HAGN: Hierarchical Attention Guided Network for Crowd Counting

ZUODONG DUAN\textsuperscript{1}, YUJUN XIE\textsuperscript{2}, AND JIAHAO DENG\textsuperscript{1}

\textsuperscript{1}School of Mechatronical Engineering, Beijing Institute of Technology, Beijing 100081, China
\textsuperscript{2}School of Computer Science and Technology, Beijing Institute of Technology, Beijing 100081, China

Corresponding author: Jiahao Deng (bitdjh@bit.edu.cn)

ABSTRACT

In recent years, deep learning based crowd counting networks have achieved significant progress. However, most of them generate rough crowd density maps due to low-resolution features used for estimating crowd distribution, which affects the performance of crowd counting. To solve this problem, in this paper, we propose a Hierarchical Attention Guided Network (HAGN) for crowd counting. We apply the first 13 layers of VGG-16 to extract base features. Then, the extracted features are processed by the Hierarchical Attention Mechanism (HAM), which guided the extracted features to enlarge step by step via our proposed attention guided branch. Finally, the outputs of HAM are fed to 1 × 1 convolutional layer for final crowd density estimation. Experiments are performed on ShanghaiTech and UCF-QNRF datasets, and our HAGN achieves promising performance compared with the other state-of-the-art methods on crowd counting and crowd localization, respectively.

INDEX TERMS

Crowd counting, crowd density estimation, crowd localization, hierarchical attention mechanism.

I. INTRODUCTION

Crowd counting is the basis of crowd analysis and scene understanding [1], [2]. However, getting accurate crowd numbers in realistic application scenarios is a challenging task due to scale variations of the crowd head.

Many efforts for improving the scale variation problem of the crowd have been studied. One approach adopts different network structures to model scale variation, such as Multi-column Convolutional Neural Network (MCNN) [25], Switching Convolutional Neural Network (Switching CNN) [26] and Contextual Pyramid CNN (CP-CNN) [57] using multi-column convolutional neural networks with different structures to model scale variation of the crowd. However, multi-column convolutional networks generate redundant features and consume a lot of computation [29]. Another approach adopts multi-scale extraction modules to model scale information, such as Context-Aware Network (CAN) [20] using spatial pyramid pooling to obtain contextual information for scale variation of the crowd. Although the above works achieve great performance, however, most of them directly upsample the output from the network to regress the final crowd density estimation, which generates a rough density map and affects the final counting performance.

Therefore, in this paper, we propose a Hierarchical Attention Guided Network (HAGN) to generate high-resolution crowd density map for crowd counting. We firstly apply the 13 layers of VGG-16 to extract features. Then, the extracted features are fed in two separate ways. One way is that the extracted features are sent to attention module for obtaining global context information. Then, the global context information guides the extracted feature to enlarge step by step via our proposed attention guided branch. Finally, the outputs of HAM are fed to 1 × 1 convolutional layer for final crowd density estimation. Experiments are performed on ShanghaiTech and UCF-QNRF datasets, and the results of the experiment demonstrate the effectiveness of our method. We also evaluate the crowd localization of our method on UCF-QNRF dataset, and we achieve superior results against the other methods.

In a word, the contributions of our method are as follows:

1) We propose the HAGN, which adopts the base feature of large receptive fields to guide the density map regression process step by step and to predict high-resolution crowd density maps for crowd counting.
2) We propose HAM module, which contains three attention guided branches to supplement rich contextual information for the generation of crowd density maps.
3) The proposed HAGN achieves good performance on ShanghaiTech Part A/B and UCF-QNRF datasets compared with other state-of-the-art methods.

The remainder of this paper is organized as follows: Section II introduces previous progress of crowd counting methods and attention module. Section III introduces HAGN and HAM. The experiment and ablation studies are performed in Section IV. We draw the conclusion of this paper in Section V.

II. RELATED WORK

A. CROWD COUNTING

Recently, the deep learning methods represented by CNNs have developed rapidly and have achieved great success in various fields of computer vision.

Many excellent crowd counting methods arise in recent years. The improvements of the proposed methods are mainly in network structure [1]–[3], [5]–[11], [13]–[23], [30], [31], [41], [44], [49], [50], [52], [54], loss function design [4], [12], [27], more challenging crowd counting datasets [24], [51], efficient crowd counting framework [32] and so on.

