TSINGHUA SCIENCE AND TECHNOLOGY
ISSN 1007-0214 09/09 pp869–879
DOI: 10.26599/TST.2021.9010097
Volume 27, Number 5, October 2022

# Object Counting Using a Refinement Network

Lehan Sun, Junjie Ma*, and Liping Jing

**Abstract:** To address the scale variance and uneven distribution of objects in scenarios of object-counting tasks, an algorithm called Refinement Network (RefNet) is exploited. The proposed top-down scheme sequentially aggregates multiscale features, which are laterally connected with low-level information. Trained by a multiresolution density regression loss, a set of intermediate-density maps are estimated on each scale in a multiscale feature pyramid, and the detailed information of the density map is gradually added through coarse-to-fine granular refinement progress to predict the final density map. We evaluate our RefNet on three crowd-counting benchmark datasets, namely, ShanghaiTech, UCF_CC_50, and UCSD, and our method achieves competitive performances on the mean absolute error and root mean squared error compared to the state-of-the-art approaches. We further extend our RefNet to cell counting, illustrating its effectiveness on relative counting tasks.

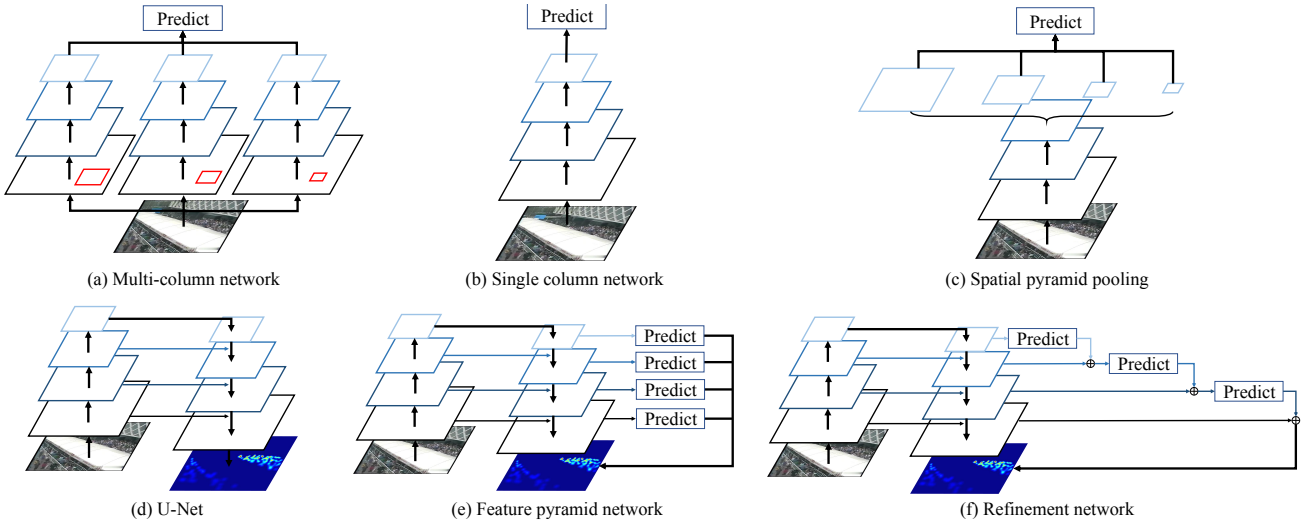
**Key words:** object counting; Refinement Network (RefNet); scale variation; uneven distribution

## 1 Introduction

As one of the important tasks in crowd scene analysis[1], crowd counting focuses on the crowd density distribution and the number of pedestrians. It helps to automatically monitor public scenarios, preventing incidents caused by extremely crowded situations. The past 10 years have witnessed the boosting development of Convolutional Neural Networks (CNNs), which have been utilized in several research aspects, such as image classification[2], face detection and alignment[3], semantic segmentation[4], and so forth. As a computer vision task, many crowd-counting approaches[5] are based on fully convolutional networks[6]. The counting accuracy has dramatically improved as compared to traditional methods. However, the crowd-counting task is still challenging mainly because of the scale variance and uneven distribution of crowds.

To deal with the scale variance issue, many crowd-counting approaches have been proposed. Multi-Column Neural Networks (MCNNs), briefly shown in Fig. 1a, are one of the widely used structures in recent crowd-counting algorithms[7–11]. A column with large kernel sizes obtains a large receptive field to extract people in large scales, whereas a column containing small kernel convolutional layers extracts small-scale heads in crowded areas. However, these scale-aware methods have several disadvantages, such as little feature sharing among columns and shallow structure property. Moreover, a single column scheme[12], briefly shown in Fig. 1b, is introduced to construct a deeper end-to-end network compared to MCNN. The method proposed by Li et al.[12] utilizes a set of atrous convolutions[13] as a regressor to enlarge receptive field, but fails to aggregate multiscale features. Inspired by the Spatial Pyramid Pooling (SPP, briefly shown in Fig. 1c)[14], an Atrous Spatial Pyramid Pooling (ASPP)[4, 15] was utilized in Ref. [16], together with a multi-column-like cascade backbone for crowd counting. Features are partially shared among columns with a deep scheme. Multiscale information is also considered in ASPP, whose structure is, however, still narrow for multiscale feature extraction.

• Lehan Sun is with School of Science, Beijing Jiaotong University, Beijing 100044, China. E-mail:19271102@bjtu.edu.cn.
• Junjie Ma is with Department of Computer Science and Technology, and Beijing National Research Center for Information Science and Technology (BNRist), Tsinghua University, Beijing 100084, China. E-mail: junjiema@tsinghua.edu.cn.
• Liping Jing is with School of Computer and Information Technology, Beijing Jiaotong University, Beijing 100044, China. E-mail:lpjing@bjtu.edu.cn.
* To whom correspondence should be addressed.
Manuscript received: 2021-11-03; revised: 2021-12-22; accepted: 2021-12-23




(a) Multi-column network (b) Single column network (c) Spatial pyramid pooling

(d) U-Net (e) Feature pyramid network (f) Refinement network

**Fig. 1 Variant structures for crowd counting.**

In addition, U-Net (Fig. 1d) has been applied to crowd counting task[17, 18], but it does not fully use multiscale semantic feature maps in the decoder module, which may be beneficial in the density map and counting prediction. Although the feature pyramid network[19] (Fig. 1e) would handle this problem, with the same issue as the above approaches, the uneven distribution issue is not well taken into account.

