JEI Journal of Electronic Imaging

RESEARCH PAPER

SDANet: scale-deformation awareness network for crowd counting

Jianyong Wang,[a] **Xiangyu Guo**,[a] **Qilei Li**,[b] **Ahmed M. Abdelmoniem**,[b] **and Mingliang Gao**[a,*]

[a]Shandong University of Technology, School of Electrical and Electronic Engineering, Zibo, China
[b]Queen Mary University of London, School of Electronic Engineering and Computer Science, London, United Kingdom

ABSTRACT. Crowd counting aims to derive information about crowd density by quantifying the number of individuals in an image or video. It offers crucial insights applicable to various domains, e.g., secure, efficient decision-making, and management. However, scale variation and irregular shapes of heads pose intricate challenges. To address these challenges, we propose a scale-deformation awareness network (SDANet). Specifically, a scale awareness module is introduced to address the scale variation. It can capture long-distance dependencies and preserve precise spatial information by readjusting weights in height and width directions. Concurrently, a deformation awareness module is introduced to solve the challenge of head deformation. It adjusts the sampling position of the convolution kernel through deformable convolution and learning offset. Experimental results on four crowd-counting datasets prove the superiority of SDANet in accuracy, efficiency, and robustness.

© 2024 SPIE and IS&T [DOI: 10.1117/1.JEI.33.4.043002]

Keywords: crowd counting; scale variation; deformable head; convolutional neural network

Paper 240308G received Mar. 27, 2024; revised May 23, 2024; accepted Jun. 10, 2024; published Jul. 2, 2024.

1 Introduction

Crowd counting constitutes a crucial and pragmatic pursuit to precisely ascertain the count of pedestrians within a static image or video sequence. Crowd counting plays crucial roles in various domains, such as urban security,[1] activity management,[2–4] and transportation planning.[5]

With the emergence of increasingly outstanding models, there is an increasing diversity of methods for crowd counting. In the early stages, the predominant crowd-counting approaches were based on detection[6,7] and regression methods.[8–11] These methods demonstrated satisfactory counting performance in scenes with sparse crowds. However, they may lose accuracy when confronted with challenges such as massive quantities and scale variations. With the advent of deep learning,[12,13] convolutional neural networks (CNNs) have become widely utilized in crowd counting. In densely populated and complex scenes, CNNs excel at capturing intricate spatial relationships within the crowd through feature learning in images. This serves to enhance the accuracy of counting.

Although the current deep learning-based methods have effectively improved counting accuracy, these methods still struggle to achieve high-precision counting when facing challenges, such as large-scale scenarios and deformations. Two primary challenges encountered in crowd counting are depicted in Fig. 1. Figure 1(a) illustrates a standard convolution operation

[*] Address all correspondence to Mingliang Gao, mlgao@sdut.edu.cn

1017-9909/2024/$28.00 © 2024 SPIE and IS&T

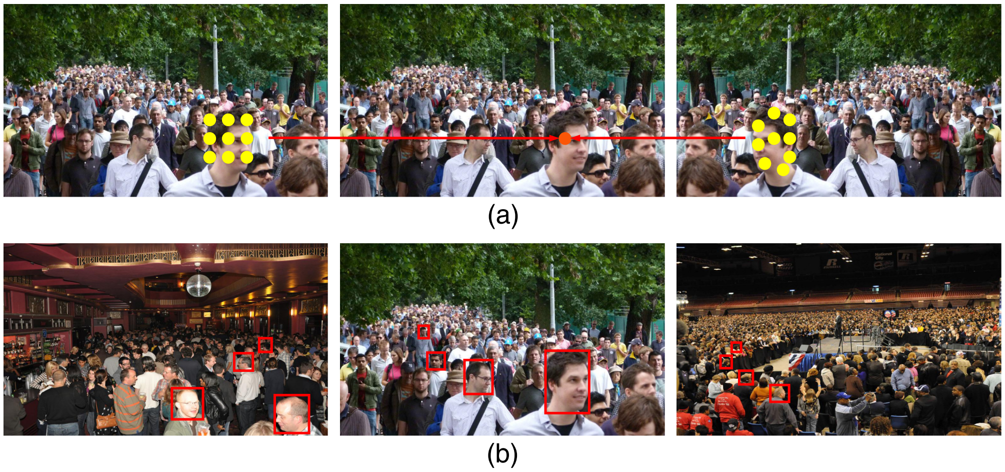

Fig. 1 Challenges of head deformation and scale variations in crowd counting. (a) Exemplars of standard convolution and deformable convolution in crowd counting. (b) Exemplars of scale variations in crowd counting.

performed by a fixed-shape (usually rectangular) filter on an input image. In the convolution process, the weight value of the filter does not change. However, for some scenes, such as images with a large variety of object shapes, traditional convolution may not be well adapted to the irregular shape of the object. Thus it can lead to inaccurate localization and counting. To overcome this limitation, we recommend using deformable convolution. As depicted in Fig. 1(a), deformable convolution autonomously adjusts the convolution sampling positions to accommodate the varying sizes of deformed heads. This convolutional approach substantially enhances counting efficiency.

Moreover, the consistent scale variation has presented a noteworthy challenge in crowd counting, especially in dense crowd scenes. Figure 1(b) illustrates the scale variations challenge in crowd counting. Multiscale feature aggregation stands out as an effective strategy to address this challenge. Guo et al.[14] introduced a multiscale aggregation module utilizing convolutions with different dilation rates to capture features across multiple scales. Cao et al.[15] proposed a scale-aggregation network, where an encoder employs a scale aggregation module to extract features across multiple scales. Furthermore, some works[16,17] introduced attention modules to deal with scale variations. Zhai et al.[18] proposed a crowd-split attention network to jointly capture spatial and channel dimension information, thereby preserving scale information effectively. Despite the efficacy of these approaches in addressing scale variations, the intensity of scale variation becomes more pronounced in dense crowd scenarios. Models need to capture and understand spatial information in images more effectively. This facilitates better adaptation to various scales of targets and scenes.

To address the problem of scale variation and irregular head shapes, we propose a scale-deformation awareness network (SDANet). The main architecture of the network has two modules. First, a scale awareness (SA) module bifurcates the channel attention into two ID feature coding processes. This approach aggregates horizontal and vertical features to capture distant dependencies and preserve precise location information. Additionally, we propose the DA module to address head deformation challenges in crowd scenes. We utilize deformable convolution to adjust the shape of the convolution kernel dynamically. This correction is applied specifically to the features in the deformation part. The contributions of this paper are summarized as follows.