Zhao et al. [1] proposed a new kind of network structure with deformable convolution for crowd counting tasks. Sang et al. [2] proposed a new model by improving the Scale-adaptive CNN (SaCNN) architecture with a backbone of fixed small receptive fields [43]. Cheng et al. [3] proposed a new kind of learning strategy named Multi-column Convolutional Neural Network (McML) for multi-column networks, which could effectively solve the multi-scale learning problem of the network, and has the advantages of less parameter and be less prone to overfitting. Sindagi and Patel [5] proposed advanced counting methods which consists of multi-level and multi-directional information fusion from multi-layer networks. Zou et al. [6] proposed a new structure named “temporal channel-aware” (TCA) block, which could use time interdependence for crowd counting of video sequences. Hossain et al. [7] proposed a new issue termed as a single scene-specific crowd count and a corresponding new learning algorithm. Shi et al. [8] proposed a novel attention module that could guide the counting network to the area of interest with supervised segmentation. Xu et al. [9] proposed Learning to Scale Module (L2SM) to solve the density variation issue in crowd counting. Xiong et al. [10] proposed Spatial Divide-and Conquer Network (S-DCNet), which has an excellent performance on counting data sets for multiple cross-categories. Liu et al. [11] proposed Deep Structured Scale Integration Network (DSSINet), which could respond to crowd density changes through structured and multi-level feature learning and corresponding loss function optimization. Yan et al. [13] proposed a new network structure named Perspective-Guided Convolution Networks (PGCNet), which could solve the perspective effect problem of crowd density changes in large scenes. Liu et al. [14] proposed a novel network architecture named Recurrent Attentive Zooming Network, which could solve the counting problem of crowded scenes by amplifying the blurred area to a high resolution by a high-resolution method and a specific loss function. Lian et al. [15] proposed a new network architecture named regression guided detection network (RDNet), which could estimate the quantity and location of heads in the crowd scene using RGB-D data. Wan et al. [16] proposed a residual regression frame-work which use relevant information between samples. Zhang and Chang [17] proposed a novel network, which could combine the information of multiple cameras to estimate the crowd density of the scene. Zhao et al. [18] proposed a new crowd counting method that could use heterogeneous attributes to solve the problem of various scales, complex backgrounds, and occlusion. Jiang et al. [19] proposed Trellis Encoder-Decoder Networks (TEDnet), which could estimate a high-quality crowd density map for counting. Liu et al. [20] proposed a novel network, which could encode context information of different scale. Shi et al. [21] proposed perspective-aware convolutional neural network (PACNN), which could add perspective information to the estimated number to effectively solve multiple-scale issues. Liu et al. [22] proposed a novel detection-based network which could estimate crowd density by means of detection. Liu et al. [23] proposed an attention-injective deformable convolutional network called ADCrowdNet, which could solve the low precision issue of noisy scenes for the high-density crowd.

B. ON-DEVICE DETECTION

Wan et al. [23] proposed a novel network architecture named SPatial Aware-ness Network (SPANet) and a new kind of loss function called Maximum Excess over Pixels Loss, which could effectively combine spatial context information for crowd counting task. Ma et al. [12] proposed a new and useful loss function named Bayesian loss, which could measure the density contribution by establishing a probability model by labeling information. Gao et al. [32] designed an open source crowd counting framework based on PyTorch named C^3 3 Framework. For the crowd localization task, Zheng et al. [64] firstly obtained the density map using the sliding window on the image and then used integer programming to locate objects on the density maps. Liu et al. [55] proposed a reliable end-to-end network called DecideNet that can estimate the localization of crowd heads and the crowd density by detecting and regression, respectively. Liu et al. [14] proposed a Recurrent Attentive Zooming Network that can simultaneously. Lian et al. [15] proposed density map regression guided detection network (RDNet), which can perform crowd counting and localization simultaneously. Liu et al. [22] proposed a new deep detection network based on point supervision, which can simultaneously detect
the localization of crowd heads and count in a crowded scene.

Different from above crowd counting methods, we propose HAGN to reuse the base feature of the first 13 layers of VGG-16 and recover the density map through the hierarchical attention mechanism.

**B. ATTENTION MODULE**

Attention modules are often used in visual tasks [60]. Mnih et al. [33] adds attention information to the RNN-based network firstly. Xu et al. [34] proposed a new pooling method that could use attention weights to select regions of interest or spatial features. Vaswani et al. [35] proposed a novel transformer structure that could effectively obtain global dependencies. Chen et al. [36] proposed a novel model which could obtain multi-level attention information. Wang et al. [37] proposed a new network architecture that could calculate the response of each point to incorporate all point weight information into the feature map. Hou et al. [39] proposed a novel attention distillation approach to enhance the representation learning of CNN-based lane detection models.