Therefore, different from the aforementioned methods, a Refinement Network (RefNet, Fig. 1f) is introduced for crowd counting in this paper. The scale aggregation bottom-up pathway is employed to improve the representation ability with more continuous multiscale feature maps without most global skip connections. The top-down pathway comprises a set of convolutions and upsampling layers to form a feature pyramid module, which represents multiscale semantic information to generate intermediate-density maps linked with low-level features by lateral connections to compensate for spatial information. Intermediate-density maps with different resolutions are sequentially generated and propagated to evaluate the final density map in a coarse-to-fine granular refinement manner, which would alleviate the uneven distribution issue. In addition, a multiresolution density regression loss is presented to gradually generate intermediate-density maps in the top-down pathway and to guide the final counting and density map regression in a coarse-to-fine manner. In the real-time cell-counting task, cells in microscopic images are usually not stained or evenly smeared. Problems that are similar to the crowd-counting task also exist, such as uneven distribution, irregular object shapes, and occlusion. We further evaluate RefNet on our collected epithelial cell dataset to illustrate its effectiveness in the cell-counting issue.

The remainder of this paper is arranged as follows: Section 2 surveys some recent literature works on crowd counting. Section 3 introduces the proposed RefNet in detail. Section 4 presents the evaluations and ablation study given by the experiments on benchmark datasets. Finally, Section 5 concludes the paper.

## 2 Related Work

As a scale-aware network, an MCNN with arbitrary input was proposed[7]. Each of its three branches corresponds to a particular head scale. Inspired by the results of a previous study[7], a patch-based switch classifier was designed in switching-CNN[8] to select an optimal operation among three independent scale-variant CNN regressors for input patches. By incorporating learning density levels as context information, a cascaded CNN was proposed by Sindagi and Patel[20] to boost the density estimation performance. Sindagi and Patel[9] further improved the network by concatenating both global and local context information with two separate subnetworks, concatenated with density features produced by MCNN, and finally regressed by a fusion CNN. The results showed the importance of context features for the crowd-counting task. To simultaneously exploit spatial and temporal information, Xiong et al.[21] proposed the Bidirectional Convolutional LSTM (Bidirectional ConvLSTM) on video sequences to access long short-range temporal information in two opposite directions. Li et al.[12] utilized a pretrained partial VGG-16 network[22] and a cascaded atrous convolution regressor to aggregate multiscale features.

Liu et al.[23] presented a multitask framework in which unlabeled crowd imagery is leveraged to address the lack of large training datasets, simultaneously ranking images and estimating crowd density maps. To reach scale diversity and high-quality density maps, Cao et al.[18] proposed a scale aggregation encoder-decoder network trained by adding local pattern consistency loss. Ranjan et al.[24] presented the iterative counting CNN (ic-CNN) of two branches, which shared features and generated low- and high-resolution density maps separately. To further improve the prediction quality, a multi-stage ic-CNN combining all the density predictions of multiple ic-CNNs was employed.

## 3 Refinement Network for Crowd Counting

This section presents our proposed RefNet for crowd counting, which consists of three modules, namely, densely connected bottom-up module, top-down refinement module, and multiresolution density regression loss, as shown in Fig. 2. $L(Y^j, Y^{GT})$ is the supplementary loss function for pyramidal density map regression, where $j \in \{1/8, 1/4, 1/2\}$. $L_{out}(Y, Y^{GT})$ means the loss function operated on the final output density map. The definition of these functions will be introduced in Section 3.3.

Features before the refinement module are extracted by a bottom-up module containing four dense encoding blocks, in which convolutional layers are densely connected, same as the connection mode in DenseNet[25]. These features are then gradually refined by a top-down scheme containing three top-down blocks, each of which comprises a set of upsampling and convolutional layers. Multiscale semantic information is sequentially extracted during this process. Guided by the multiresolution density regression loss, the output density map is refined, sequentially aggregating fine granular information during coarse-to-fine intermediate density map generation process in the refinement module.

### 3.1 Densely connected module

Features among different scales appear to be similar without a clear difference. As DenseNet[25] has a competitive performance on feature extraction and generalization and shares extracted information among convolutional layers, it can be used to represent continuous receptive fields without many global skip connections. It is also claimed to help train deep networks and reduce overfitting with small training datasets.

The upper left part of Fig. 2 shows the constructs of a bottom-up pathway. Considering the advantages discussed above, motivated by Ref. [26], a modified DenseNet is deployed as a backbone. We retain the main structure with the first 3×3 convolutionsl layer and the following four dense blocks, removing all the fully connected layers after the fourth dense block. The growth rate in each dense block is set to 16. The number of output channels for each 1×1 bottleneck convolutional layer is 4 times that of the growth rate. Layers between two adjacent dense blocks are referred to as a transition layer and a max-pooling layer, which aims to maintain the translation invariance to some extent. A transition layer is added after the third and the fourth dense blocks. The compression rates of all transition layers are set to 0.5. Each dense block has (4,12,8,8) dense layers.

