing methods are mainly based on training-testing pattern and rely on large data training, which fails to accurately count the crowd in real-world scenes because of the limitation of model's generalization capability. To alleviate this issue, a scene-adaptive crowd counting method based on meta-learning with Dual-illumination Merging Network (DMNet) is proposed in this paper. The proposed method based on learning-to-learn and few-shot learning is able to adapt different scenes which only contain a few labeled images. To generate high quality density map and count the crowd in low-lighting scene, the DMNet is proposed, which contains Multi-scale Feature Extraction module and Element-wise Fusion Module. The Multi-scale Feature Extraction module is used to extract the image feature by multi-scale convolutions, which helps to improve network accuracy. The Element-wise Fusion module fuses the low-lighting feature and illumination-enhanced feature, which supplements the missing illumination in low-lighting environments. Experimental results on benchmarks, WorldExpo'10, DISCO, USCD, and Mall, show that the proposed method outperforms the existing state-of-the-art methods in accuracy and gets satisfied results.

Keywords crowd counting, meta-learning, scene-adaptive, Dual-illumination Merging Network

1 Introduction

Crowd counting is becoming an important technology to many applications, such as city security and crowd protection. The aim of this task is to count the number of the people automatically [1]. There are many excellent works have been proposed. These methods usually use deep learning methods to learn crowd feature with the help of density map [2,3]. The deep-learning-based methods need large dataset for training. But in real world, the data is expensive and inaccessible. And such models sometimes don't have good ability of scene generalization. It will lead to that the trained model cannot applied in real-world scenes [4]. It likes the great difference of the scenes between park and street.

To transfer the trained model to a real-world scene, several works tried self-supervised and meta-learning to solve such problem. Liu et al. [5] leveraged abundantly available unlabeled crowd imagery in a learning-to-rank framework. Zhang et al. [6] and Loy et al. [7] used semi-supervised method to finish the cross-scene challenge. Reddy et al. [4] proposed a scene adaptive crowd counting to deploy a crowd counting model specially adapted to a target camera. Although these methods got some good results, they fail to count the crowd when facing complex conditions, such as extremely low-lighting environments.

To alleviate the above problems, a novel scene-adaptive crowd counting method based on meta-learning with Dual-illumination Merging Network (DMNet) is proposed. This work is mainly based on learning-to-learn and few-shot learning, which lead to fast learning on a real-world scene. When the method is proposed and designed, the structures of the Chelsea et al. [8] have been analyzed and researched. To generate high quality density map and count the crowd in low-lighting scene, the DMNet is proposed as the backbone in scene-adaptive crowd counting method. The DMNet contains two submodules, named Multi-scale Feature Extraction module and Element-wise Fusion module. The Multi-scale Feature Extraction module extracts the multi-scale feature of image, which is adaptive to large variation in people size. The Element-wise Fusion module fuses the low-lighting feature and illumination-enhanced feature, which supplements the missing illumination in low-lighting environments. Thereinto, the illumination of image in low-lighting is enhanced to get the illumination-enhanced feature. Experimental results on benchmarks, WorldExpo'10, DISCO, USCD, and Mall, show that the proposed method outperforms the existing state-of-the-art methods in accuracy and gets satisfied results.

The main contributions of this study are summarized as follows.

- With the idea of "learning-to-learn", the proposed

Received April 28, 2021; accepted August 3, 2021

E-mail: minweidong@ncu.edu.cn

scene-adaptive crowd counting method can deal with different complex scenes.

- To estimate the density map of the crowd and count the number of people, the DMNet is proposed. The network is designed as the backbone in the scene-adaptive crowd counting method.
- Considering the influence of the illumination in complex scenes, the Element-wise Fusion module is proposed. This module can express the illumination information better, which is helpful to estimate the density map with higher accuracy.

2 Related work

There are many approaches are proposed to solve the crowd counting problem. These methods are mainly based on deep learning and get good performances. With the help of meta-learning, the models are further trained only using few images.

2.1 Crowd counting based on deep learning

The problems of apparent perspective distortion, dense crowd, and illumination variations affect the accuracy of the crowd counting results. Several works [9−11] have contributed to this problem. Some works tried to solve this problem in low-lighting conditions. Liu et al. [12] found the illumination variations further restrict the performance improvement of the counting method. Wu et al. [13] proposed an adaptive scenario discovery framework for counting crowds with varying densities, which is also helpful for non-uniform illumination aspect. Hu et al. [14] combined audio information and visual information to increase the accuracy of the counting task when facing extreme conditions, such as low illumination and occlusion. Some works proposed new approaches to solve the high density of the crowd. Such as Zhao et al. [15] trained an end-to-end approach which combines multi-scale features using multiple receptive field sizes and learns the feature-aware information on each image. Zheng et al. [16], Liu et al. [5], and Shen et al. [17] considered the relationship between part and global of the crowded image. These methods count the number of individuals and gathered them. Yang et al. [18] and Zou et al. [19] used multi-scale convolution network to estimate the density of the crowd. And some other methods, such as [20−22], also tried to remove the interference of background, locate every person in the crowd or use dedicated network to predict the object mask. The above methods are mainly proposed for general crowd counting framework. In real world, the crowd scenes are abundant and the training data are not available. So, the scene-adaptive crowd method with few training data is needed.

2.2 Meta-learning

With the development of deep learning [23−25], some new technologies, such as meta-learning and few-shot learning are also proposed. Different from machine learning, meta-learning aims to make machines learn how to learn. Hence, meta-learning is widely used. Ma et al. [26] proposed a novel generative adversarial network CGAN. The meta-learning structure is an auxiliary network to provide deconvolutional weights for CGAN. Jung et al. [27] proposed a novel meta-learning framework for real-time object tracking, which is trained by updating its meta-parameters for initial parameters. Elsken et al. [28] and Xu et al. [29] decreased the training cost with only few images with few-shot learning based on meta-learning, which got the good performances. Ye et al. [30] proposed an adaptively approach for meta-learning, which obtained high-quality model solutions efficiently. This work incorporated task context into the determination of the model initialization. Nichol et al. [31] considered meta-learning problems and obtained an agent that performs well when presented with a previously unseen task sampled from this distribution. Except for the applications of the meta-learning, some researchers also explored the fundamental research about meta-learning. Wang et al. [32] designed a hybrid meta-learning model which is able to handle flexible numbers of classes and combines the merits of both optimization-based methods. Lai et al. [33] proposed a novel meta-learning method to learn how to learn task-adaptive classifier-predictor to generate classifier weights for few-shot classification. These methods can handle flexible numbers of classes and resolve classification across tasks.