CBAM [38] and SEblock [40] are most extensive attention modules in recent years. Woo et al. [38] proposed the Convolutional Block Attention Module (CBAM), which could achieve attentional calculation of spatial and channel dimensions in the forward process of convolutional neural networks. Hu et al. [40] proposed an attention module named Squeeze-and-Excitation block (SEblock), which could be adaptive adjustment the channel feature response by modeling the dependencies between channels.

**III. METHOD**

In this section, we firstly introduce the overview of the proposed HAGN. Then, we demonstrate the details of the hierarchical attention mechanism used in the HAGN and the associated attention modules.

**A. OVERVIEW**

The overview of HAGN is shown in Fig.1. The HAGN mainly consists of two parts: the encoder part and the decoder part. The details of HAGN are shown in Table 1. \(k(3, 3) - c_{64} - s_{1} - R\) means that the layer contains \(3 \times 3\) kernel size convolutional operation, 64 output channels and stride size of 1. The \(R\) represents the ReLU.

**TABLE 1. The network architecture of the proposed HAGN (k = kernal size, c = output channels, s = stride size, R = ReLu Layer).**

| Main network structure of the HAGN |
|-----------------------------------|
| **Encoder module**               |
| conv: \(k(3,3)-c_{64}-s_{1}-R\) \times 2                        |
| Max pooling                      |
| conv: \(k(3,3)-c_{128}-s_{1}-R\) \times 2                       |
| Max pooling                      |
| conv: \(k(3,3)-c_{128}-s_{1}-R\) \times 3                       |
| Max pooling                      |
| conv: \(k(3,3)-c_{512}-s_{1}-R\) \times 3                       |
| Max pooling                      |
| conv: \(k(3,3)-c_{512}-s_{1}-R\) \times 3                       |

| **Decoder module**               |
|----------------------------------|
| 1st attention guided branch      |
| Upsampling                       |
| 2nd attention guided branch      |
| Upsampling                       |
| 3rd attention guided branch      |
| Upsampling                       |
| conv: \(k(1,1)-c_{1}-s_{1}-R\)  |
| Upsampling                       |

**Encoder part:** The first 13 convolutional layers of VGG-16 are used as local feature extractor.

**Decoder part:** We design the attention guided branch to recover the loss of contextual information as shown in Fig.1.
and Fig.2. Each attention guided branch will concatenate feature maps from the output of the attention module and the upper layer. Besides, we add a residual connection after the fusion of the feature maps at each branch. The proposed HAGN adopts step-by-step upsampling mode, and the ablation studies on different upsampling modes are shown in Section IV-E. Finally, the high-resolution crowd density maps are regressed by one $1 \times 1$ convolutional layer.

**B. HIERARCHICAL ATTENTION MECHANISM**

Hierarchical attention mechanism used in the proposed HAGN, as shown in Fig.2.

Hierarchical attention mechanism mainly consists of three attention guide branch, as shown in Fig.2. The workflow of each attention guided branch is as follows:

$$F'' = F^{RA} \otimes F'.$$

$$F^R = F'' \oplus F'.$$  \hspace{1cm} (1)

where $F$ represents the input feature map. $F'$ represents the feature map predicted by $3 \times 3$ convolution layer. $F^A$ represents refined feature map predicted by attention module. $F^{RA}$ represents refined feature map predicted by one $1 \times 1$ convolutional layer and one upsampling layer. $\otimes$ represents element-wise multiplication. $F''$ represents the attention refined feature map, which is the element-wise multiplication of $F'$ and $F^{RA}$. $\oplus$ represents element-wise summation. $F^R$ represents the final refined feature map, which is the element-wise summation of $F'$ and $F''$.

For the attention module, We apply CBAM and SEblocks to get contextual information, respectively. CBAM module can encode spatial-wise attention information and channel-wise attention information at the same time. The operation process of CBAM is shown as follows:

$$U = M_C(F) \otimes F.$$  \hspace{1cm} (2)

where the refined feature map $F^A$ obtains channel-wise attention information and spatial-wise attention information in a specific order. $F$ represents the input feature map. $M_C(F)$ represents the predicted channel attention feature map by the channel attention module of CBAM. $U$ represents the refined feature map, which is the element-wise multiplication of $F$ and $M_C(F)$. $M_S(U)$ represents the predicted spatial attention feature map by the spatial attention module of CBAM. $F^A$ represents the channel and spatial refined feature map, which is the element-wise multiplication of $U$ and $M_S(U)$.

where $S$ represents the input feature map. $C(S)$ represents the predicted channel attention feature map by the channel attention module of SEblock. $S^A$ represents the refined feature map, which is the element-wise multiplication of $S$ and $C(S)$.