(1) An SA module is built to address the challenge of scale variations by preserving accurate spatial information.
(2) A DA module is introduced to tackle the issue of head deformation in complex scenes.
(3) The precision and resilience of SDANet are demonstrated via various experimental results across various datasets. Meanwhile, the ablation experiments have further substantiated the effectiveness of the proposed method.

2 Related Work

2.1 Head Deformation

The challenge of handling deformed heads poses a formidable obstacle for conventional CNNs. Head deformations can hinder the accurate extraction of facial features in crowd images. This problem consequently impacts the counting precision of the network. To address this issue, Luo et al.[19] introduced the concept of an effectively receptive field. This concept enables the network to respond appropriately to distinct regions of varying sizes within the input image. Jia et al.[20] proposed a dynamic filter network, where filters adjust dynamically based on various inputs to accommodate local spatial transformations. Although the filters employed here demonstrate adaptability to spatial variations, their deformation capabilities are constrained. In response, Yu and Koltun[21] proposed an expansive convolution. This convolution preserves resolution while accommodating different-sized receptive fields based on the expansion factor.

With the advent of expansive convolutions, deformable convolutions have gradually found applications in object detection.[22,23] This architecture is built upon the notion of enhancing spatial sampling positions and learning spatial offsets at each location without prior knowledge.

The above methods provide diverse technologies for tackling head deformation issues. Nevertheless, these methods fall short of attaining outstanding deformation performance. When confronted with irregular head shapes, they often fail to achieve superior counting efficacy. To overcome this issue, we propose the DA module. This module integrates the deformable convolution method to dynamically adapt the convolution kernel shape, thereby enhancing its compatibility with deformable regions.

2.2 Scale Variation

Scale variation is a challenging factor in image processing tasks, such as image quality assessment, and image fusion. Inspired by the perception of the human visual system combined with multiscale features, Zhou and Chen[24] utilized pyramid feature learning to construct a DNN with layered multiscale features to predict distorted image quality. Sun et al.[25] proposed multiscale network (MCnet) to achieve high-quality image fusion. The MCnet could process the image feature fusion of different scales from coarse to fine and adaptively provide information on different scales and images.

Scale variation is also a challenging factor in crowd-counting tasks. To address this issue, recent crowd-counting models often adopted the multiscale pyramid structure. For example, Guo et al.[26] proposed the ghost attention pyramid network, which features a pyramid fusion module and establishes a four-branch architecture to obtain features across different scales. Zhao et al.[27] introduced a multiscale residual feature attention network. The key innovation of this network lies in optimizing losses for each scale to enable the network to adapt to scale variations. Additionally, attention mechanisms have been introduced to address large-scale variations in crowd counting. Guo et al.[28] proposed a multiscale perception attention fusion module. This module captures visual granularity information at different scales of the crowd region to generate high-quality density maps. To tackle the large-scale variations in complex crowds, Guo et al.[10] introduced a bottleneck spatial attention module and a multispectral channel attention module and combined the two attention modules. This approach effectively allocates spatial attention to different scales of crowd regions.

Although the above methods achieved excellent counting performance, they failed to reconcile global and local relationships while addressing spatial information at different scales. To address this challenge, we propose an SA module that can capture relationships between the global and local aspects while perceiving information about the image scale.

3 Methodology

3.1 Overview

The overall structure of the proposed SDANet is depicted in Fig. 2. We adopted focal inverse distance transform map (FIDTM)[29] as the baseline framework. The FIDTM adopts a local maximum detection strategy by adaptive threshold and max pooling to achieve more accurate counting. Overall, the SDANet comprises four components, namely the front-end network,

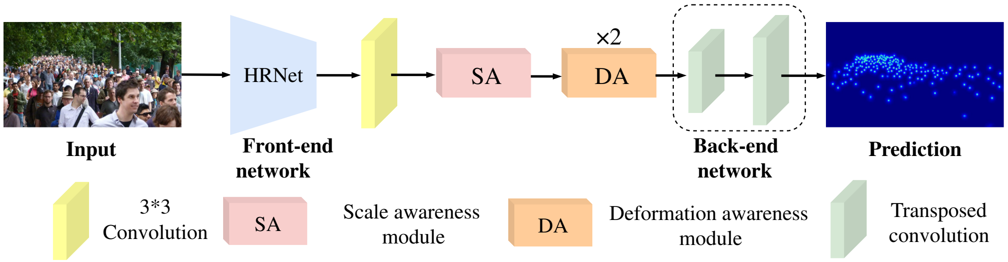

Fig. 2 Architecture of the SDANet for crowd counting.

SA module, DA module, and back-end network. Specifically, the high-resolution network (HRNet)[30] serves as the front-end network. An SA module is utilized to mitigate the influence of address scale variations. A DA module handles irregular head shapes during counting. Finally, a back-end network incorporating a series of transposed convolutions is equipped to output the estimated density map.

3.2 Scale Awareness Module

The schematic of the SA module is delineated in Fig. 3. The SA module is an attention module, which generates weights through dual dimension-pooling and convolution to extract multiscale features. This module accommodates input in the form of any intermediate feature tensor, denoted as $G = [g_1, g_2, \dots, g_C] \in \mathbb{R}^{C\times H\times W}$, and yields an enriched representation of equivalent dimensions, denoted as $R = [r_1, r_2, \dots, r_C]$. Existing global average pooling is based on channel dimension. Nonetheless, in crowd counting, the precise localization of human heads necessitates effectively preserving positional information inherent in feature images. Consequently, global pooling is dissected along two directions, and each channel is individually encoded. This process ensures the preservation of spatial information. The formulation governing the vertical and horizontal directions is denoted as

$$x_c^h(h) = \frac{1}{W} \sum_{0\leq i<W} f_c(h, i), \tag{1}$$

$$x_c^w(w) = \frac{1}{H} \sum_{0\leq j<H} f_c(j, w), \tag{2}$$

where $f_c(h, i)$ and $f_c(j, w)$ represent the spatial information encoding of the c channel along the horizontal and vertical directions, respectively.