In conclusion, the crowd counting methods based on deep learning mainly rely on big data training. The generalization ability of these models usually is weak when facing complex conditions. Such as the illumination variations in new scenes is still problems. Meta-learning aims to train a pre-trained model, which can update to a new task with a few training data and is suitable for adaption problem. So, with the development of the meta-learning, a scene-adaptive crowd counting method is needed.

3 Overview of the proposed scene-adaptive crowd counting method

3.1 The framework of the proposed method

To enhance the generalization capability of crowd counting model in real-world scene, this work proposes a scene-adaptive crowd counting method based on meta-learning with Dual-illumination Merging Network (DMNet). The whole method is seen as Fig. 1, which is mainly based on the mechanism of learning-to-learn and few-shot learning. The proposed method based on meta-learning divides into two parts, the meta training part and the meta testing part. The meta training part is used to train a pre-trained model which is further trained in meta testing part with a few labeled images. To generate high quality density map and count the crowd in low-lighting scene, the DMNet is proposed. It is designed as backbone in scene-adaptive crowd counting method with dual inputs.

This section mainly introduces the whole process of the scene-adaptive crowd counting method, including meta training process and meta testing process. The details of the proposed DMNet are introduced in Section 4.

3.2 The training and testing processes of the proposed method

Meta training process The meta training process for crowd counting aims to learn a pre-trained model $\mathcal{M}_p$ in some source scenes and fine-turned in target scenes. When this method is

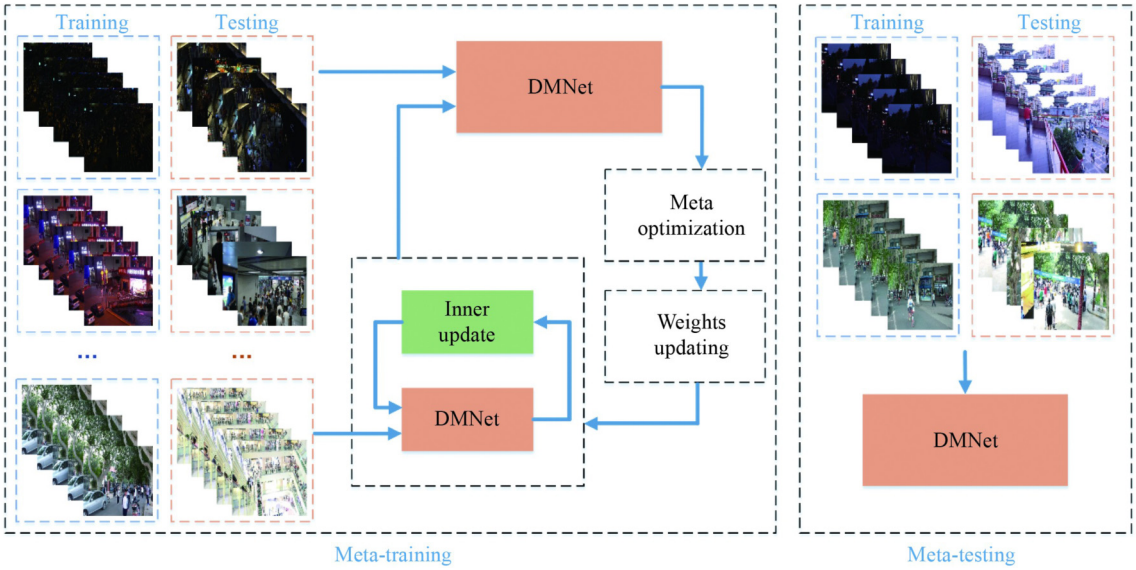


Fig. 1 The scene-adaptive crowd counting method with meta-learning. The meta-training step aims to train a pre-trained model. The meta-testing step aims to fine-turned the pre-trained model for new scenes

improved and designed, the structures of the MAML [8] have been analyzed and researched. Given a dataset D, the dataset is divides into training data D_1 and testing data D_2. As Fig. 1 shows, this work chooses different images in different scenes and different level of illumination. The D_1 and the D_2 are selected randomly. They are not from the same source. These images are from DISCO dataset [14]. Compared with the same illumination conditions in dataset [6,7,34], these images are more challenging. D_1 is used to train the pre-trained model. It contains training data $D_{1_{train}}$ and testing data $D_{1_{test}}$. D_2 is used to further fine-turn the pre-trained model. The D_2 also contains training data $D_{2_{train}}$ and testing data $D_{2_{test}}$. The model generalization is tested on D_2. The $\mathcal{M}_p$ is trained on D_1 and it transfers learned knowledge on D_2. In training data, each image x has a ground-truth density map y. The $\mathcal{M}_p$ gets an estimated density map for x.

Considering the expensive of the dataset, the pattern of few-shot learning is adopted in this method. The training data D_1 contains N tasks. The testing data D_2 contains M tasks. The number of tasks in D_1 and D_2 are not the same. Each training task is defined as $D_1^i, (i \in N)$. Each testing task is defined as $D_2^i, (i \in M)$. Each of training scene has several Q labeled images. Reddy et al. [4] randomly sample a small number $K \in \{1,5\}$ and $k \ll Q$ labeled images for each task. This work also follows this setup. Few-shot learning means to train the model with a few samples. It reflects the real-world scene of having to learn the model from a few labeled images. Such works also save the cost of labeling images.

In the meta training process, pre-trained model $\mathcal{M}_p$ learns a serious of parameters $\vartheta_N = \{\vartheta_1, \vartheta_2, \ldots, \vartheta_n\}$. These parameters which are learned by convolutional neural network reflect the feature mapping of crowd. When learning to adapt to a new real-world scene, these parameters are updated. From Fig. 1 shows, the training data $D_{1_{train}}$ are sent into DMNet to estimate the crowd density map. DMNet is a backbone, which will be introduced in Section 4. After DMNet, the meta optimization will optimize the loss function. The training parameters ϑ_N is updated in each training task D_1^i, as Eq. (1) shows,

$$\vartheta_N^{'} = \vartheta_N - \alpha \nabla_{\vartheta_N} \mathcal{L}_{D_1^i}\left(\mathcal{M}_p\right). \tag{1}$$

This optimization happened in inner update. α is the step size, which is fixed as a hyper-parameter. α is the meta step size.$\nabla_{\vartheta_N} \mathcal{L}_{D_1^i}\left(\mathcal{M}_p\right)$ is the gradient of the loss function in training task D_1^i. This work chooses the L2 loss function $loss()$ to finish the update. For image x, its ground ground-truth density map is y. The estimated density map $\mathcal{M}_p(y)$ and ground truth are calculated as Eq. (2),

$$loss\left(\mathcal{M}_p\right) = \sum\nolimits_{(x_i,y_i)D_1} \|\mathcal{M}_p(y_i) - y_i\|^2. \tag{2}$$

The parameters ϑ_N are trained by optimizing for the performance of $\mathcal{M}_p$. The updated parameters are then sent into next training step. In the training inner loop, the parameters of the model will be updated until the network is convergent. It represents that the pre-trained model $\mathcal{M}_p$ has a good generalization performance when facing the similar scenes.