**C. LOSS FUNCTION**

$L2$ loss is chosen as loss function.

$$L = \frac{1}{2N} \sum_{i=0}^{N} \| Y_i - f(x_i) \|^2.$$  \hspace{1cm} (3)

where $L$ represents the loss of the HAGN. $N$ represents the number of images. $i$ represents the $i$ th serial number of test images. $Y_i$ represents the $i$ th ground truth density map. $x_i$ represents the $i$ th image of test images. $f(x_i)$ represents the $i$ th predicted density map by the HAGN.

**IV. EXPERIMENT**

In this section, we first introduce datasets, evaluation metrics, and implementation details. Then, we show the experimental
TABLE 2. Performance comparison with the state-of-the-art methods on ShanghaiTech and UCF-QNRF datasets. ("--" denotes the results are not available).

| Method               | ShanghaiTech A |       | ShanghaiTech B |       | UCF-QNRF |       |
|----------------------|----------------|-------|----------------|-------|----------|-------|
|                      | MAE ↓          | MSE ↓ | MAE ↓          | MSE ↓ | MAE ↓    | MSE ↓ |
| Idrees et al. [48]   | -              | -     | -              | -     | 315.0    | 508.0 |
| MCNN [25]            | 110.2          | 173.2 | 26.4           | 41.3  | 227.6    | -     |
| DecideNet [55]       | 101.3          | 152.4 | 21.5           | 32.0  | 252.0    | 514.0 |
| CMTL [28]            | 97.5           | 145.1 | 20.0           | 31.1  | 220.0    | 445.0 |
| TDF-CNN [42]         | 90.4           | 135.0 | 21.6           | 33.4  | 221.0    | 444.0 |
| Switching CNN [26]   | 86.8           | 139.2 | 16.2           | 25.8  | -        | -     |
| SaCNN [43]           | -              | -     | 20.2           | 35.6  | -        | -     |
| BSAD [50]            | 73.6           | 112.0 | 13.7           | 21.4  | 132.0    | 191.0 |
| CL [51]              | 73.5           | 112.3 | 18.7           | 26.0  | -        | -     |
| SL2R [58]            | 73.5           | 118.2 | 13.6           | 21.1  | -        | -     |
| D-ConvNet-v1 [46]    | 72.0           | 106.6 | 14.4           | 23.8  | -        | -     |
| IC-GCNN [47]         | 68.5           | 116.2 | 10.7           | 16.0  | -        | -     |
| CSRNet [59]          | 68.2           | 115.0 | 10.6           | 16.0  | -        | -     |
| HAGN(Ours)           | 65.3           | 106.4 | 10.0           | 17.6  | 107.7    | 183.1 |

The qualitative results are displayed in Fig.4. From left to right, they are input images, ground truth, the results of HAGN, and pretrained VGG-16.

Following [14], [22], [51], we adopt the precision, recall and F-measure to evaluate the localization performance of our method.

B. IMPLEMENTATION DETAILS

All images are uniformly resized as 576 × 768. We choose Adam algorithm as the optimizer. The initial learning rate is set as 1 × 10⁻⁵. The learning rate decay is set as 0.995. All experiments are performed on one NVIDIA GTX TITAN Xp. The batch size is set as 2, 6, 1 on ShanghaiTech Part A/B and UCF-QNRF datasets, respectively.

C. CROWD COUNTING

For counting, we compare HAGN with other state-of-the-art methods [25], [26], [28], [29], [42], [43], [45]–[48], [51], [55], [56], [58], [59] on ShanghaiTech Part A/B and UCF-QNRF datasets, as shown in Table 2. We see that the HAGN gets the first in terms of MAE and MSE on ShanghaiTech Part A dataset and gets the first in terms of MAE and second place in terms of MSE on ShanghaiTech Part B dataset. The HAGN gets the first in terms of MAE and MSE on UCF-QNRF dataset.

The qualitative results are displayed in Fig.4. From left to right, they are input images, ground truth, the results of HAGN, and pretrained VGG-16.