These two transformations aggregate spatial features in vertical and horizontal directions, respectively. They enable attention blocks to capture long-range dependencies in one spatial direction while preserving accurate location information in the other spatial direction. These two transformations are concatenated and used as inputs to the shared convolution transform function. It is formulated as

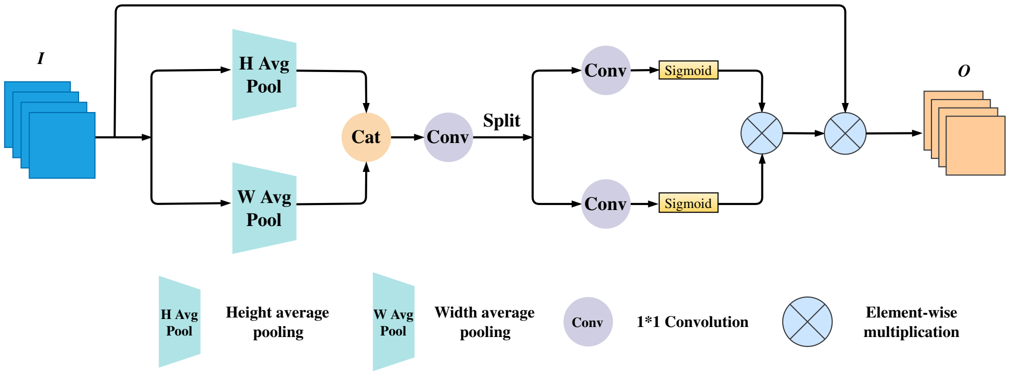

Fig. 3 Architecture of the SA module.

$$\mathbf{z} = \delta(F_1([\mathbf{g}^h, \mathbf{g}^w])), \tag{3}$$

where δ denotes the nonlinear activation function, F_1 represents 1×1 convolutional transformation function, and z represents the intermediate feature tensor encoding spatial information along two spatial directions. The spatial dimension's feature tensor is partitioned into two disjoint tensors, $\mathbf{n}^h \in \mathbb{R}^{C/r \times H}$ and $\mathbf{n}^w \in \mathbb{R}^{C/r \times W}$, with r denoting the reduction ratio of the control block's size. Subsequently, two 1×1 convolution operations are applied to preserve the same final output as the input:

$$\mathbf{q}^h = \sigma(F_h(\mathbf{n}^h)), \tag{4}$$

$$\mathbf{q}^w = \sigma(F_w(\mathbf{n}^w)), \tag{5}$$

where σ symbolizes the sigmoid function, F_h and F_w represent 1×1 convolutional transformations along the horizontal and vertical directions, respectively. Subsequently, the acquired $\mathbf{q}^h$ and $\mathbf{q}^w$ serve as attention weights. Finally, the result of the attention obtained is as follows:

$$o_c(i,j) = f_c(i,j) \times q_c^h(i) \times q_c^w(j), \tag{6}$$

where q^h and q^w represent attention weights in horizontal and vertical directions, respectively.

3.3 Deformation Awareness Module

The structure of the DA module is depicted in Fig. 4. The DA module incorporates deformable convolution and attention mechanisms to capture long-range dependencies. In the case of conventional 2D convolution, the process is essentially executed in two steps. Initially, the input feature map undergoes sampling using a regular grid denoted as $\mathcal{P}$. Subsequently, the weighted sample values are aggregated. The grid $\mathcal{P}$ defines the size and extent of the receptive field. For each position $\mathbf{r}_0$ on the input feature map z, it can be denoted as

$$\mathbf{z}(\mathbf{r}_0) = \sum_{\mathbf{r}_n \in \mathcal{P}} \mathbf{w}(\mathbf{r}_n) \cdot \mathbf{x}(\mathbf{r}_0 + \mathbf{r}_n). \tag{7}$$

In deformable convolution, the regular grid $\mathcal{P}$ is replaced by an offset $\{\Delta\mathbf{r}_n | n = 1, \dots, N\}$. The above position $\mathbf{z}(\mathbf{r}_0)$ can be expressed as

$$\mathbf{z}(\mathbf{r}_0) = \sum_{\mathbf{r}_n \in \mathcal{P}} \mathbf{w}(\mathbf{r}_n) \cdot \mathbf{x}(\mathbf{r}_0 + \mathbf{r}_n + \Delta\mathbf{r}_n). \tag{8}$$

Moreover, deformable convolution is sampled at the offset position $\mathbf{r}_n + \Delta\mathbf{r}_n$ when sampling is performed. As the offset $\Delta\mathbf{r}_n$ is typically fractional, Eq. (8) is implemented via bilinear interpolation as

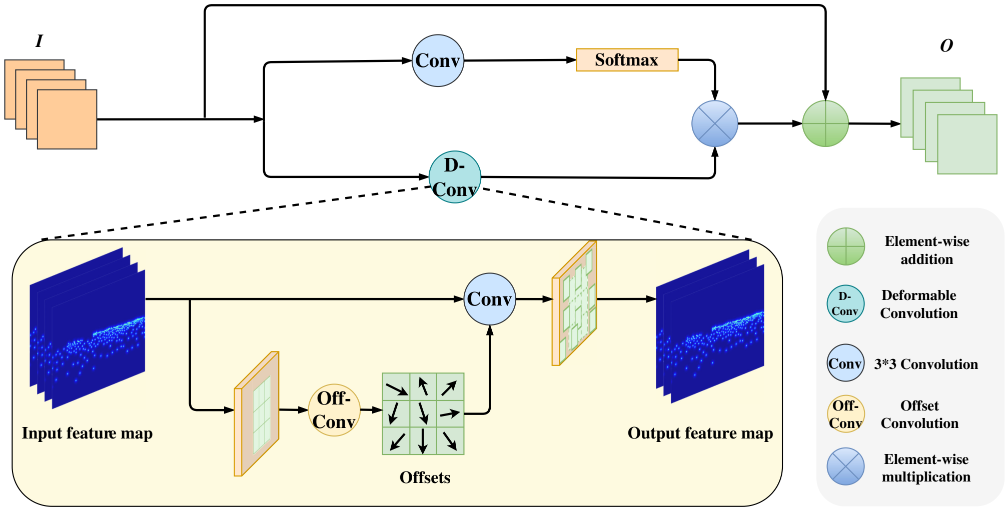

Fig. 4 Architecture of the DA module.

$$\mathbf{x}(\mathbf{r}) = \sum_{\mathbf{k}} F(\mathbf{k}, \mathbf{r}) \cdot \mathbf{x}(\mathbf{k}), \tag{9}$$

where **r** represents an arbitrary position in the feature map. The function F represents bilinear interpolation.