Meta testing process After training inner loop, the pre-trained model $\mathcal{M}_p$ is fine-turned with DMNet in meta testing process. In testing step, the new target task D_2^i also has a few K images for training and other images are for testing. The proposed method is a learning-to-learn algorithm. The testing data is different form the training data, which contains different scenes. It quickly adapts the $\mathcal{M}_p$ to $\mathcal{M}_p^{test}$. It has a loss function to train the $\mathcal{M}_p^{test}$, as Eq. (3) shows.

$$loss\left(\mathcal{M}_p^{test}\right) = \sum\nolimits_{(x_i,y_i)D_2} \|\mathcal{M}_p^{test}(y_i) - y_i\|^2. \tag{3}$$

The meta testing optimization across tasks is dealt with Stochastic Gradient Descent (SGD). The updated parameters perform well on test images. The optimization process is defined as Eq. (4). β is the meta step size.

$$\vartheta_N = \vartheta_N - \beta \nabla_{\vartheta_N} \mathcal{L}_{D_2^i}\left(\mathcal{M}_p^{test}\right). \tag{4}$$

After the meta training and meta testing processes, this crowd counting method achieves the scene-adaptive crowd counting. In addition, because the DMNet is able to handle the low-lighting scenes, this method has better scene-adaptive capacity, especially in complex illumination conditions.

4 The proposed Dual-illumination Merging Network (DMNet)

4.1 Overview of the DMNet

The DMNet is designed as backbone in proposed scene-adaptive crowd counting method with meta-learning. The Fig. 2 shows the structure of the DMNet. This network is designed as a dual inputs structure. The initial input image is dealt with illumination-enhanced approach. The enhanced image and the initial image are sent into the DMNet. DMNet mainly contains two submodules, the Multi-scale Feature Extraction (MFE) module and the Element-wise Fusion Module (EFM). The MFE is used to extract the image feature on the multiple scales, which helps to improve network accuracy. The EFM fuses the low-lighting feature and illumination-enhanced feature, which is beneficial to extract the illumination feature in low-lighting environments. The two modules can also be applied to classification or object detection approaches, which can improve the detection accuracy and handle multi-scale problems.

4.2 The process of dataset illumination enhanced

The low-lighting images always cannot be proceeded by image processing methods. Such as the initial image in Fig. 2, the crowd is hidden in the darkness. These parts cannot be detected by neural network. To solve this problem, this work uses the illumination enhanced approach to recover the illumination information in darkness. The Fig. 3 shows this process.

The initial image is firstly decomposed in RGB space. Then, the image is transformed from RGB to YUV with Eq. (5). The value of Y represents the illumination grad of the image. If the value of Y is lower than $\sigma \in (10,90)$, it means that this image is taken in low-lighting scene.

$$Y = 0.30R + 0.59G + 0.11B. \tag{5}$$

The other images are remaining unchanged and two same images are sent into next step. This image will be enhanced with DUAL [35]. The DUAL is an automatic exposure correction method to enhance the image's illumination. It can firstly generate two intermediate exposure correction results, i.e., the underexposed regions and the overexposed regions, for the input image. Then, a multi-exposure image fusion technique is employed to adaptively blend the visually best exposed parts in the two intermediate exposure correction images and the input image into a global image [35]. To choose the best model, this work also tested with other three methods [36,37,38]. Figure 4 shows the tested results.

Because the results mainly rely on visual comparison [35], from Fig. 4, it is found that the DUAL has the best illumination enhanced performance. The quality of the images after processing by these methods is very different, which can be found in Fig. 4. The methods KinD [36] and RetinexNet [37] changed the style of the images, just like the cartoon style. And some subtle features of image are lost after LIMI [38]. So, this work chooses the DUAL to enhance the illumination of the crowd dataset.

4.3 The proposed MFE

After the illumination enhanced process, the initial image and the illumination enhanced image are sent into the frontend network. The frontend network mainly contains convolutional layers (*Conv.*) and ReLU layers (*ReLu.*). The dimension of the input image is $(1, w_1, h_1, N)$. The dimension of the output feature is (n, w_2, h_2, M). The processing can be represented as $ReLu(Conv(1, w_1, h_1, N))$. It is consisted of convolutional layers and ReLU layers. The images which are not changed with illumination enhanced step are also sent their copies into frontend network. After the frontend network, the initial image's feature f_i and enhanced image's feature f_e are got. Then, the f_i is dealt with the proposed MFE. Figure 5 shows the structure of the MFE.

This module extracts the image feature f_i from the multi-scale channels. The MFE outputs the feature $f_{i_{MFE}}$. The whole process is presented as Eq. (6).

$$f_{i_{MFE}} = \sum\nolimits_{k*k} U\left(\mathcal{F}_{3*3}(\mathcal{F}_{k*k}(f_i))\right), k \in (2,3,5). \tag{6}$$

The f_i is put into three different convolutional layers $\mathcal{F}_{k*k}$ with different kernel sizes. Because the illumination enhanced methods maybe fail to handle some conditions, the enhanced process will also break the initial image feature. To keep the initial image feature, the MFE is dealt with f_i, instead of f_e.