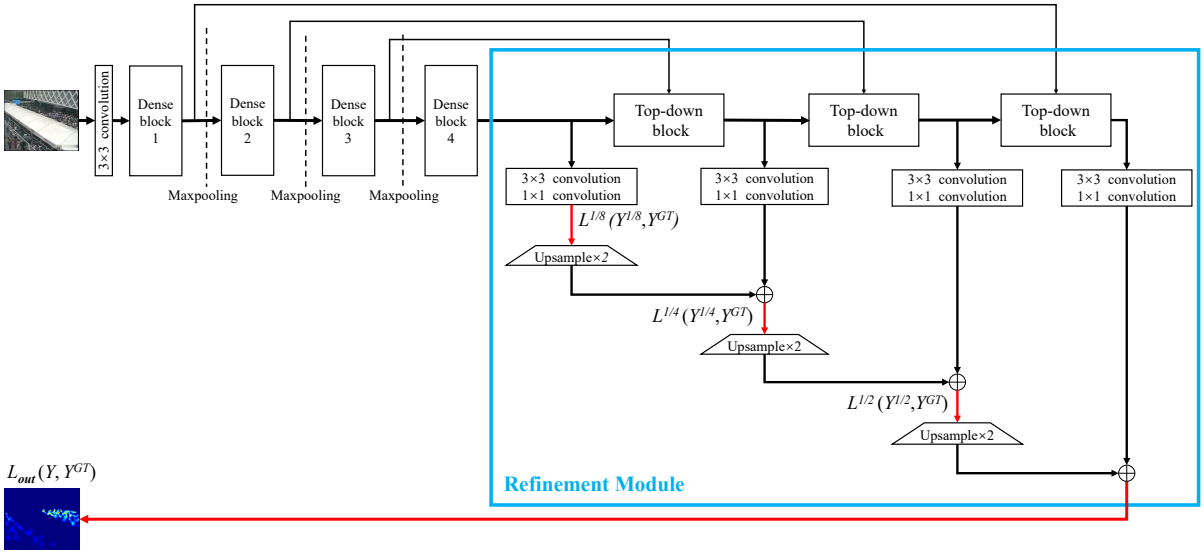



**Fig. 2 Architecture of our proposed method.**

### 3.2 Refinement module

In a bottom-up framework, high-resolution feature maps contain low-level information that may harm their representational capacity for object recognition. Low-resolution features have high-level semantic granularity but lack spatial information. Moreover, because of dense connections in the densely connected module, there is no need to replicate the features from a certain layer to the following layers through global skip connections for multiscale feature concatenation. However, the pooling layers may hamper the feature propagation by the downsampling layer between the two nearest dense blocks. To further improve model compactness, we compensate the fine but low-level information blocked by the pooling layer between each dense block by laterally connecting the top-down stage with the same resolution. To fuse multiscale information in the prediction stage, considering the uneven distribution of the object in a certain scenario, we further employ a multiresolution refinement module to sequentially aggregate spatial granular information, reaching a higher-quality density map and more precise counting value.

The right part of Fig. 2 in the blue line shows the proposed refinement structure of a top-down module. For coarse-resolution feature maps, we first upsample the spatial resolution by a factor of 2 (using bilinear upsampling). The upsampled feature maps are then put into a 3×3 convolution to refine themselves. Here we concatenate the refined upsampled feature maps with the corresponding bottom-up feature maps proposed by the transition layer after a dense block to retain the information flow instead of the element-wise summation proposed in Ref. [19]. Then a set of convolutions containing 1×1 and two 3×3 layers are used for feature fusion and channel compression by half. The extracted features are utilized as the input of the next top-down block and the information for the intermediate-density prediction. The predicted density map on a certain resolution in the feature pyramid is estimated by 3×3 and a 1×1 convolution layers. This process is iterated in the top-down pathway until the finest-resolution density map is generated.

Although this feature pyramid network concatenates multiscale information, each intermediate-density map is regressed separately, and the density map with fine spatial information does not fully utilize the coarse granular information estimated in previous top-down blocks. Hence, alleviating the uneven distribution issue is essential. Therefore, different from the feature pyramid network, the refinement scheme is proposed to sequentially aggregate in a coarse-to-fine granular process. The previously estimated intermediate-density map with coarse granular information is first upsampled by a factor of 2 and then is added to generate a density map containing fine granular information. In addition, this refined intermediate-density map is trained by the corresponding regression loss and is observed for the information that is used for the density map estimation with a high resolution. The finest density map is finally regressed by iteratively operating this process.

### 3.3 Multiresolution density regression loss

Most state-of-the-art crowd-counting methods are trained using the Euclidean loss between the ground-truth and estimated density maps. In the present study, we follow the work in Ref. [16], which focuses on not only the counting accuracy but also the quality of density maps. In addition, because of the multiresolution expectation and sequential fusion for the density map generation in the refinement module, the intermediate-density map prediction for every spatial level contributes to the final output and should also be individually considered to further enhance the counting performance. Hence, each of the spatial levels in the proposed refinement module should be added and guided by their own regression loss.

Therefore, different from ACSPNet[16], the multiresolution density regression loss function $L(Y_i, Y_i^{GT})$ in this work is introduced as follows:

$$L(Y_i, Y_i^{GT}) = L_{out}(Y_i, Y_i^{GT}) + \gamma \sum_{j \in J} L^j(Y_i^j, Y_i^{GT}) \tag{1}$$

where $\gamma$ is a constant weighting factor with a value 0.01. $J = \{1/8, 1/4, 1/2\}$ corresponds to the label of various resolution outputs generated by the refinement module. $Y_i$ means the finest density map estimated by the proposed RefNet for an input image $X_i$, of which the corresponding ground-truth density map is $Y_i^{GT}$. $L_{out}(\cdot,\cdot)$ means the loss function proposed in Ref. [16] that operates on the final output density map,

$$L_{out}(Y_i, Y_i^{GT}) = L_E(Y_i, Y_i^{GT}) + \alpha_1 L_P(Y_i, Y_i^{GT}) + \alpha_2 L_A(Y_i, Y_i^{GT}) \tag{2}$$

$$L_E(Y_i, Y_i^{GT}) = \frac{1}{2M} \sum_{i=1}^{M} \left\| Y_i - Y_i^{GT} \right\|_2^2 \tag{3}$$

$$L_P(Y_i, Y_i^{GT}) = \frac{1}{M}\sum_{i=1}^{M}\sum_{w=1,h=1}^{W,H}\left|Y_i^{w,h} - (Y_i^{GT})^{w,h}\right| \tag{4}$$

$$L_A(Y_i, Y_i^{GT}) = \frac{1}{M}\sum_{i=1}^{M}\left|C(Y_i) - C(Y_i^{GT})\right| \tag{5}$$

where $M$ is the size of the training set. $\alpha_1$ and $\alpha_2$ are weighting factors of pixel-wise $l_1$ loss and mean absolute error loss (counting loss) between the final output density map and the ground-truth, respectively, which are both set as 0.1. $H$ and $W$ represent the height and width of the density map $Y_i^{GT}$ or $Y_i$, respectively. $C(Y_i)$ is the estimated number of pedestrians in the imput image $X_i$, which is defined as

$$C(Y_i) = \sum_{w=1}^{W}\sum_{h=1}^{H} Y_i^{h,w} \tag{6}$$

where $Y_i^{h,w}$ equals to the value of pixel at $(h, w)$ in $Y_i$. $C(Y_i^{GT})$ is the real quantity value of the objects in the imput image $X_i$. For the predictions on level $j \in J$ of pyramidal features supervised by $L^j(Y_i^j, Y_i^{GT})$, the supplementary loss function is as follows:

$$L^j(Y_i^j, Y_i^{GT}) = \frac{1}{2M}\sum_{i=1}^{M}\left\|Y_i^j - Y_i^{GT}\right\|_2^2 \tag{7}$$

## 4 Experiments on Benchmark Datasets

We first introduce the training configurations that contain ground-truth generation, evaluation metrics, and details of the training procedure. Then, the proposed RefNet is evaluated and compared to the state-of-the-art methods on three standard benchmark datasets[7, 27, 28]. We further extend our RefNet to cell-counting tasks. The ablation study of the feature pyramid network, refinement module, and multiresolution density regression loss on RefNet is finally introduced to show their effectiveness on object counting.