D. CROWD LOCALIZATION

We perform crowd localization task on UCF-QNRF dataset. Firstly, we adopt nonmaximal suppression on the crowd density map to extract the local maximums, which are the output of the head locations. When local maximums are found, we use 1-1 matching between predicted heads locations and ground truth locations to calculate precisions and recall. We use different pixel thresholds to evaluate the matching results. If the distance between the predicted head location and the ground truth location is less than the pixel threshold, it is marked as True Positive. If the distance

The mean absolute error (MAE) and mean squared error are chosen as the evaluation metrics to evaluate the performance of our method. The definitions of MAE and MSE are as follows:

\[
MAE = \frac{1}{N} \sum_{i=0}^{N} |y_i - \hat{y}_i|.
\]

\[
MSE = \sqrt{\frac{1}{N} \sum_{i=0}^{N} (y_i - \hat{y}_i)^2}.
\]

where \(N\) denotes the number of all images in the dataset, \(y_i\) and \(\hat{y}_i\) represent respectively the ground-truth and predicted number of crowd in the \(i\)th image.

A. DATASETS AND EVALUATION METRICS

We perform experiments on the ShanghaiTech [25] and UCF-QNRF [51] datasets.

- ShanghaiTech dataset [25] consists of Part A and Part B. It consists of 1,198 annotated images with 330,165 labels. Specifically, the images in Part A are selected from the internet and have different resolutions, which contains 300 train images and 182 test images with crowd head numbers varying from 33 to 3139. Part B is obtained from street view cameras on the sidewalk in Shanghai, China. And it contains 400 train images and 316 test images with crowd head numbers varying from 9 to 578.

- UCF-QNRF dataset [51] consists of 1535 high-resolution crowd images with altogether 1.25 million labels, which includes 1201 training images and 224 test images. The number of crowd heads in images varies from 49 to 12865.

The mean absolute error (MAE) and mean squared error are chosen as the evaluation metrics to evaluate the performance of our method. The definitions of MAE and MSE are as follows:

\[
MAE = \frac{1}{N} \sum_{i=0}^{N} |y_i - \hat{y}_i|.
\]

\[
MSE = \sqrt{\frac{1}{N} \sum_{i=0}^{N} (y_i - \hat{y}_i)^2}.
\]

where \(N\) denotes the number of all images in the dataset, \(y_i\) and \(\hat{y}_i\) represent respectively the ground-truth and predicted number of crowd in the \(i\)th image.
between the predicted head location and the ground truth location is greater than the pixel threshold, it is marked as False Positive. The ground truth location without matching is marked as False Negative. For the evaluation on the UCF-QNRF dataset, we follow [51] and show the average precision, average recall at different distance thresholds $t_d : 1, 2, 3, \ldots, 100$ pixels.

The localization results are shown in Table 3. Our proposed method outperforms MCNN [25], ResNet74 [61], DenseNet63 [62], Encoder-Decoder [63] and CL [51] in
terms of average precision and F-measure and gets the second place in terms of average recall, which demonstrates the effectiveness of our method.

The qualitative results are shown in Fig.5, where red points represent ground truth locations, and green dots represent head locations estimated by our method.

### E. ABLATION STUDIES

We perform two ablation experiments on UCF-QNRF dataset. The details are shown in Table 4 and Table 5.

### TABLE 3. Performance comparison of crowd localization on UCF-QNRF dataset.

| Method                  | Av.Precision | Av.Recall | F-measure |
|-------------------------|--------------|-----------|-----------|
| MCNN [25]               | 59.93%       | 63.50%    | 61.68%    |
| ResNet74 [61]           | 61.60%       | 66.90%    | 64.14%    |
| DenseNet63 [62]         | 70.91%       | 58.10%    | 63.87%    |
| Encoder-Decoder [63]    | 71.80%       | 62.98%    | 67.10%    |
| CL. [51]                | 75.80%       | 59.75%    | 66.82%    |
| HAGN                    | 83.99%       | 59.23%    | 68.56%    |

### TABLE 4. Ablation studies on the effectiveness of the selection of upsampling modes and attention modules for crowd counting on UCF-QRNF dataset.

| Method                                      | UCF-QRNF |
|---------------------------------------------|----------|
| VGG-16 baseline model                       | 113.1    |
| VGG-16+SE+FU                                | 129.1    |
| VGG-16+SE+SU                                | 112.5    |
| VGG-16+CBAM+FU                              | 113.9    |
| HAGN (VGG-16+CBAM+SU)                       | 107.7    |

### TABLE 5. Ablation studies on the effectiveness of the proposed base feature extraction structure for crowd counting on UCF-QRNF dataset.

| Method                                      | MAB ↓ | MSE ↓ |
|---------------------------------------------|-------|-------|
| HAGN w/ the first 10 layers of VGG-16       | 110.6 | 191.0 |
| HAGN                                         | 107.7 | 183.1 |

### FIGURE 5. Qualitative visualization results of crowd localization of our HAGN. Red points and green points denote the ground-truth and predicted locations, respectively.