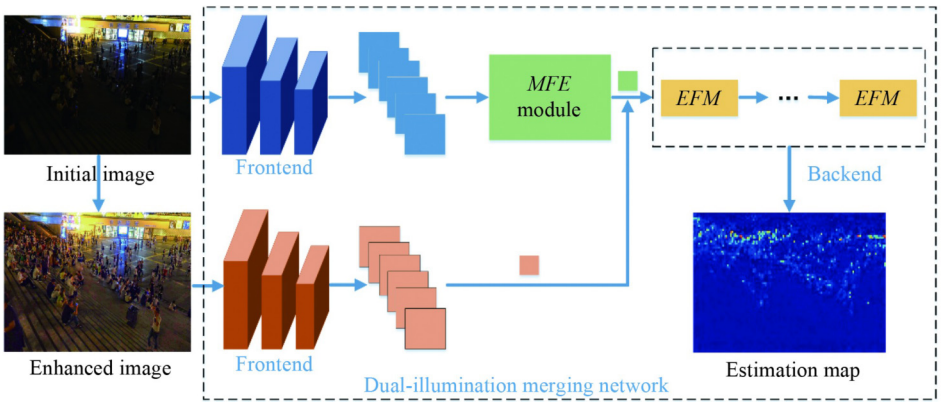


Fig. 2 The structure of the proposed DMNet

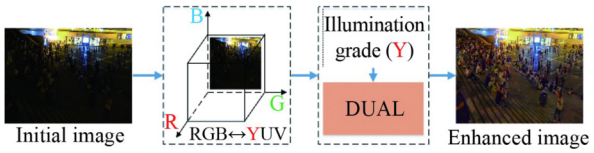


Fig. 3 The process of image illumination enhanced

Such structure is also used in some previous works, e.g., [9,10,15,18]. These convolutional layers have different receptive fields, which can extract features and avoid leaving out significant information. This work chooses three sizes kernels, i.e., $2*2$, $3*3$, and $5*5$. The $\mathcal{F}_{2*2}(f_i)$, $\mathcal{F}_{3*3}(f_i)$, and $\mathcal{F}_{5*5}(f_i)$ represent the convolutional results. The three features are sent into $3*3$ convolutional operations $\mathcal{F}_{3*3}()$ to further extract the image feature. This operation can help the network to express the high-level features. Because convolutional function reduces the feature channel, to recover the feature size, this work adds the up-sampling operation $U()$ behind the convolutional layers. The three channels' feature are added to get the output $f_{i_{MFE}}$.

4.4 The proposed EFM

The initial image after frontend is dealt with MFE to get the feature $f_{i_{MFE}}$. Then, $f_{i_{MFE}}$ and the enhanced image feature f_e are sent into EFM. EFM module fuses the illumination feature and the initial image feature. Due to lacking illumination information in initial image, this work tries to extract missing feature with EFM. The initial image lacks of illumination information and the illumination enhanced image maybe break the initial image feature. Due to the small size of the people in crowd, the subtle feature of people would be lost in illumination enhanced process. So, the EFM aims to fuse the initial image feature and the illumination enhanced image's feature. The structure of EFM is shown in Fig. 6.

The EFM has two inputs, i.e., $f_{i_{MFE}}$ and f_e. The two features are dealt by two convolutional layers with dilation rate 2. The dilated convolution can enlarge the receptive field, meanwhile avoiding the loss of information in pooling operation. Each output of the dilate convolution contains a wide range of information. Then, the Batch Normalizing (BN) is used to accelerate the training speed and prevent the gradient from disappearing of $f_{i_{MFE}}$. To get the fusion feature $f_{i_{MFE}+e}$, this work defines the fusion process as Eq. (7).

$$f_{i_{MFE}+e} = ReLU(f_{i_{MFE}} - ((f_{i_{MFE}} \odot f_e) \oplus f_e)). \quad (7)$$

$\odot$ represents the Hadamard product. The corresponding elements do multiply operation in $f_{i_{MFE}}$ and f_e. Then, the $(f_{i_{MFE}} \odot f_e)$ adds the f_e to fuse the illumination feature. To improve the convergence speed of the network and reduce the parameters, the $f_{i_{MFE}} - ((f_{i_{MFE}} \odot f_e) \oplus f_e)$ represents the feature after fusion step. The final output $f_{i_{MFE}+e}$ is got by ReLU. The new $f_{i_{MFE}+e}$ and the f_e are sent into next EFMs. The operation in Eq. (7) fuses the initial image and the illumination enhanced image through the element-wise operations. Some other methods tried such network structure to improve the accuracy of the model, such as [10]. And the experimental results also show that the influence of the EFM module.

In DMNet, this work uses several EFMs to fuse the illumination feature. Some experiments are tested with different number of EFMs. In the final EFM, the $f_{i_{MFE}+e}$ is dealt with backend network to generate the estimated density map. The backend network is an up-sampling network which

Fig. 4 The illumination enhanced results with different approaches. The first row is the initial images. The second row is the results with KinD [36]. The third row is the results with RetinexNet [37]. The fourth row is the results with LIMI [38]. The fifth row is the results with DUAL [35]

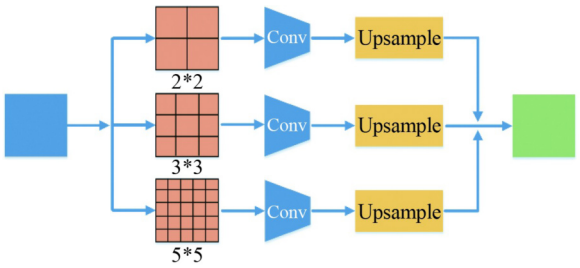


Fig. 5 The structure of the proposed the Multi-scale Feature Extraction module

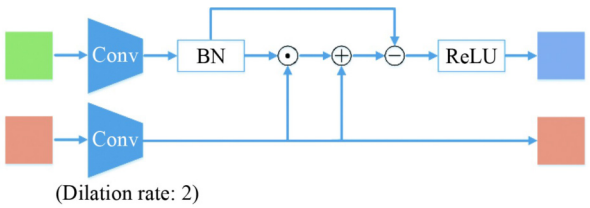


Fig. 6 The structure of the proposed the Element-wise Fusion Module

also contains convolutional layers and ReLu layers. The output feature maps of all convolutional layers are stacked and they are mapped to a density map. And the number of the crowd can be got by integrating the density map. To train the DMNet, the L2 loss function is used.

5 Experiments

The scene-adaptive crowd counting by DMNet with meta-learning is implemented under the Windows 10 and Pytorch 1.4.0 experimental environment. The hardware environments are Inter Xeon E-2136 3.3 GHz and Quadro P5000.

5.1 Evaluation metrics

Two standard evaluation metrics to test the proposed method is used, i.e., Mean Absolute Error (MAE) and the Root Mean Squared Error (RMSE) [2,3]. The evaluation metrics are defined as Eqs. (8) and (9).

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}\left|y_i - \widehat{y_i}\right|, \tag{8}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \widehat{y_i})^2}. \tag{9}$$

The N represents the total number of the test images, y_i is the ground-truth number of people inside the whole i image and $\widehat{y_i}$ is the estimated number of people.

5.2 Comparison of the proposed method on benchmarks

In this section, this work illustrates the influence of the illumination information for crowd counting. The efficiency of the proposed scene-adaptive crowd counting method is also proved.

The influence of the illumination information The initial image and the illumination enhanced image are sent into frontend network to get the features. The frontend network consists convolutional layers and ReLU layers. This work does the visualization operation of the feature in frontend network, as Fig. 7 shows.