As depicted in Fig. 4, the offset is derived through offset convolution. Subsequently, the obtained offset is employed in the 3×3 convolution operation to accomplish the convolutional deformation. The attention acquired through convolution and softmax is multiplied to yield an offset feature map. Finally, the conclusive output is attained via the residual connection.

3.4 Ground Truth Generation

The FIDTM[29] is employed to derive ground truth values. This approach facilitates a more accurate head annotation and density representation. The formulation for the generation of the distance transform graph is denoted as follows:

$$P(x, y) = \min_{(x', y') \in B} \sqrt{(x - x')^2 + (y - y')^2}, \tag{10}$$

where B represents all header comments, and (x, y) represents arbitrary pixel. Then this method uses the inverse function to control the distance change and generates the inverse distance transform map:

$$I = \frac{1}{P(x, y)^{(\alpha \times P(x,y) + \beta)} + C}, \tag{11}$$

where α and β are set to 0.02 and 0.75, respectively. C is a nonzero constant.

3.5 Loss Function

The mean squared error (MSE) loss function is employed to minimize the disparity between the predicted and actual values. It is formulated as follows:

$$L_{\mathrm{MSE}} = \frac{1}{N} \sum_{i=1}^{N} \| y_i - \hat{y}_i \|_2^2, \tag{12}$$

where N denotes the total number of crowd in the image, y_i represents the predicted value, and $\hat{y}_i$ represents the ground true value.

4 Experiments

4.1 Datasets

The performance of the proposed method was assessed using five datasets. ShanghaiTech[8] dataset is bifurcated into two parts, denoted as part A and part B, comprising a cumulative total of 1198 images. Specifically, part A encompasses 300 images in the training set and 182 images in the test set, whereas part B incorporates 400 images for the training set and 316 images for the test set. UCF_CC_50[31] dataset contains 50 grayscale crowded images with a count range of 96 to 4633. UCF-QNRF[32] dataset contains 1535 images, including many high-resolution in diverse scenarios. The minimum and maximum counts are 49 and 12,865, respectively. JHU-Crowd++[33] dataset comprises a total of 4822 images. It is categorized into three sets: a training set with 2722 images, a test set with 1600 images, and a validation set with 500 images. This dataset encapsulates diverse scenarios, with the counting number from 0 to 25,791.

4.2 Implementation Details

The training process commences with random cropping and horizontal flipping. In ShanghaiTech datasets, the image cropping size is configured as 256×256. Conversely, for the UCF_CC_50, UCF-QNRF, and JHU-Crowd++ datasets, the cropping size is set to 512×512. During the training phase, the batch size for ShanghaiTech is defined as 8, and the other three datasets adopt a batch size of 4. To optimize the experimental execution and achieve superior results, we set the decay rate to 0.995 and the learning rate to 1×10^{-4}. The evaluation was executed in PyTorch framework equipped with two NVIDIA GTX 3090 GPUs.

4.3 Evaluation Protocols

In assessing the experiment's performance, we employ the mean absolute error (MAE) and the root-mean-square error (RMSE). They are formulated as follows:

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N} |y_i - \hat{y}_i|, \tag{13}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} |y_i - \hat{y}_i|^2}, \tag{14}$$

where N represents the number of test images. y_i and $\hat{y}_i$ represent the predicted and ground truth values for the i'th image, respectively.

To evaluate the efficiency of the models, we adopt the indicator of parameters, FLOPs,[34,35] inferring time, and frames per second (FPS).

4.4 Comparison on Crowd Counting

4.4.1 *Comparison of crowd counting accuracy*

We compared the obtained results with those achieved by competitive models. The detailed comparison results are presented in Table 1. On the part A dataset, SDANet scores 54.9 and 90.4 in MAE and RMSE surpass all the competitors. Compared with PESSNet,[5] the MAE and RMSE of SDANet decreased by 2.4 and 5.5, respectively. On the part B dataset, SDANet scores 7.1 and 12.0 in MAE and RMSE, only inferior to the PESSNet method. The MAE and RMSE are increased by 0.7 and 2.1, respectively.

On the UCF-QNRF dataset, the proposed SDANet achieves competitive scores of 107.3 and 195.5 in MAE and RMSE. The proposed SDANet scored 5.1% higher in MAE and 14.0% higher in RMSE compared to the first-placed SFCN.[4] The reason is that SFCN adopts dilation to enlarge the receptive field, to capture a larger range crowd. However, the SA module cannot be aware of large receptive fields. The results on the UCF-QNRF dataset are inferior to SFCN's results.

Table 1 Objective results on crowd counting datasets.

	Part A		Part B		UCF_CC_50		UCF-QNRF		JHU++	
Method	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
MCNN[8]	110.2	173.2	26.4	41.3	377.6	509.1	277.0	426.0	188.9	483.4
Switch-CNN[36]	90.4	135.0	21.6	33.4	318.1	439.2	—	—	—	—
A-CCNN[37]	85.4	124.6	19.2	31.5	—	—	367.3	—	171.2	453.1
SFCN[4]	64.8	107.5	7.6	13.0	214.2	318.2	**102.0**	**171.4**	77.5	297.6
RAZ[38]	65.1	106.7	8.4	14.1	—	—	116.0	195.0	—	—
DA^{2}Net[3]	74.1	128.4	7.9	13.2	169.5	237.0	111.7	204.3	—	—
PESSNet[5]	57.3	95.9	**6.4**	**9.9**	—	—	—	—	**70.91**	**256.33**
TEDNet[39]	64.2	109.1	8.2	12.8	249.4	354.5	113.0	188.0	75.0	299.9
LSC-CNN[40]	66.4	117.0	8.1	12.7	—	—	120.5	218.2	112.7	454.4
MUD-iKNN[41]	68.0	117.7	13.4	21.4	237.7	305.7	104.0	172.0	—	—
DUBNet[42]	64.6	106.8	7.7	12.5	243.8	329.3	105.6	180.5	133.5	416.5
SUA-fully[43]	66.9	125.6	12.3	17.9	—	—	119.2	213.3	—	—
PCCNet[34]	73.5	124.0	11.0	19.0	240.0	315.5	148.7	247.3	—	—
CG-DRCN[44]	64.0	98.4	8.5	14.4	—	—	112.2	176.3	71.0	278.6
SDANet (ours)	**54.9**	**90.4**	7.1	12.0	**104.1**	**154.4**	107.3	195.5	71.8	287.0

The best results are highlighted in bold.