### 4.1 Settings

#### 4.1.1 Ground-truth generation

To generate ground-truth density maps for training our proposed RefNet, we follow the same configurations as those presented in Ref. [29]. However, several datasets, such as ShanghaiTech[7] and UCF_CC_50[28], are congested datasets with head-size variation problems, and do not provide perspective information. Following the method proposed by Zhang et al.[7], geometry-adaptive Gaussian kernels are utilized to generate density maps. For the input image $X_i$, an annotation $s_k$ ($K$ in total) in annotation set $S$ is denoted. Then, $Y_i^{GT}$ is calculated by convolving a Dirac function $\delta(k)$ with a normalized Gaussian kernel $G_{\sigma_k}$,

$$Y_i^{GT} = \sum_{s_k \in S}\delta(X_i - s_k) \times \beta G_{\sigma_k} \tag{8}$$

where $\beta$ is the constant parameter. The integral of $Y_i^{GT}$ is equal to the number of people $K$ in the input image $X_i$.

#### 4.1.2 Evaluation matrix

In this study, the Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) are adopted to evaluate diverse counting methods, which are defined as

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}\left|C(Y_i) - C(Y_i^{GT})\right| \tag{9}$$

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left[C(Y_i) - C(Y_i^{GT})\right]^2} \tag{10}$$

where $N$ is the number of test images.

#### 4.1.3 Training setting

The implementation of our model is under the Pytorch framework (Version 1.2). The version of CUDA is 10.0. For all the convolution layers, the initial values derive from a Gaussian initialization with 0.01 standard deviation. Adam optimization[30] is applied with a fixed learning rate $10^{-5}$ for a total 300 epochs during training. Because of variant resolution inputs, the batch size is set to 1. The nonlinear activation function Rectified Linear Unit (ReLU) is utilized before each convolution. Because the values of pixels belonging to the generated density maps are also non-negative, we also add a ReLU function after the final fusion layer. No batch normalizations are utilized in the network.

### 4.2 ShanghaiTech dataset

The ShanghaiTech dataset[7] is a large-scale crowd-counting dataset. It contains two parts: Part A, with 482 images, is randomly collected from the Internet, and Part B, including 716 images, is taken from urban areas in Shanghai. For Part A, $G_{\sigma_k}$ for each annotation $k$ is calculated by the average distance among its three nearest heads using the k-nearest neighbor method, and $\beta$ is set to 0.3. For Part B, $\beta G_{\sigma_k}$ is set as a fixed value 4.

We compare our RefNet to state-of-the-art methods recently published on this dataset. All the detailed results for each method are shown in Table 1, which indicates that our RefNet can reach a competitive counting performance compared to the state-of-the-art methods. In particular, our proposed RefNet outperforms ACSPNet[16] with 12.6/14.5 points

**Table 1 Estimation errors on the ShanghaiTech dataset[7].**

| Method | Part A | | Part B | |
|---|---|---|---|---|
| | MAE | RMSE | MAE | RMSE |
| Zhang et al.[29] | 181.8 | 277.7 | 32.0 | 49.8 |
| MCNN[7] | 110.2 | 173.2 | 26.4 | 41.3 |
| Switching-CNN[8] | 90.4 | 135.0 | 21.6 | 33.4 |
| Huang et al.[31] | – | – | 20.2 | 35.6 |
| CP-CNN[9] | 73.6 | 106.4 | 20.1 | 30.1 |
| ACSCP[17] | 75.7 | **102.7** | 17.2 | 27.4 |
| ACSPNet[16] | 85.2 | 137.1 | 15.4 | 23.1 |
| CSRNet[12] | **68.2** | 115.0 | 10.6 | 16.0 |
| RefNet (ours) | 72.6 | 122.6 | **10.1** | **15.5** |

Note: The bold and underline formats represent the best and second-best counting results, respectively. Tables 2 and 3 are the same.

(approximately 14.8%/10.6%) decrease on MAE/RMSE on ShanghaiTech-A dataset, with 5.3/7.6 points (approximately 34.4%/32.9%) decrease on MAE/RMSE on ShanghaiTech-B dataset. Though CSRNet[12] reached better results than RefNet (with 4.4/7.6 points increase on MAE/RMSE on ShanghaiTech-A dataset), the resolution of the density map in CSRNet is one-eighth of the original input, missing considerable spatial information. In addition, CSRNet uses a partial pretrained model on VGG-Net[22]. Moreover, our proposed RefNet outperforms CSRNet with 0.5/0.5 points (approximately 4.7%/3.1%) decrease on MAE/RMSE on ShanghaiTech-B dataset.

**Table 2 Estimation errors on the UCF_CC_50 dataset[28].**

| Method | MAE | RMSE |
|---|---|---|
| Idrees et al.[28] | 419.5 | 541.6 |
| Zhang et al.[29] | 467.0 | 498.5 |
| MCNN[7] | 377.6 | 509.1 |
| Hydra-CNN[32] | 333.7 | 425.2 |
| Cascaded-MTL[20] | 322.8 | 397.9 |
| Switching-CNN[8] | 318.1 | 439.2 |
| CP-CNN[9] | 295.8 | **320.9** |
| IG-CNN[33] | 291.4 | 349.4 |
| ACSCP[17] | 291.0 | 404.6 |
| ACSPNet[16] | 275.2 | 383.7 |
| RefNet (ours) | **258.6** | 373.2 |

**Table 3 Estimation errors on the UCSD dataset[27].**

| Method | MAE | RMSE |
|---|---|---|
| Zhang et al.[29] | 1.60 | 3.31 |
| MCNN[7] | 1.07 | 1.35 |
| FCN-rLSTM[34] | 1.54 | 3.02 |
| Bidirectional ConvLSTM[21] | 1.13 | 1.43 |
| CSRNet[12] | 1.16 | 1.47 |
| ACSCP[17] | 1.04 | 1.35 |
| SANet[18] | 1.02 | 1.29 |
| ACSPNet[16] | 1.02 | 1.28 |
| RefNet (ours) | **0.94** | **1.24** |