The first column is the initial image, the second column is the illumination enhanced image, and the third column is the illumination enhanced image with gray operation. The red part in second row is the activate crowd feature. The bule part in second row means the negative crowd feature. The blue part would not be detected with DMNet. It is found that the crowd in darkness is not activated. When enhancing the illumination of the image, the crowd in darkness is detected and activated. To prove the efficiency of the illumination information, the illumination enhanced image is also be grayed as the third image in first row. Note that the grayed image is not used for training and just used to prove the efficiency of the illumination enhanced image. It is found that the crowd in darkness can also be detected. It illustrates the significant influence of the illumination information. So, this work adopts the DUAL to recover such illumination information and proposes illumination feature fusion module to express the image feature better.

The performances of the proposed method To test the performance of the proposed scene-adaptive crowd counting by DMNet with meta-learning, this work tests four benchmarks, i.e., WorldExpo's10 [6], DISCO [14], UCSD [34], and Mall [7] datasets, as Fig.8 shows.

The WorldExpo's10 contains of 3,980 labeled images from 1,132 video sequences based on 108 different scenes. Following the setup of the "Meta+CSRNet" [4], this work uses 103 scenes for meta training and the remaining five scenes for meta testing. The DISCO contains 1,935 images and 170,270 annotated instances from 65 different scenes. As the black dotted box in Fig. 8 shows, each scene has many images with different illumination. It makes this dataset is more challenging. This work considers 60 scenes for meta training and the remaining five scenes for meta testing. UCSD and Mall datasets both only contain one scene. The Mall includes 2,000 images from the same camera in a mall. The UCSD consists 2000 images from the same camera to capture a pedestrian scene. This work uses $K \in \{1,5\}$ images to train the model on these benchmarks. It is assumed that the number of labeled training data is small, which aims to finish this task with a few data. The number one and five are typical values, which is also used in other works, such as "Meta+CSRNet".

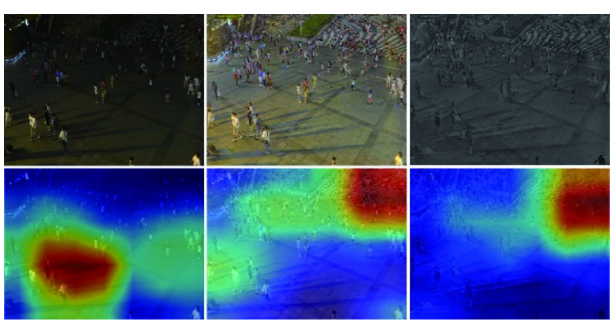


Fig. 7 The visualization results of the feature in frontend network. The first column is the result of image in darkness. The second column is the result of illumination enhanced image. The third column is the result of illumination enhanced image with gray operation

Fig. 8 Some examples of four different datasets. The images in black dotted box are from DISCO. The images in red dotted box are fromWorldExpo'10. The images in orange dotted box are from UCSD. The images in blue dotted box are from Mall

To prove the efficiency of the proposed method, two main methods are compared. "Meta+CSRNet" also proposed a scene-adaptive method for crowd counting. "Meta+CSRNet" used CSRNet [39] as backbone to train the model. In addition, this work replaces the CSRNet with CANNet [40] in "Meta+CSRNet" for comparing. The "Meta+CSRNet" is defined as "Meta+CSRNet". "Meta+CSRNet" with CANNet is defined as "Meta+CANNet". Our work is defined as "Meta+DMNet". The pre-trained model of Meta+CANNet is defined as "CANNet Pre-", which means it does not fine-turned in meta-testing process. The pre-trained model of "Meta+DMNet" is defined as "DMNet Pre-". These methods are tested with 1-shot and 5-shot. 1-shot means only one image of each task for training. 5-shot means five images of each task for training.

Table 1 shows the performances of these methods with 1-&5- shots in five scenes on WorldExpo's10. From the Table 1, it is found that the proposed scene-adaptive crowd counting method gets better performances than others. On average, the proposed method gets 19.88 with MAE and 26.49 with RMSE in 5-shot settings. "Meta+CSRNet" gets 23.94 with MAE and 32.01 with RMSE. In addition, it is found that model with 1-shot is sometimes better than model with 5-shot, such as the "Meta+CANNet" in scene 2. The "Meta+CANNet" model gets 18.69 with 5-shot and gets 17.60 with 1-shot in scene 2. Such results are also tested in the paper of "Meta+CSRNet". It is inferred that more data makes the parameters learned by the model ineffective in some scenes. But as a whole, more training data tends to lead to better results [41,42].

Table 2 shows the performances of these methods with 1-&5- shots in five scenes on DISCO. DISCO has more scenes and the images are more challenging. On average, the proposed method gets 50.80 with MAE and gets 72.99 with RMSE. The "Meta+CANNet" gets 58.05 with MAE and 77.80 with RMSE. The accuracy of the proposed method is 7.25 with MAE higher than that of "Meta+CANNet" method. "Meta+CSRNet" gets 74.02 with MAE and 97.01 with RMSE. The accuracy of the proposed method is 23.22 with MAE higher than that of the method in "Meta+CSRNet".

Table 3 shows the performances of these methods with 1-&5- shots in five scenes on UCSD dataset. Table 4 shows the performances of these models with 1-&5- shots in five scenes on Mall dataset. The UCSD and the Mall both contain single scenes. The proposed method gets 4.01 with MAE and gets 5.17 with RMSE in UCSD. The 3.05 with MAE and 4.13 with RMSE are got in Mall dataset. It is also found that the "Meta+CSRNet" and the "Meta+CSRNet" also get high accuracy on UCSD and Mall dataset.

Note that Tables 1−4 mainly tested on different dataset, i.e., WorldExpo's10, DISCO, UCSD, and Mall. The scenes in these datasets are very different. So, it is reasonable for one method performs different in these datasets and this work does not compare the performance cross these tables. And compared to other approaches, the method which this paper proposed, can get satisfactory results.

To prove the generalization capability of the proposed model, this work also tested the images of scene 1−5 in Table 1 with three traditional crowd counting models, MSR-FAN [15], CSRNet [39] and the CANNet [40]. The generalization capability represents that the trained model can get good performance on new scenes [43]. The MSR-FAN, CSRNet, and CANNet are three general crowd counting method, which have been verified that have good performances. Table 5 shows the comparison results. It is found that the proposed method can get best performance on average. The "Meta+DMNet" gets 19.88 with MAE and gets 26.49 with RMSE. It can prove the proposed method has a good generalization capability.