But the score shows SDANet is still competitive against a high-density and large-scale variations crowd. On the UCF_CC_50 dataset, SDANet scores the best results of 104.1 and 154.4 in MAE and RMSE, respectively. Compared with SFCN,[4] it improves the MAE and RMSE by 51.4% and 51.5%, respectively.

On the JHU-Crowd++ dataset, the SDANet scores 71.8 and 287 in MAE and RMSE, respectively. Although the results are not optimal compared with the CG-DRCN and PESSNet, the proposed method is still competitive. It is worth mentioning that the CG-DRCN[44] model was presented alongside the JHU-Crowd++ dataset, including images captured under extreme weather conditions. It incorporates uncertainty-guided residual estimation conditioned on image-level labels to integrate weather data into training. Furthermore, additional image-level labels in the dataset refine the uncertainty-based confidence weighting model, limiting it to specific tags for enhanced accuracy in difficult weather. The CG-DRCN method is superior to the SDANet on the JHU-Crowd++ but inferior to SDANet on other datasets.

Figure 5 illustrates the visualized results generated by SDANet across various datasets. It proves that the estimated map and counting number are closely approximated to the ground

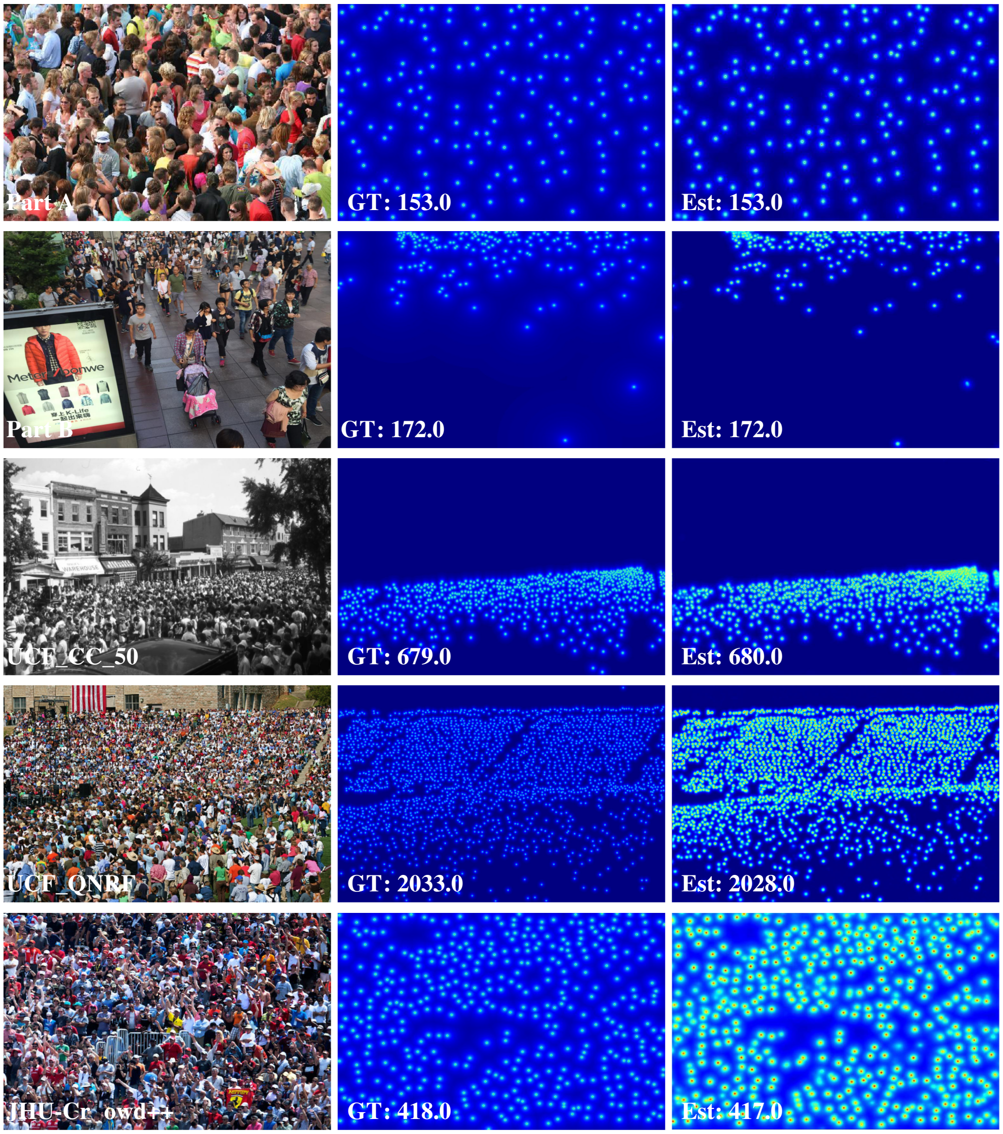

Fig. 5 Visualization result of the proposed SDANet on different datasets. The sequence from left to right: input image, the corresponding ground truth, and the predicted density map. The sequence from top to bottom: visual results of the Shanghai part A, Shanghai part B, UCF_CC_50, UCF-QNRF, and JHU-Crowd++ datasets.