### FIGURE 6. The architectures about the VGG-16 baseline model and variant model with the final upsampling mode. In (a), the VGG-16 baseline model adopts a simple decoder structure to regress crowd density maps. While in (b), the variant model take out all upsampling operations during feature maps processing and take up one upsampling at the end.

1) **ABLA TION EXPERIMENTS ON NETWORK STRUCTURE**

We explore the selection of attention modules and up-sampling modes of HAGN for crowd counting, the results are displayed in Table 4. SU represents that upsampling mode is the step to step upsampling after every attention guided branch, as shown in Fig.1. FU represents that upsampling mode is the final upsampling at the end of the network, as shown in Fig.6 (b).
VGG-16 baseline model consists of VGG-16 first 13 convolutional layers as the feature extraction module, as shown in Fig. 6 (a). VGG-16+SE+SU is the model which adopts SEblock attention module in each attention guided branch and step by step upsampling mode. VGG-16+SE+FU is the model which adopts SEblock attention module and the final upsampling mode. VGG-16+CBAM+FU is the model which adopts CBAM attention module and the final upsampling mode.

From Table 4, we see that HAGN gets better results than VGG-16+SE+FU: 107.7/183.1 v.s. 112.5/191.3 in terms of MAE/MSE. It indicates that the CBAM module is more effective than the SEblock module on contextual information encoding because the CBAM module adopts both spatial and channel attention mechanisms for modeling contextual information effectively, which is helpful to improve the crowd counting accuracy. The HAGN and VGG-16+SE+SU model get better results than the VGG-16+CBAM+FU model and VGG-16+SE+FU model. It indicates that the step by step upsampling mode is more effective than the final upsampling mode because the step by step upsampling mode integrates the contextual information in the process of generating the crowd density maps, which increases the ability of the network to recognize the crowd.

2) ABLATION EXPERIMENTS ON FEATURE EXTRACTION SELECTION

Ablation experiments on feature extraction selection is shown in Table 5. HAGN w/ the first 10 layers of VGG-16 is a model that has the decoder network structure with HAGN but adopts the first 10 layers of VGG-16 to extract base feature. The HAGN gets a better result than the model using the first 10 convolution layers of VGG-16. It indicates that the base feature using the first 13 convolution layers of VGG-16 is more effective because the perceptive field of the first 13 convolution layers is larger than that of the first 10 convolution layers, which obtains more contextual information and guides network generating crowd density maps more effectively.

V. CONCLUSION

In this paper, we proposed an effective crowd counting network named HAGN to solve the scale variation problem in crowd counting. The HAGN consists of an encoder network, which is the first 13 layers of VGG-16 and a decoder network named HAM. The HAGN has achieved excellent performance on challenge crowd counting datasets against the other state-of-the-art methods. Finally, we designed the ablation experiments in attention module selection, upsampling strategy selection, and base feature extraction selection.

REFERENCES

[1] Z. Zou, X. Su, X. Qu, and P. Zhou, “DA-net: Learning the fine-grained density distribution with deformation aggregation network,” IEEE Access, vol. 6, pp. 60745–60756, Oct. 2018.

[2] J. Sang, W. Wu, H. Luo, H. Xiang, Q. Zhang, H. Hu, and X. Xia, “Improved crowd counting method based on scale-adaptive convolutional neural network,” IEEE Access, vol. 7, pp. 24411–24419, 2019.

[3] Z.-Q. Cheng, J.-X. Li, Q. Dai, X. Wu, J.-Y. He, and A. G. Hauptmann, “Improving the learning of multi-column convolutional neural network for crowd counting,” in Proc. 27th ACM Int. Conf. Multimedia, Nice, France, Oct. 2019.

[4] Z. Cheng, J. Li, Q. Dai, X. Wu, and A. Hauptmann, “Learning spatial awareness to improve crowd counting,” in Proc. IEEE Int. Conf. Comput. Vis., Seoul, South Korea, Jun. 2019, pp. 6152–6161.