Table 1 The performances of different models with 1-&5- shots in five scenes on WorldExpo's10

Target	Methods	1-shot (K=1)		5-shot (K=5)	
		MAE	RMSE	MAE	RMSE
Scene 1	Meta+CSRNet	19.57	26.17	21.79	28.35
	CANNet Pre-	20.72	21.96	21.49	24.27
	Meta +CANNet	18.69	22.20	18.83	22.34
	DMNet Pre-	18.96	22.64	20.66	23.89
	Ours	**18.45**	**21.56**	**18.14**	**21.19**
Scene 2	Meta +CSRNet	20.39	31.86	20.77	32.44
	CANNet Pre-	20.67	23.17	23.06	25.09
	Meta +CANNet	17.60	28.54	18.69	**22.20**
	DMNet Pre-	22.55	27.23	18.57	29.77
	Ours	**16.65**	**27.61**	**16.66**	27.48
Scene 3	Meta +CSRNet	16.82	27.11	20.94	31.51
	CANNet Pre-	26.66	32.51	18.39	28.83
	Meta +CANNet	16.04	25.52	16.13	24.75
	DMNet Pre-	24.28	28.70	19.21	29.09
	Ours	**16.47**	**25.29**	**16.09**	**24.69**
Scene 4	Meta +CSRNet	22.63	27.33	21.35	25.53
	CANNet Pre-	22.61	24.01	22.83	26.88
	Meta +CANNet	21.52	27.41	21.34	27.00
	DMNet Pre-	22.69	23.98	23.27	25.75
	Ours	**21.30**	**23.76**	**21.18**	**25.30**
Scene 5	Meta +CSRNet	27.49	33.82	34.86	42.22
	CANNet Pre-	29.35	34.88	28.94	34.72
	Meta +CANNet	28.36	34.80	28.36	34.79
	DMNet Pre-	30.49	36.12	29.63	35.95
	Ours	**27.10**	**33.45**	**27.35**	**33.80**
Average	Meta +CSRNet	21.38	29.26	23.94	32.01
	CANNet Pre-	24.00	27.30	22.94	27.96
	Meta +CANNet	20.44	27.69	20.64	**26.21**
	DMNet Pre-	23.79	27.73	22.27	28.89
	Ours	**19.99**	**26.33**	**19.88**	26.49

Table 2 The performances of different models with 1-&5- shots in five scenes on DISCO

Target	Methods	1-shot (K=1)		5-shot (K=5)	
		MAE	RMSE	MAE	RMSE
Scene 1	Meta +CSRNet	153.65	210.45	149.91	200.15
	CANNet Pre-	132.46	160.22	129.71	156.56
	Meta +CANNet	126.40	161.32	127.58	160.68
	DMNet Pre-	125.02	154.78	121.92	150.59
	Ours	**106.30**	**143.61**	**104.40**	**141.75**
Scene 2	Meta +CSRNet	66.39	103.46	64.08	100.70
	CANNet Pre-	83.56	109.78	86.17	114.83
	Meta +CANNet	45.36	96.19	42.53	89.44
	DMNet Pre-	57.64	**70.29**	54.26	**71.30**
	Ours	**40.14**	86.72	**42.50**	89.16
Scene 3	Meta +CSRNet	**53.64**	**64.29**	**50.54**	**61.30**
	CANNet Pre-	65.72	74.20	67.51	74.36
	Meta +CANNet	63.18	70.45	61.58	69.25
	DMNet Pre-	58.37	67.60	60.55	70.12
	Ours	54.15	67.00	59.49	68.31
Scene 4	Meta +CSRNet	34.64	43.89	33.31	42.11
	CANNet Pre-	35.03	41.78	35.56	42.86
	Meta +CANNet	31.25	39.32	30.86	38.56
	DMNet Pre-	28.33	**33.14**	29.57	34.97
	Ours	**27.01**	36.78	**26.02**	**32.56**
Scene 5	Meta +CSRNet	78.11	86.34	72.26	80.77
	CANNet Pre-	40.90	51.37	38.88	48.31
	Meta +CANNet	24.19	**29.06**	27.72	**31.05**
	DMNet Pre-	33.71	42.50	32.24	41.76
	Ours	**20.15**	32.99	**21.58**	33.17
Average	Meta +CSRNet	77.29	101.69	74.02	97.01
	CANNet Pre-	71.53	87.47	71.57	87.38
	Meta +CANNet	58.08	79.27	58.05	77.80
	DMNet Pre-	60.61	73.66	59.71	73.75
	Ours	**49.55**	**73.42**	**50.80**	**72.99**

5.3 Ablation studies

To prove the performance of the proposed method, this part analyzes the efficient of the MFE module and how many EFM modules is best. The proposed method without MFE module is denoted as "Meta-DMNet-no-MFE". The Table 6 shows the performance between the "Meta-DMNet-no-MFE" and "Meta-DMNet" on UCSD dataset. The proposed method without MFE module gets 9.43 with MAE and gets 11.75 with RMSE. The proposed method gets 4.01 with MAE and gets 5.17 with RMSE, which is better than "Meta-DMNet-no-MFE".

In addition, this work also explores how many EFMs performs best on UCSD dataset. The models with number of $\{2,3,4,5,6\}$ are defined as "Meta-DMNet-2", "Meta-DMNet-3", "Meta-DMNet-4", "Meta-DMNet-5", and "Meta-DMNet-6". Table 7 shows the experimental results. From the Table 7, it is found that the "Meta-DMNet-4" gets the best performance. So, the number of EFM module in this work is set as four. The "Meta-DMNet-4" gets the 4.01 with MAE and 5.17 with RMSE, which outperforms other models. It is assumed that the performances of the proposed method with different EFMs fit the Gaussian distribution. Too many EFM modules will lead to the loss of the useful feature, which will decrease the accuracy of the proposed method.