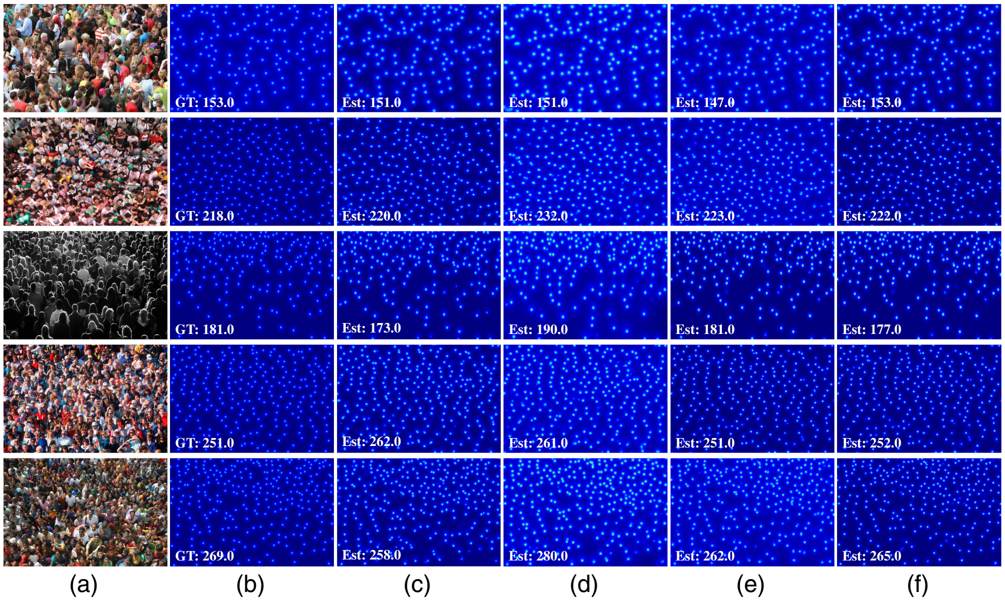

Fig. 6 Subjective results of different models on Shanghai part A datasets: (a) test images, (b) ground truth, (c) SA^{2}Net, (d) GGANet, (e) SCPNet, and (f) SDANet (ours).

truth values. The proposed SDANet is a density map-based model. It works by converting the image into a density map and then estimating the number of people by integrating the density map. However, the predicted crowd distribution is still not entirely aligned with that in the real world, which is unavoidable in crowd counting tasks. The task of crowd location belongs to a high-level crowd analysis task, i.e., crowd localization.[45,46] Compared with crowd localization, crowd counting tasks are primarily concerned with capturing the overall characteristics of a group rather than accurately identifying spatial location information. Especially in dense crowd regions, the density map cannot accurately reflect the sample location due to the severe occlusion. Therefore, there is a deviation between the predicted crowd density distribution and the actual spatial location.

We made a comparative analysis with SOTA models, i.e., SA^{2}Net,[47] GGANet,[48] and SCPNet[46] in Shanghai part A. Figure 6 shows the subjective results of SDANet with different models. The images encompass scale variation and head deformation issues. It proves that the proposed SDANet achieves accurate counting, and it is competitive in dense crowd counting.

4.4.2 *Comparison of crowd counting efficiency*

To evaluate the efficiency, we compared the SDANet with FIDTM,[29] GGANet,[48] SCPNet,[46] and SA^{2}Net[47] on an RTX 3090 GPU. The input size is set to 576 × 768. The complexity analysis results are shown in Table 2. It shows that the results obtained by the proposed SDANet are

Table 2 Comparison results of different models in Params, FLOPs, inferring time, and FPS on Shanghai part A.

Methods	Params (M)	FLOPs (G)	Time (ms)	FPS
FIDTM[29]	**66.6**	**240.3**	62.3	16.1
GGANet[48]	66.9	240.3	30.2	**33.1**
SCPNet[46]	75.4	449.0	97.6	10.2
SA^{2}Net[47]	79.2	589.7	117.7	8.5
SDANet (ours)	68.5	254.2	**38.0**	26.3

The best results are highlighted in bold.

Table 3 Objective ablation studies on the DA and SA modules.

Methods	MAE	RMSE
Baseline	62.1	108.8
Baseline + SA	61.4	108.7
Baseline + DA	57.9	108.8
Baseline + SA + DA(1)	59.3	102.8
Baseline + SA + DA(2)	**54.9**	**90.4**
Baseline + SA + DA(3)	57.6	94.7

The best results are highlighted in bold.

competitive. The SDANet has 68.5 M parameters and 254.2 M in FLOPs. Specifically, compared to FIDTM,[29] the proposed method reduces the inferring time by 39% and increase the FPS by 63.3%, with a minimal increase of 2.8% and 5.7% in Params and FLOPs. Although the proposed method increases the number of parameters and the amount of computation compared with FIDTM, it improves the inference time and FPS and reduces the complexity.

4.5 Ablation Studies

We conducted ablation experiments on the Shanghai part A dataset to assess the practical utility of the SA and DA modules. The comparative results are delineated in Table 3. "baseline + SA + module(·)" represents the basic network that adopts HRNet with SA module, and adds n module(s) to the HRNet.

We investigate the influence of the SA and DA modules on counting performance. As shown in Table 3, adding both SA and DA modules helps to enhance MAE and RMSE. Specifically, adding the SA module to the baseline achieved 1.1% and 0.9% improvements in MAE and RMSE, respectively. The addition of the DA module improves the MAE by 6.6%. Then the SA module and one DA module are connected in series. Counting performance has been further improved, and MAE and RMSE have been improved by 4.5% and 5.5%, respectively. When the number of DA is 2, the performance is further improved with MAE and RMSE being 54.9 and 90.4, respectively. However, when the number of DA increases to 3, the performance decreases. Thus the configuration of "baseline + SA + DA(2)" is the optimal model.

5 Failure Cases

As mentioned in Sec. 4.4.1, the proposed SDANet ranks third place on JHU-Crowd++, which is characterized by weather changes. The result demonstrates that the SDANet is hard to deal with this challenge. The subjective results of images with bad weather are shown in Fig. 7. It proves that there is a gap between the estimation and ground truth. In the case of weather changes, the environmental background may become more complex which increases the background interference. Moreover, the physical features of the crowd may be blurred or partially obscured, making it more difficult for the SDANet to detect and distinguish the crowd accurately. Considering that the proposed SDANet addresses the scale variation and head deformation, the counting performance is not optimal in scenarios with bad weather. Crowd counting under harsh weather will be our further research direction.

6 Conclusion and Perspective

This paper builds an SDANet to address the challenges of scale variations and head deformation. The SDANet consists of two key modules, i.e., the SA module and the DA module. The SA module decomposes channel attention into two spatial directions to capture long-distance dependencies and preserve precise spatial information. It effectively mitigates the impact of scale variations on counting accuracy during the feature extraction stage. Simultaneously, the DA module adapts the deformable convolution to resolve the issue of head deformation.