[5] V. A. Sindagi and V. M. Patel, “Multi-level bottom-up and top-bottom feature fusion for crowd counting,” in Proc. IEEE Int. Conf. Comput. Vis., Seoul, South Korea, Oct. 2019, pp. 1002–1012.

[6] Z. Zou, H. Shao, X. Qu, W. Wei, and P. Zhou, “Enhanced 3D convolutional networks for crowd counting,” in Proc. 30th Brit. Mach. Vis. Conf., Cardiff, U.K., 2019, pp. 1–11.

[7] M. A. Hessain, M. K. Kumar, M. Hosseinzadeh, O. Chanda, and Y. Wang, “One-shot scene-specific crowd counting,” in Proc. 30th Brit. Mach. Vis. Conf., Cardiff, U.K., 2019, pp. 1–11.

[8] Z. Shi, P. Mettes, and C. G. M. Snoek, “Counting with focus for free,” in Proc. IEEE Int. Conf. Comput. Vis., Seoul, South Korea, 2019, pp. 4200–4209.

[9] C. Xu, K. Qiu, J. Fu, S. Bai, Y. Xu, and X. Bai, “Learn to scale: Generating multipolar normalized density maps for crowd counting,” in Proc. IEEE Int. Conf. Comput. Vis., Seoul, South Korea, Oct. 2019, pp. 8382–8390.

[10] H. Xiong, H. Lu, C. Liu, L. Liu, Z. Cao, and C. Shen, “From open set to closed set: Counting objects by spatial divide-and-conquer,” in Proc. IEEE Int. Conf. Comput. Vis., Seoul, South Korea, Oct. 2019, pp. 8362–8371.

[11] L. Liu, Z. Qiu, G. Li, S. Liu, W. Ouyang, and L. Lin, “Crowd counting with deep structured scale integration network,” in Proc. IEEE Int. Conf. Comput. Vis., Seoul, South Korea, Oct. 2019, pp. 1774–1783.

[12] Z. Ma, X. Wei, X. Hong, and Y. Gong, “Bayesian loss for crowd count estimation with point supervision,” in Proc. IEEE Int. Conf. Comput. Vis., Seoul, South Korea, Oct. 2019, pp. 6142–6151.

[13] Z. Yan, Y. Yuan, W. Zuo, X. Tan, Y. Wang, S. Wen, and E. Ding, “Perspective-guided convolution networks for crowd counting,” in Proc. IEEE Int. Conf. Comput. Vis., Seoul, South Korea, Oct. 2019, pp. 952–961.

[14] C. Liu, X. Weng, and Y. Mu, “Recurrent attentive zooming for joint crowd counting and precise localization,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Long Beach, CA, USA, Jun. 2019, pp. 1217–1226.

[15] D. Lian, J. Li, J. Zheng, W. Luo, and S. Gao, “Density map regression guided detection network for RGB-D crowd counting and localization,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Long Beach, CA, USA, Jun. 2019, pp. 1821–1830.

[16] J. Wan, W. Luo, B. Wu, A. B. Chan, and W. Xu, “Residual regression with semantic prior for crowd counting,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Long Beach, CA, USA, Jun. 2019, pp. 4036–4045.

[17] Q. Zhang and A. B. Chan, “Wide-area crowd counting via ground-plane density maps and multi-view fusion CNNs,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Long Beach, CA, USA, Jun. 2019, pp. 8297–8306.

[18] M. Zhao, J. Zhang, C. Zhang, and W. Zhang, “Leveraging heterogeneous auxiliary tasks to assist crowd counting,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Long Beach, CA, USA, Jun. 2019, pp. 12736–12745.

[19] X. Jiang, Z. Xiao, B. Zhang, X. Zhen, X. Cao, D. Doermann, and L. Shao, “Crowd counting and density estimation by trellis encoder-decoder networks,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Long Beach, CA, USA, Jun. 2019, pp. 6133–6142.

[20] W. Liu, M. Salzmann, and P. Fua, “Context-aware crowd counting,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Long Beach, CA, USA, Jun. 2019, pp. 5099–5108.

[21] M. Shi, Z. Yang, C. Xu, and Q. Chen, “Revisiting perspective information for efficient crowd counting,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Long Beach, CA, USA, Jun. 2019, pp. 7279–7288.

[22] Y. Liu, M. Shi, Q. Zhao, and X. Wang, “Point in, box out: Beyond counting persons in crowds,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Long Beach, CA, USA, Jun. 2019, pp. 6469–6478.