### 4.3 UCF_CC_50 dataset

The UCF_CC_50 dataset collected by Idrees et al.[28] contains 50 annotated images of different resolutions and aspect ratios in a variety of event scenarios. It is an extremely challenging dataset with a large counting variation. Following the standard protocol introduced in Ref. [28], fivefold cross-validation is performed for evaluating the proposed method. To generate the ground-truth density maps, we follow the settings in ShanghaiTech-A and the same configurations as in Ref. [7]. Table 2 shows that our network surpasses all of the state-of-the-art methods in terms of the MAE and reaches a competitive RMSE result. In particular, our proposed RefNet outperforms ACSPNet[16] with 16.6/10.5 points (approximately 6.0%/2.7%) decrease on MAE/RMSE. Hence, our method can obtain very competitive counting results with a small number of training samples compared to the state-of-the-art approaches.

### 4.4 UCSD dataset

The UCSD dataset[27] contains 2000 annotated frames of pedestrians on a walkway. This dataset also provides the region of interest. Among these frames, we use frames 601 to 1400 for training and the rest for testing in accordance with Ref. [27] and fix $\beta G_\sigma = 3$ of the ground-truth density maps. Table 3 shows that our proposed method reaches the best MAE and RMSE performances among all the state-of-the-art methods. In particular, our proposed RefNet outperforms ACSPNet[16] with 0.08/0.04 points (approximately 7.8%/3.1%) decrease on MAE / RMSE.

### 4.5 Epithelial cell dataset

To demonstrate the counting accuracy of the proposed RefNet on relative counting tasks, we further extend our approach to cell-counting tasks. The epithelial cell dataset contains 138 annotated frames, which have a resolution of 3648× 2736 or 2736× 3648. All the images in the epithelial cell dataset are captured by an endoscopic optical microscope (Supereyes B013). The annotated information represents the center point of each cell. Following the same ground-truth generation method as crowd counting, the annotated points are calculated using geometry-adaptive Gaussian kernels to generate density maps with parameter $\sigma_k$ valued 25. A total number of ten cell images are selected for testing, which are all hard cases for cell counting, such as

complex background, out of focus, overexposed, and cell stacking. The remaining parts are utilized for training. To evaluate the counting accuracy of the algorithm on a single image, the Absolute Error (AE) and Relative Error (RE) between the true value and estimated value of the number of cells in each image are analyzed, which are defined as

$$\mathrm{AE} = \left| C(Y_i) - C(Y_i^{GT}) \right| \tag{11}$$

$$\mathrm{RE} = \frac{\left| C(Y_i) - C(Y_i^{GT}) \right|}{C(Y_i^{GT})} \times 100\% \tag{12}$$

Table 4 illustrates the counting performances. In addition, to intuitively show the relationship between the true values and estimated values of the number of cells, data in Table 4 are visualized in Fig. 3, where the horizontal axis represents the true value of the number of cells and the vertical axis represents the number of cells estimated by the proposed algorithm. The red straight line represents the identity mapping relationship between the true values and estimated values. Figure 4

**Table 4 Estimation errors on the epithelial cell dataset.**

| Image sequence | Ground-truth | Estimated value | AE | RE (%) |
|---|---|---|---|---|
| 1 | 17 | 38 | 21 | 123 |
| 2 | 81 | 95 | 14 | 17 |
| 3 | 93 | 94 | 1 | 1 |
| 4 | 24 | 32 | 8 | 33 |
| 5 | 23 | 29 | 6 | 26 |
| 6 | 72 | 54 | 18 | 25 |
| 7 | 49 | 49 | 0 | 0 |
| 8 | 175 | 171 | 4 | 2 |
| 9 | 168 | 169 | 1 | 0.6 |
| 10 | 69 | 64 | 5 | 7 |
| MAE | | 7.8 | – | – |
| RMSE | | 10.5 | – | – |

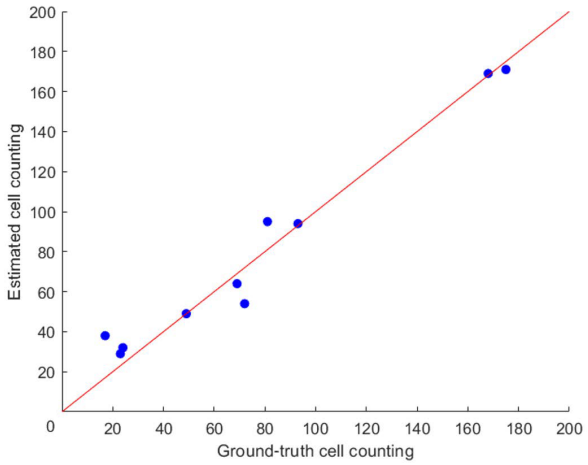



**Fig. 3 Comparison between ground-truth values and the estimated counting values.**

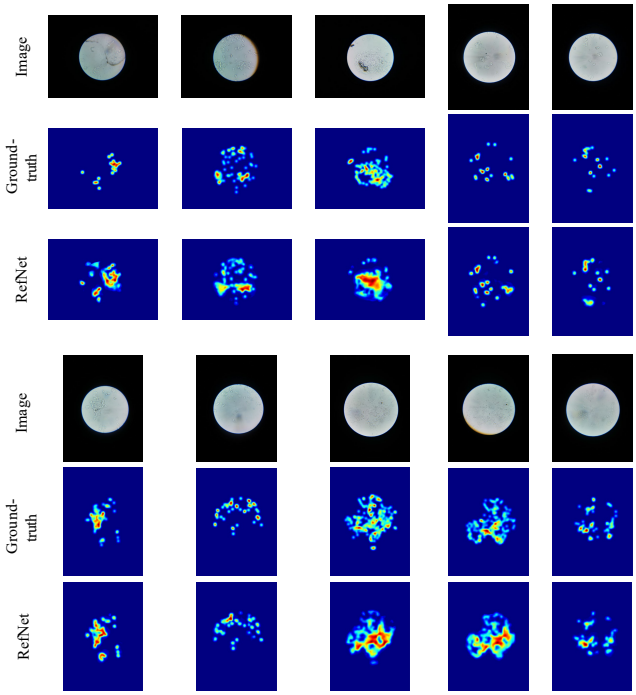



**Fig. 4 Results of the proposed network on the epithelial cell dataset.**

demonstrates the comparative results between the ground-truth density maps and output density maps.