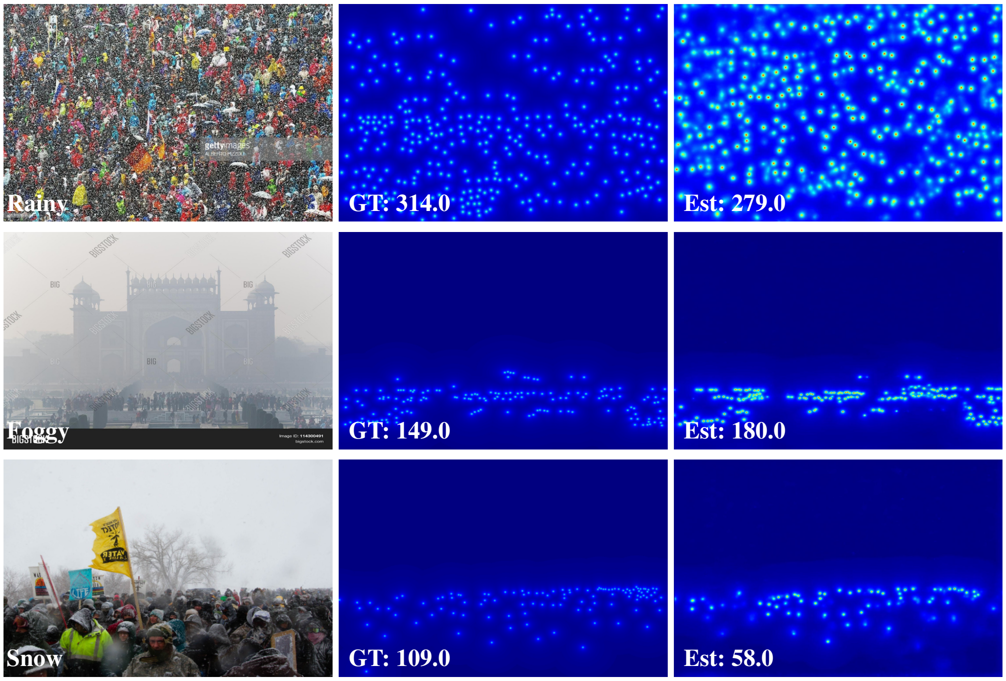

Fig. 7 Subjective results on challenging scenarios with harsh weather.

Experimental results on four benchmark datasets verify the superiority of the proposed model. In future work, we intend to improve the proposed counting model in scenarios with bad weather by introducing image enhancement modules.

Code and Data Availability

The code is available at https://github.com/sdutwjy/SDANet

References

1. J. Chen et al., "Object counting in remote sensing via selective spatial-frequency pyramid network," *Softw. Pract. Exp.* (2023).
2. V. Sindagi and V. Patel, "Generating high-quality crowd density maps using contextual pyramid CNNs," in *Proc. Int. Conf. Comput. Vis. (ICCV)*, pp. 1879–1888 (2017).
3. W. Zhai et al., "DA^2 Net: a dual attention-aware network for robust crowd counting," *Multimedia Syst.* **29**(5), 3027–3040 (2023).
4. Q. Wang et al., "Learning from synthetic data for crowd counting in the wild," in *Proc. IEEE Conf. Comput. Vis. and Pattern Recognit. (CVPR)*, pp. 8190–8199 (2019).
5. J. Yi et al., "A perspective-embedded scale-selection network for crowd counting in public transportation," *IEEE Trans. Intell. Transport. Syst.,* **25**(5), 3420–3432 (2023).
6. M. Rodriguez et al., "Density-aware person detection and tracking in crowds," in *Int. Conf. Comput. Vis.*, IEEE, pp. 2423–2430 (2011).
7. Á. García-Martín and J. M. Martínez, "People detection in surveillance: classification and evaluation," *IET Comput. Vis.* **9**(5), 779–788 (2015).
8. Y. Zhang et al., "Single-image crowd counting via multi-column convolutional neural network," in *Proc. IEEE Conf. Comput. Vis. and Pattern Recognit.*, pp. 589–597 (2016).
9. D. Babu Sam, S. Surya, and R. Venkatesh Babu, "Switching convolutional neural network for crowd counting," in *Proc. IEEE Conf. Comput. Vis. and Pattern Recognit.*, pp. 5744–5752 (2017).
10. X. Guo et al., "Spatial-frequency attention network for crowd counting," *Big Data* **10**(5), 453–465 (2022).
11. I. S. Topkaya, H. Erdogan, and F. Porikli, "Counting people by clustering person detector outputs," in *11th IEEE Int. Conf. Adv. Video and Signal Based Surveillance (AVSS)*, IEEE, pp. 313–318 (2014).
12. A. Patwal et al., "Crowd counting analysis using deep learning: a critical review," *Proc. Comput. Sci.* **218**, 2448–2458 (2023).
13. 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.* **107**, 3–16 (2018).