Table 3 The performances of different models with 1-&5- shots on UCSD

Methods	1-shot (K=1)		5-shot (K=5)	
	MAE	RMSE	MAE	RMSE
Meta +CSRNet	5.63	6.92	5.25	6.63
CANNet Pre-	4.59	5.48	4.21	5.37
Meta +CANNet	4.26	5.42	**3.95**	**4.92**
DMNet Pre-	4.31	5.14	4.26	5.26
Ours	**4.18**	**4.92**	4.01	5.17

Table 4 The performances of different models with 1-&5- shots on Mall

Methods	1-shot (K=1)		5-shot (K=5)	
	MAE	RMSE	MAE	RMSE
Meta +CSRNet	4.69	5.56	4.52	5.36
CANNet Pre-	3.62	4.51	3.36	4.20
Meta +CANNet	3.23	**4.45**	3.17	4.39
DMNet Pre-	3.50	4.67	3.11	**4.02**
Ours	**3.19**	4.49	**3.05**	4.13

6 Conclusion

This study proposes a scene-adaptive crowd counting method based on meta-learning with Dual-illumination Merging Network (DMNet) to adapt different real-world scenes which only contain a few labeled images. The proposed method is mainly based on meta-learning, which can lead to fast learning on a new real-world scene. To generate high quality density map and count the crowd in low-lighting scene, the DMNet is proposed, which contains two submodules, Multi-scale Feature Extraction module and Element-wise Fusion Module. The Multi-scale Feature Extraction module is used to extract

Table 5 The performances of different crowd counting models on scene 1−5 in Table 1

Methods	Evaluation	CSRNet [39]	CANNet [40]	MSR-FAN	**Ours**
Scene 1	MAE	24.62	19.44	**14.38**	18.14
	RMSE	29.21	22.59	**17.25**	21.19
Scene 2	MAE	25.14	19.91	**15.28**	16.65
	RMSE	29.63	30.69	**26.86**	27.61
Scene 3	MAE	28.70	22.07	22.19	**16.09**
	RMSE	39.21	31.49	33.27	**24.69**
Scene 4	MAE	20.72	**10.72**	13.19	21.18
	RMSE	26.45	**13.89**	17.55	25.30
Scene 5	MAE	46.89	39.38	37.38	**27.10**
	RMSE	50.10	46.52	46.44	**33.45**
Average	MAE	29.21	22.30	20.48	**19.88**
	RMSE	34.92	29.04	28.27	**26.49**

Table 6 The influence of MFE module in the proposed method on UCSD dataset

Methods	1-shot (K=1)		5-shot (K=1)	
	MAE	RMSE	MAE	RMSE
Meta-DMNet-no-MFE	10.76	13.49	9.43	11.75
Meta-DMNet	**4.18**	**4.92**	**4.01**	**5.17**

Table 7 The experimental results of the proposed method with different EFMs

Methods	1-shot (K=1)		5-shot (K=1)	
	MAE	RMSE	MAE	RMSE
Meta-DMNet-2	8.36	10.48	9.42	11.25
Meta-DMNet-3	6.46	7.52	6.02	7.61
Meta-DMNet-4	**4.18**	**4.92**	**4.01**	**5.17**
Meta-DMNet-5	4.63	5.28	4.50	5.29
Meta-DMNet-6	4.50	5.11	4.96	5.33

the image feature by multi-scale convolutions, which helps to improve network accuracy. The Element-wise Fusion module fuses the low-lighting feature and illumination-enhanced feature, which supplements the missing illumination in low-lighting environments. Experimental results on benchmarks, WorldExpo'10, DISCO, USCD, and Mall, show that the proposed method outperforms the existing state-of-the-art methods in accuracy and can get satisfied results.

In the future, the more advanced meta-learning method will be applied to crowd counting. The unsupervised methods will be considered.

Acknowledgements This work was supported by the National Natural Science Foundation of China (Grant Nos. 62076117 and 61762061), the Natural Science Foundation of Jiangxi Province, China (20161ACB20004) and Jiangxi Key Laboratory of Smart City (20192BCD40002).

References

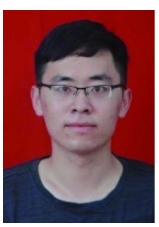
1. Wang Q, Gao J, Lin W, Li X. NWPU-crowd: a large-scale benchmark for crowd counting and localization. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 43(6): 2141–2149
2. Liu Y, Wen Q, Chen H, Liu W, Qin J, Han G, He S. Crowd counting via cross-stage refinement networks. IEEE Transactions on Image Processing, 2020, 29: 6800–6812
3. Gao J, Wang Q, Li X. PCC Net: perspective crowd counting via spatial convolutional network. IEEE Transactions on Circuits and Systems for Video Technology, 2020, 30(10): 3486–3498
4. Reddy M K K, Hossain M A, Rochan M, Wang Y. Few-shot scene adaptive crowd counting using meta-learning. In: Proceedings of the 2020 IEEE Winter Conference on Applications of Computer Vision (WACV). 2020, 2803−2812
5. Liu X, Van De Weijer J, Bagdanov A D. Leveraging unlabeled data for crowd counting by learning to rank. In: Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2018, 7661−7669
6. Zhang C, Li H, Wang X, Yang X. Cross-scene crowd counting via deep convolutional neural networks. In: Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2015, 833−841
7. Loy C C, Gong S, Xiang T. From semi-supervised to transfer counting of crowds. In: Proceedings of the 2013 IEEE International Conference on Computer Vision. 2013, 2256−2263
8. Finn C, Abbeel P, Levine S. Model-agnostic meta-learning for fast adaptation of deep networks. In: Proceedings of the 34th International Conference on Machine Learning. 2017, 1126−1135
9. Zhao M, Zhang C, Zhang J, Porikli F, Ni B, Zhang W. Scale-aware crowd counting via depth-embedded convolutional neural networks. IEEE Transactions on Circuits and Systems for Video Technology, 2020, 30(10): 3651–3662
10. Fang Y, Gao S, Li J, Luo W, He L, Hu B. Multi-level feature fusion based Locality-Constrained Spatial Transformer network for video crowd counting. Neurocomputing, 2020, 392: 98–107
11. Sam D B, Peri S V, Sundararaman M N, Kamath A, Babu R V. Locate, size, and count: accurately resolving people in dense crowds via detection. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 43(8): 2739–2751
12. Liu L, Lu H, Xiong H, Xian K, Cao Z, Shen C. Counting objects by blockwise classification. IEEE Transactions on Circuits and Systems for Video Technology, 2020, 30(10): 3513–3527
13. Wu X, Zheng Y, Ye H, Hu W, Ma T, Yang J, He L. Counting crowds with varying densities via adaptive scenario discovery framework. Neurocomputing, 2020, 397: 127–138
14. Hu D, Mou L, Wang Q, Gao J, Hua Y, Dou D, Zhu X X. Ambient sound helps: audiovisual crowd counting in extreme conditions. 2020, arXiv preprint arXiv: 2005.07097
15. Zhao H, Min W, Wei X, Wang Q, Fu Q, Wei Z. MSR-FAN: multi-scale residual feature-aware network for crowd counting. IET Image Processing, 2021, 15(14): 3512–3521
16. Zheng H, Lin Z, Cen J, Wu Z, Zhao Y. Cross-line pedestrian counting based on spatially-consistent two-stage local crowd density estimation and accumulation. IEEE Transactions on Circuits and Systems for Video Technology, 2019, 29(3): 787–799
17. Shen Z, Xu Y, Ni B, Wang M, Hu J, Yang X. Crowd counting via adversarial cross-scale consistency pursuit. In: Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2018, 5245−5254
18. Yang B, Zhan W, Wang N, Liu X, Lv J. Counting crowds using a scale-distribution-aware network and adaptive human-shaped kernel. Neurocomputing, 2020, 390: 207–216
19. Zou Z, Cheng Y, Qu X, Ji S, Guo X, Zhou P. Attend to count: crowd counting with adaptive capacity multi-scale CNNs. Neurocomputing, 2019, 367: 75–83
20. Wang L, Yin B, Tang X, Li Y. Removing background interference for crowd counting via de-background detail convolutional network. Neurocomputing, 2019, 322: 360–371
21. Chen J, Wang Z. Crowd counting with segmentation attention convolutional neural network. IET Image Processing, 2021, 15(6): 1221–1231
22. Jiang S, Lu X, Lei Y, Liu L. Mask-aware networks for crowd counting. IEEE Transactions on Circuits and Systems for Video Technology, 2020, 30(9): 3119–3129