Fivefold cross-validation[35, 36] is utilized for the RefNet evaluation on our proposed epithelial cell dataset to illustrate the counting performance on the whole dataset. The dataset is divided into five groups according to the serial number of images. The number of images contained in each group is 28, 28, 28, 28, and 26. The MAE and RSME counting results are concluded in Table 5.

We also analyze the RE between the true value and the estimated value of the number of cells in the images of each group. The RE results, as well as the average number of cells in each cross-validation group, are presented in Table 6. The proposed RefNet infers an accurate number of cells on a high-density level of images but with worse counting inference on sparse cell situations. The counting accuracy may be increased if

**Table 5 Fivefold cross-validation performance for the RefNet evaluation on the epithelial cell dataset.**

| Fivefold cross-validation | MAE | RSME |
|---|---|---|
| Group 1 | 12.0 | 18.2 |
| Group 2 | 15.6 | 21.7 |
| Group 3 | 12.2 | 16.8 |
| Group 4 | 18.9 | 29.8 |
| Group 5 | 11.5 | 15.1 |
| Average | 14.1 | 21.1 |

**Table 6 RE performance for fivefold cross-validation RefNet evaluation on the epithelial cell dataset.**

| Fivefold cross-validation | Average value | MAE | RE (%) |
|---|---|---|---|
| Group 1 | 39.5 | 12.0 | 30.4 |
| Group 2 | 85.4 | 15.6 | 18.3 |
| Group 3 | 73.8 | 12.2 | 16.5 |
| Group 4 | 75.2 | 18.9 | 25.1 |
| Group 5 | 77.0 | 11.5 | 14.9 |
| Average | 70.1 | 14.1 | 21.2 |

**Table 7 Comparison of the estimation errors with different module configurations on ShanghaiTech-A dataset[7].**

| Method | MAE | RMSE |
|---|---|---|
| Baseline | 89.8 | 145.1 |
| MFPNet without multiresolution density regression loss | 79.7 | 129.5 |
| MFPNet with multiresolution density regression loss | 75.2 | 124.1 |
| RefNet | **72.6** | **122.6** |

the detection-based counting method is utilized.

### 4.6 Ablation study

We compare our RefNet with the baseline model and Multiresolution Feature Pyramid Network (MFPNet)[26] without and with the multiresolution density regression loss function on the effectiveness of the refinement module and the modified benchmark ShanghaiTech-A dataset[7] to clarify the multiresolution density regression loss function. The results of the comparative experiments are illustrated in Table 7. By utilizing the feature pyramid module, multiscale features are captured and gathered. The performance becomes better, with MAE/RMSE evaluations 10.1/15.6 points lower than that of the baseline on the ShanghaiTech-A dataset[7].

By adding the multiresolution density regression loss function in the MFPNet, multiple resolution density maps are produced, which are supervised by each level of the feature pyramid. The performance is improved, with the MAE/RMSE evaluations 14.6/21.0 points lower than that of the baseline and 4.5/5.4 points lower than that of the MFPNet without the multiresolution density regression loss function on the ShanghaiTech-A dataset[7]. Instead of the MFPNet, by utilizing the refinement module in a top-down scheme, not only the scale variation issue is concerned but also the uneven distribution of the crowd is taken into consideration. The performance reaches the best, with the MAE/RMSE evaluations 17.2/22.5 points lower than that of the baseline, 7.1/6.9 points lower than that of the MFPNet without the multiresolution density regression loss function, and 2.6/1.5 points lower than that of the MFPNet with the multiresolution density regression loss function on the ShanghaiTech-A dataset[7].

To illustrate the function of the modules in RefNet, some estimated density maps from the above various configurations on the input images from ShanghaiTech-A dataset[7] are shown in Fig. 5. The top row demonstrates the sample input test images from the ShanghaiTech-A[7]. The ground-truth density maps of the first-row images are shown in the second row. Rows 3 to 6 represent the corresponding density maps generated by a baseline, the MFPNet without the multiresolution density regression loss function, the MFPNet with the multiresolution density regression loss function, and RefNet. Hence, the use of a refinement module and multiresolution density regression loss function contribute to improving the quality of generated density maps and reducing the count error.

## 5 Conclusion

In this paper, an algorithm named RefNet is introduced for object counting. To sequentially concatenate multiscale semantic features in a top-down decoder procedure, a refinement module is proposed, guided by a modified multiresolution density regression loss function, for high-quality density map generation and counting estimation to alleviate the scale variation and uneven distribution issues. Extensive experiments show that our approach achieves competitive performance compared to recent state-of-the-art methods. In future works, as background noise in crowd scenarios may influence counting accuracy, methods to alleviate noisy backgrounds should be considered. Deep metric learning[37, 38] approaches may be concerned with effectively distinguishing between heads and background information.

## References

[1] T. Li, H. Chang, M. Wang, B. B. Ni, R. C. Hong, and S. C. Yan, Crowded scene analysis: A survey, *IEEE Trans. Circuits Syst. Video Technol.*, vol. 25, no. 3, pp. 367–386, 2015.

[2] Y. M. Rao, J. W. Lu, J. Lin, and J. Zhou, Runtime network routing for efficient image classification, *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 10, pp. 2291–2304, 2019.

[3] H. Liu, J. W. Lu, J. J. Feng, and J. Zhou, Two-stream transformer networks for video-based face alignment, *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 11, pp. 2546–2554, 2018.