14. X. Guo et al., “Multiscale aggregation network via smooth inverse map for crowd counting,” *Multimedia Tools Appl.*, 1–15 (2022).
15. X. Cao et al., “Scale aggregation network for accurate and efficient crowd counting,” *Lect. Notes Comput. Sci.* **11209**, 734–750 (2018).
16. Z. Shi, P. Mettes, and C. G. Snoek, “Counting with focus for free,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, pp. 4200–4209 (2019).
17. X. Jiang et al., “Attention scaling for crowd counting,” in *Proc. IEEE/CVF Conf. Comput. Vis. and Pattern Recognit.*, pp. 4706–4715 (2020).
18. W. Zhai et al., “Group-split attention network for crowd counting,” *J. Electron. Imaging* **31**(4), 041214 (2022).
19. W. Luo et al., “Understanding the effective receptive field in deep convolutional neural networks,” in *Adv. in Neural Inf. Process. Syst.*, Vol. **29** (2016).
20. X. Jia et al., “Dynamic filter networks,” in *Adv. in Neural Inf. Process. Syst.*, Vol. **29** (2016).
21. F. Yu and V. Koltun, “Multi-scale context aggregation by dilated convolutions,” arXiv:1511.07122 (2015).
22. N. Bodla et al., “Soft-NMS–improving object detection with one line of code,” in *Proc. IEEE Int. Conf. Comput. Vis.*, pp. 5561–5569 (2017).
23. X. Zhu et al., “Flow-guided feature aggregation for video object detection,” in *Proc. IEEE Int. Conf. Comput. Vis.*, pp. 408–417 (2017).
24. W. Zhou and Z. Chen, “Deep multi-scale features learning for distorted image quality assessment,” in *IEEE Int. Symp. Circuits and Syst. (ISCAS)*, IEEE, pp. 1–5 (2021).
25. L. Sun et al., “MCNet: multiscale visible image and infrared image fusion network,” *Signal Process.* **208**, 108996 (2023).
26. X. Guo et al., “Crowd counting in smart city via lightweight ghost attention pyramid network,” *Future Gen. Comput. Syst.* **147**, 328–338 (2023).
27. H. Zhao et al., “MSR-fan: multi-scale residual feature-aware network for crowd counting,” *IET Image Process.* **15**, 3512–3521 (2021).
28. D. Guo et al., “DADNet: dilated-attention-deformable convnet for crowd counting,” in *Proc. 27th ACM Int. Conf. Multimedia*, pp. 1823–1832 (2019).
29. D. Liang et al., “Focal inverse distance transform maps for crowd localization,” *IEEE Trans. Multimedia* **25**, 6040–6052 (2022).
30. J. Wang et al., “Deep high-resolution representation learning for visual recognition,” *IEEE Trans. Pattern Anal. Mach. Intell.* **43**(10), 3349–3364 (2020).
31. H. Idrees et al., “Multi-source multi-scale counting in extremely dense crowd images,” in *Proc. IEEE Conf. Comput. Vis. and Pattern Recognit.*, pp. 2547–2554 (2013).
32. H. Idrees et al., “Composition loss for counting, density map estimation and localization in dense crowds,” *Lect. Notes Comput. Sci.* **11206**, 532–546 (2018).
33. V. A. Sindagi, R. Yasarla, and V. M. Patel, “JHU-Crowd++: large-scale crowd counting dataset and a benchmark method,” *IEEE Trans. Pattern Anal. Mach. Intell.* **44**(5), 2594–2609 (2020).
34. J. Gao, Q. Wang, and X. Li, “PCC Net: perspective crowd counting via spatial convolutional network,” *IEEE Trans. Circuits Syst. Video Technol.* **30**, 3486–3498 (2020).
35. P. Wang et al., “MobileCount: an efficient encoder-decoder framework for real-time crowd counting,” *Neurocomputing* **407**, 292–299 (2020).
36. D. B. Sam, S. Surya, and R. V. Babu, “Switching convolutional neural network for crowd counting,” in *Proc. IEEE Conf. Comput. Vis. and Pattern Recognit. (CVPR)*, pp. 4031–4039 (2017).
37. S. A. Kasmani et al., “A-CCNN: Adaptive CCNN for density estimation and crowd counting,” in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, pp. 948–952 (2018).
38. C. Liu, X. Weng, and Y. Mu, “Recurrent attentive zooming for joint crowd counting and precise localization,” in *CVPR*, pp. 1217–1226 (2019).
39. X. Jiang et al., “Crowd counting and density estimation by trellis encoder-decoder networks,” in *Proc. IEEE Conf. Comput. Vis. and Pattern Recognit. (CVPR)*, pp. 6126–6135 (2019).
40. D. B. Sam et al., “Locate, size, and count: accurately resolving people in dense crowds via detection,” *IEEE Trans. Pattern Anal. Mach. Intell.* **43**, 2739–2751 (2021).
41. G. Olmschenk, H. Tang, and Z. Zhu, “Improving dense crowd counting convolutional neural networks using inverse k-nearest neighbor maps and multiscale upsampling,” in *15th Int. Joint Conf. Comput. Vis., Imaging and Comput. Graph. Theory and Appl.*, Vol. **5** (2020).
42. M. Hwan Oh, P. Olsen, and K. Ramamurthy, “Crowd counting with decomposed uncertainty,” in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, pp. 11799–11806 (2020).
43. Y. Meng et al., “Spatial uncertainty-aware semi-supervised crowd counting,” in *Proc. Int. Conf. Comput. Vis. (ICCV)*, pp. 15549–15559 (2021).
44. V. A. Sindagi, R. Yasarla, and V. M. Patel, “JHU-Crowd++: large-scale crowd counting dataset and a benchmark method,” *IEEE Trans. Pattern Anal. Mach. Intell.* **44**(5), 2594–2609 (2022).

45. D. Liang, W. Xu, and X. Bai, "An end-to-end transformer model for crowd localization," *Lect. Notes Comput. Sci.* **13661**, 38–54 (2022).
46. W. Zhai et al., "Scale-context perceptive network for crowd counting and localization in smart city system," *IEEE Internet Things J.* **10**(21), 18930–18940 (2023).
47. W. Zhai et al., "Scale attentive aggregation network for crowd counting and localization in smart city," *ACM Trans. Sens. Netw.* (2024).
48. X. Guo et al., "Object counting via group and graph attention network," *IEEE Trans. Neural Netw. Learn. Syst.*, 1–12 (2023).

Jianyong Wang is pursuing his MS degree at the School of Electrical and Electronic Engineering, Shandong University of Technology, Zibo, China. His research interests include crowd counting, computer vision, and deep learning.

Xiangyu Guo is pursuing his MS degree at the School of Electrical and Electronic Engineering, Shandong University of Technology, Zibo, China. His research interests include smart city systems, computer vision, and deep learning.

Qilei Li is a PhD student in computer science at Queen Mary University of London. Previously, he received his MS degree from Sichuan University in 2020. His research interests include computer vision and deep learning, particularly focusing on person ReID, video/image enhancement. He is a student member of IEEE, and he serves as a reviewer for *Information Fusion, IEEE TIM, IEEE Access, Concurrency, and Computation: Practice and Experience*, and *Multimedia System*.

Ahmed M. Abdelmoneim is a member of the School of Electronic Engineering and Computer Science at Queen Mary University of London. He has a PhD in computer science and engineering from Hong Kong University of Science and Technology advised by Prof. Brahim Bensaou. His current research focus involves designing and prototyping networked and distributed systems of the future, in particular, he is interested in developing methods and techniques to improve and enhance the performance of networked and distributed systems.

Mingliang Gao received his PhD in communication and information systems from Sichuan University. He is now an IEEE senior member and an associate professor at Shandong University of Technology. He has been the principal investigator for a variety of research funding, including the National Natural Science Foundation, the China Postdoctoral Foundation, and National Key Research Development Project. His research interests include computer vision, machine learning, and intelligent optimal control.