23. Min W, Fan M, Guo X, Han Q. A new approach to track multiple vehicles with the combination of robust detection and two classifiers. IEEE Transactions on Intelligent Transportation Systems, 2018, 19(1): 174–186
24. Yang H, Liu L, Min W, Yang X, Xiong X. Driver yawning detection based on subtle facial action recognition. IEEE Transactions on Multimedia, 2020, 23: 572–583
25. Wang Q, Min W, He D, Zou S, Huang T, Zhang Y, Liu R. Discriminative fine-grained network for vehicle re-identification using two-stage re-ranking. Science China Information Sciences, 2020, 63(11): 212102
26. Ma Y, Zhong G, Liu W, Wang Y, Jiang P, Zhang R. ML-CGAN: conditional generative adversarial network with a meta-learner structure for high-quality image generation with few training data. Cognitive Computation, 2021, 13(2): 418–430
27. Jung I, You K, Noh H, Cho M, Han B. Real-time object tracking via meta-learning: efficient model adaptation and one-shot channel pruning. In: Proceedings of the 34th AAAI Conference on Artificial Intelligence. 2020, 11205−11212, doi: 10.1609/aaai.v34i07.6779
28. Elsken T, Staffler B, Metzen J H, Hutter F. Meta-learning of neural architectures for few-shot learning. In: Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2020, 12362−12372
29. Xu C, Shen J, Du X. A method of few-shot network intrusion detection based on meta-learning framework. IEEE Transactions on Information Forensics and Security, 2020, 15: 3540–3552
30. Ye H J, Sheng X R, Zhan D C. Few-shot learning with adaptively initialized task optimizer: a practical meta-learning approach. Machine Learning, 2020, 109(3): 643–664
31. Nichol A, Achiam J, Schulman J. On first-order meta-learning algorithms. 2018, arXiv preprint arXiv: 1803.02999v3
32. Wang D, Cheng Y, Yu M, Guo X, Zhang T. A hybrid approach with optimization-based and metric-based meta-learner for few-shot learning. Neurocomputing, 2019, 349: 202–211
33. Lai N, Kan M, Han C, Song X, Shan S. Learning to learn adaptive classifier–predictor for few-shot learning. IEEE Transactions on Neural Networks and Learning Systems, 2021, 32(8): 3458–3470
34. Chan A B, Liang Z S J, Vasconcelos N. Privacy preserving crowd monitoring: counting people without people models or tracking. In: Proceedings of the 2008 IEEE Conference on Computer Vision and Pattern Recognition. 2008, 1−7
35. Zhang Q, Nie Y, Zheng W S. Dual illumination estimation for robust exposure correction. Computer Graphics Forum, 2019, 38(7): 243–252
36. Zhang Y, Zhang J, Guo X. Kindling the darkness: a practical low-light image enhancer. In: Proceedings of the 27th ACM International Conference on Multimedia. 2019, 1632−1640
37. Wei C, Wang W, Yang W, Liu J. Deep Retinex decomposition for low-light enhancement. 2018, arXiv preprint arXiv: 1808.04560
38. Guo X, Li Y, Ling H. LIME: low-light image enhancement via illumination map estimation. IEEE Transactions on Image Processing, 2017, 26(2): 982–993
39. Li Y, Zhang X, Chen D. CSRNet: dilated convolutional neural networks for understanding the highly congested scenes. In: Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2018, 1091−1100
40. Liu W, Salzmann M, Fua P. Context-aware crowd counting. In: Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2019, 5094-5103
41. Chu J, Guo Z, Leng L. Object detection based on multi-layer convolution feature fusion and online hard example mining. IEEE Access, 2018, 6: 19959–19967
42. Zhang Y, Chu J, Leng L, Miao J. Mask-Refined R-CNN: a network for refining object details in instance segmentation. Sensors, 2020, 20(4): 1010
43. Zhang Y, Zhou D, Chen S, Gao S, Ma Y. Single-image crowd counting via multi-column convolutional neural network. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016, 589−597

Haoyu Zhao obtained the BE degree of computer science and technology at Nanchang University in China in 2019. He is a post-graduate at Nanchang University in China now. His research interests include computer vision and deep learning.

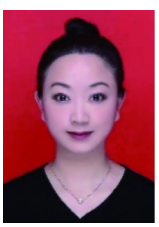
Weidong Min received the BE, ME and PhD degrees in computer application from Tsinghua University, China in 1989, 1991 and 1995, respectively. He is currently a Professor and the Dean, School of Software, Nanchang University, China. He is an Executive Director of China Society of Image and Graphics. His current research interests include image and video processing, artificial intelligence, big data, distributed system and smart city information technology.

Jianqiang Xu obtained the ME degree from Information Engineering School of Nanchang University, China in 2010. He is currently pursuing the PhD degree with the Information Engineering School of Nanchang University, China. His research interests include computer vision, pattern recognition, machine learning, computer image and video processing.



Qi Wang obtained the ME degree in computer science and technology from Nanchang University, China in 2017. He is currently pursuing the PhD degree at Nanchang University, China. His current research focuses on computer vision, particularly vehicle re-identification.

Yi Zou obtained the BE degree of computer science and technology at Nanchang University, China in 2021. She is a post-graduate at Nanchang University in China now. Her research interests include image processing and deep learning.



Qiyan Fu received the ME degree in Electronic and Communication Engineering from Nanchang University, China in 2017. She is currently pursuing the PhD degree at Nanchang University, China. Her current research focuses on artificial intelligence and computer vision.