[4] L. C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully

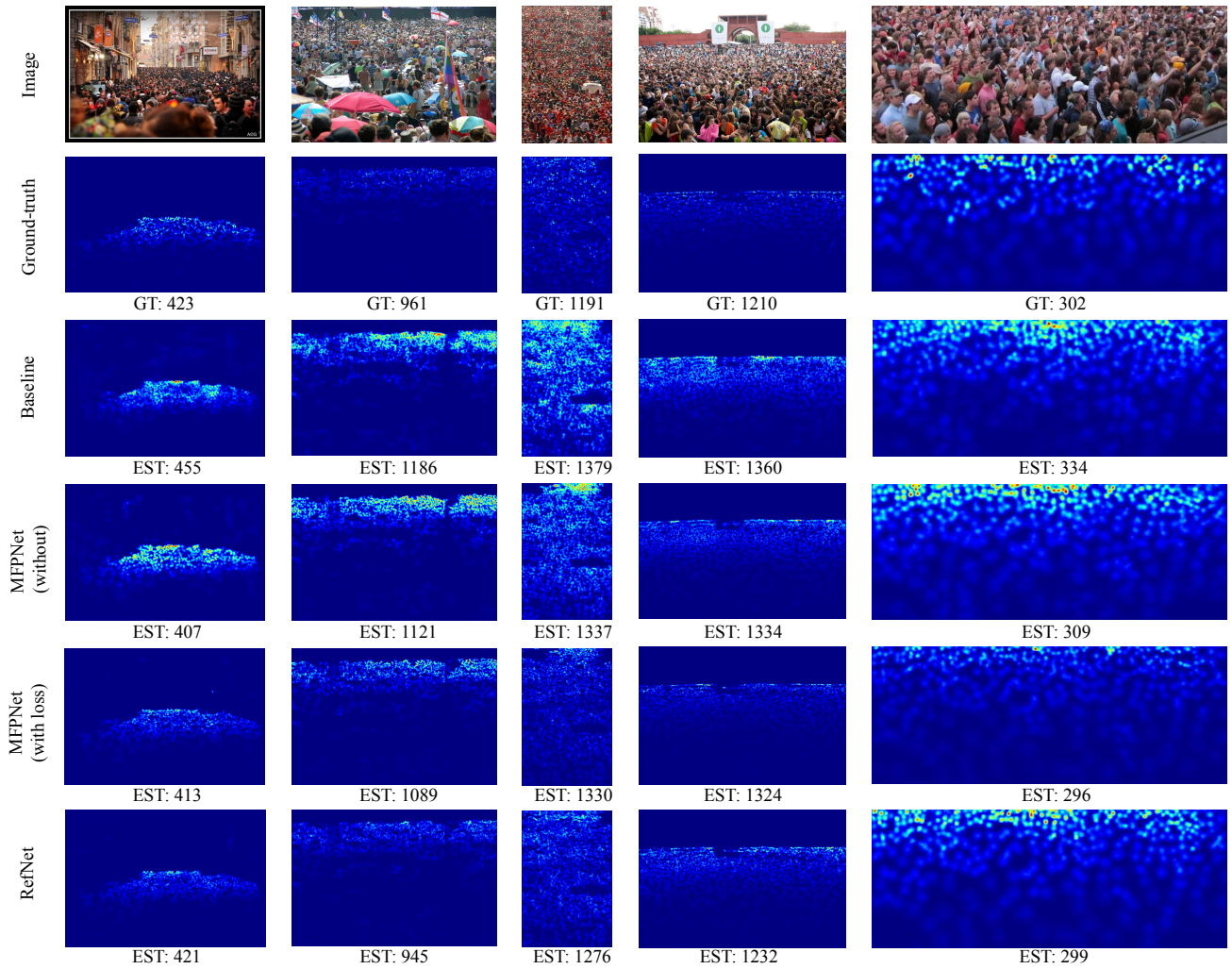



**Fig. 5 Comparison of results from different configurations of the proposed network on ShanghaiTech-A dataset[7].**

connected CRFs, *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, 2018.

[5] V. A. Sindagi and V. M. Patel, A survey of recent advances in CNN-based single image crowd counting and density estimation, *Pattern Recognit. Lett.*, vol. 107, pp. 3–16, 2018.

[6] J. Long, E. Shelhamer, and T. Darrell, Fully convolutional networks for semantic segmentation, in *Proc. 2015 IEEE Conf. Computer Vision and Pattern Recognition*, Boston, MA, USA, 2015, pp. 3431–3440.

[7] Y. Y. Zhang, D. S. Zhou, S. Q. Chen, S. H. Gao, and Y. Ma, Single-image crowd counting via multi-column convolutional neural network, in *Proc. 2016 IEEE Conf. Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, 2016, pp. 589–597.

[8] D. B. Sam, S. Surya, and R. V. Babu, Switching convolutional neural network for crowd counting, in *Proc. 2017 IEEE Conf. Computer Vision and Pattern Recognition*, Honolulu, HI, USA, 2017, pp. 4031–4039.

[9] V. A. Sindagi and V. M. Patel, Generating high-quality crowd density maps using contextual pyramid CNNs, in *Proc. 2017 IEEE Int. Conf. Computer Vision*, Venice, Italy, 2017, pp. 1879–1888.

[10] D. Deb and J. Ventura, An aggregated multicolumn dilated convolution network for perspective-free counting, in *Proc. 2018 IEEE/CVF Conf. Computer Vision and Pattern Recognition Workshops*, Salt Lake City, UT, USA, 2018, pp. 195–204.

[11] M. Hossain, M. Hosseinzadeh, O. Chanda, and Y. Wang, Crowd counting using scale-aware attention networks, in *Proc. 2019 IEEE Winter Conf. Applications of Computer Vision*, Waikoloa, HI, USA, 2019, pp. 1280–1288.

[12] Y. H. Li, X. F. Zhang, and D. M. Chen, CSRNet: Dilated convolutional neural networks for understanding the highly congested scenes, in *Proc. 2018 IEEE/CVF Conf. Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, 2018, pp. 1091–1100.

[13] F. Yu and V. Koltun, Multi-scale context aggregation by dilated convolutions, in *Proc. of the* $4^{th}$ *Int. Conf. Learning Representations*, arXiv preprint arXiv: 1511.07122, 2015.

[14] K. M. He, X. Y. Zhang, S. Q. Ren, and J. Sun, Spatial pyramid pooling in deep convolutional networks for visual recognition, *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1904–1916, 2015.

[15] L. C. Chen, G. Papandreou, F. Schroff, and H. Adam, Rethinking atrous convolution for semantic image segmentation, arXiv preprint arXiv: 1706.05587, 2017.

[16] J. J. Ma, Y. P. Dai, and Y. P. Tan, Atrous convolutions spatial pyramid network for crowd counting and density estimation, *Neurocomputing*, vol. 350, pp. 91–101, 2019.

[17] Z. Shen, Y. Xu, B. B. Ni, M. S. Wang, J. G. Hu, and X. K.

Yang, Crowd counting via adversarial cross-scale consistency pursuit, in *Proc. 2018 IEEE/CVF Conf. Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, 2018, pp. 5245–5254.
[18] X. K. Cao, Z. P. Wang, Y. Y. Zhao, and F. Su, Scale aggregation network for accurate and efficient crowd counting, in *Proc. 15th European Conf. Computer Vision*, Munich, Germany, 2018, pp. 757–773.
[19] T. Y. Lin, P. Dollár, R. Girshick, K. M. He, B. Hariharan, and S. Belongie, Feature pyramid networks for object detection, in *Proc. 2017 IEEE Conf. Computer Vision and Pattern Recognition*, Honolulu, HI, USA, 2017, pp. 936–944.
[20] V. A. Sindagi and V. M. Patel, CNN-based cascaded multi-task learning of high-level prior and density estimation for crowd counting, in *Proc. 14th IEEE Int. Conf. Advanced Video and Signal Based Surveillance*, Lecce, Italy, 2017, pp. 1–6.
[21] F. Xiong, X. J. Shi, and D. Y. Yeung, Spatiotemporal modeling for crowd counting in videos, in *Proc. 2017 IEEE Int. Conf. Computer Vision*, Venice, Italy, 2017, pp. 5161–5169.
[22] K. Simonyan and A. Zisserman, Very deep convolutional networks for large-scale image recognition, in *Proc. of the 3$^{rd}$ Int. Conf. Learning Representations*, arXiv preprint arXiv: 1409.1556, 2014.
[23] X. L. Liu, J. van de Weijer, and A. D. Bagdanov, Leveraging unlabeled data for crowd counting by learning to rank, in *Proc. 2018 IEEE/CVF Conf. Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, 2018, pp. 7661–7669.
[24] V. Ranjan, H. Le, and M. Hoai, Iterative crowd counting, in *Proc. 15th European Conf. Computer Vision*, Munich, Germany, 2018, pp. 278–293.
[25] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, Densely connected convolutional networks, in *Proc. 2017 IEEE Conf. Computer Vision and Pattern Recognition*, Honolulu, HI, USA, 2017, pp. 2261–2269.
[26] J. Ma, Y. Dai, Z. Jia, and K. Hirota, Multi-resolution feature pyramid network for crowd counting, in *Proc. Int. WorkshopAdv. Comput. Intell. Intell. Inf.*, Chengdu, China, 2019, pp. 6–11.
[27] A. B. Chan, Z. S. J. Liang, and N. Vasconcelos, Privacy preserving crowd monitoring: Counting people without people models or tracking, in *Proc. 2008 IEEE Conf. Computer Vision and Pattern Recognition*, Anchorage, AK, USA, 2008, pp. 1–7.
[28] H. Idrees, I. Saleemi, C. Seibert, and M. Shah, Multi-source multi-scale counting in extremely dense crowd images, in *Proc. 2013 IEEE Conf. Computer Vision and Pattern Recognition*, Portland, OR, USA, 2013, pp. 2547–2554.
[29] C. Zhang, H. S. Li, X. G. Wang, and X. K. Yang, Cross-scene crowd counting via deep convolutional neural networks, in *Proc. 2015 IEEE Conf. Computer Vision and Pattern Recognition*, Boston, MA, USA, 2015, pp. 833–841.
[30] D. P. Kingma and J. Ba, Adam: A method for stochastic optimization, in *Proc. of the 3$^{rd}$ Int. Conf. Learning Representations*, arXiv preprint arXiv: 1412.6980, 2014.
[31] S. Y. Huang, X. Li, Z. F. Zhang, F. Wu, S. H. Gao, R. R. Ji, and J. W. Han, Body structure aware deep crowd counting, *IEEE Trans. Image Process.*, vol. 27, no. 3, pp. 1049–1059, 2018.
[32] D. Oñoro-Rubio and R. J. López-Sastre, Towards perspective-free object counting with deep learning, in *Proc. 14th European Conf. Computer Vision*, Amsterdam, The Netherlands, 2016, pp. 615–629.
[33] D. Babu Sam, N. N. Sajjan, R. Venkatesh Babu, and M. Srinivasan, Divide and grow: Capturing huge diversity in crowd images with incrementally growing CNN, in *Proc. 2018 IEEE/CVF Conf. Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, 2018, pp. 3618–3626.
[34] S. H. Zhang, G. H. Wu, J. P. Costeira, and J. M. F. Moura, FCN-rLSTM: Deep spatio-temporal neural networks for vehicle counting in city cameras, in *Proc. 2017 IEEE Int. Conf. Computer Vision*, Venice, Italy, 2017, pp. 3687–3696.
[35] J. J. Tie, X. J. Lei, and Y. Pan, Metabolite-disease association prediction algorithm combining DeepWalk and random forest, *Tsinghua Science and Technology*, vol. 27, no. 1, pp. 58–67, 2022.
[36] Y. H. Yu and J. Li, Residuals-based deep least square support vector machine with redundancy test based model selection to predict time series, *Tsinghua Science and Technology*, vol. 24, no. 6, pp. 706–715, 2019.
[37] J. L. Hu, J. W. Lu, Y. P. Tan, and J. Zhou, Deep transfer metric learning, *IEEE Trans. Image Process.*, vol. 25, no. 12, pp. 5576–5588, 2016.
[38] J. L. Hu, J. W. Lu, and Y. P. Tan, Sharable and individual multi-view metric learning, *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 9, pp. 2281–2288, 2018.

**Liping Jing** received the PhD degree from University of Hong Kong, China in 2007. She was a postdoctor researcher at Hong Kong Baptist University, University of Texas at Dallas, and University of California at Berkeley. She is currently a professor at Beijing Key Laboratory of Traffic Data Analysis and Mining, School of Computer and Information Technology, Beijing Jiaotong University, Beijing, China. She serves as a regular reviewer and program committee member for a number of international journals and conferences. Her research focuses on machine learning and pattern recognition.

**Junjie Ma** received the BEng degree from Taiyuan University of Technology, Taiyuan, China in 2012, and the PhD degree from Beijing Institute of Technology, Beijing, China in 2020. He was an internship student at Nanyang Technological University, Singapore from 2017 to 2019. He is currently an assistant researcher at the Department of Computer Science and Technology, Tsinghua University, Beijing, China. His research interests include computer vision and deep learning.

**Lehan Sun** is an undergraduate student at the School of Science, Beijing Jiaotong University, China. His research interests include data classification and particle counting in a complex environment.