[23] N. Liu, Y. Long, C. Zou, Q. Niu, L. Pan, and H. Wu, “ADCrowdNet: An attention-injective deformable convolutional network for crowd understanding,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Long Beach, CA, USA, Jun. 2019, pp. 3225–3234.
Q. Wang, J. Gao, W. Lin, and Y. Yuan, “Learning from synthetic data for crowd counting in the wild,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, Jun. 2019, pp. 8198–8207.

Y. Zhang, D. Zhou, S. Chen, S. Gao, and Y. Ma, “Single-image crowd counting via multi-column convolutional neural network,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 589–597.

D. B. Sam, S. Surya, and R. V. Babu, “Switching convolutional neural network for crowd counting,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 5744–5752.

H. Idrees, M. Tayyab, K. Athre, D. Zhang, S. Al-Maadeed, N. Rajpoot, and M. Shah, “Composition loss for counting, density map estimation and localization in dense crowds,” 2018, arXiv:1808.01050. [Online]. Available: http://arxiv.org/abs/1808.01050

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. Adv. Video Signal Based Survell. (AVSS)*, Lecce, Italy, Aug. 2017, pp. 1–6.

Y. Li, X. Zhang, and D. Chen, “CSRNet: Dilated convolutional neural networks for understanding the highly congested scenes,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 1091–1110.

L. Liu, H. Wang, G. Li, W. Ouyang, and L. Lin, “Crowd counting using deep recurrent spatial-aware network,” 2018, arXiv:1807.00601. [Online]. Available: http://arxiv.org/abs/1807.00601

Y. Zhang, F. Chang, M. Wang, F. Zhang, and C. Han, “Auxiliary learning for crowd counting via count-net,” *Neurocomputing*, vol. 273, pp. 190–198, Jan. 2018.

J. Gao, W. Lin, B. Zhao, D. Wang, C. Gao, and J. Wen, “[C] framework: An open-source PyTorch code for crowd counting,” 2019, arXiv:1907.02724. [Online]. Available: http://arxiv.org/abs/1907.02724

V. Mnih, N. Heess, A. Graves, and K. Kavukcuoglu, “Recurrent models of visual attention,” presented at the 28th Conf. Neural Information Process. Syst. (NeurIPS), Dec. 2014.

K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. Zemel, and Y. Bengio, “Show, attend and tell: Neural image caption generation with visual attention,” 2015, arXiv:1502.03044. [Online]. Available: http://arxiv.org/abs/1502.03044

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” 2017, arXiv:1706.03762. [Online]. Available: http://arxiv.org/abs/1706.03762

L. Chen, H. Zhang, J. Xiao, L. Nie, J. Shao, W. Liu, and T.-S. Chua, “SCA-CNN: Spatial and channel-wise attention in convolutional networks for image captioning,” 2018, arXiv:1705.03044. [Online]. Available: http://arxiv.org/abs/1705.03044

X. Wang, R. Girshick, A. Gupta, and K. He, “Non-local neural networks,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 7794–7803.

S. Woo, J. Park, J.-Y. Lee, and I. So Kweon, “CBAM: Convolutional block attention module,” in *Proc. IEEE Eur. Conf. Comput. Vis.*, Munich, Germany, vol. 11211, Sep. 2018, pp. 3–19.

Y. Hou, Z. Ma, C. Liu, and C. C. Loy, “Learning lightweight lane detection CNNs by self attention distillation,” in *Proc. IEEE Int. Conf. Comput. Vis.*, Seoul, South Korea, Oct. 2019, pp. 1013–1021.

J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, “Squeeze-and-excitation networks,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 7132–7141.

D. B. Sam, N. N. Sajjan, H. Maurya, and R. V. Babu, “Almost unsupervised learning for dense crowd counting,” in *Proc. 33rd AAAI Conf. Artif. Intell.*, Honolulu, HI, USA, Jan./Feb. 2019, pp. 8868–8875.

D. B. Sam and R. V. Babu, “Top-down feedback for crowd counting convolutional neural network,” in *Proc. 32nd AAAI Conf. Artif. Intell.*, New Orleans, LA, USA, 2018, pp. 7323–7330.

L. Zhang, M. Shi, and Q. Chen, “Crowd counting via scale-adaptive convolutional neural network,” in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2018, pp. 1113–1121.

Z. Shen, Y. Xu, B. Ni, M. Wang, J. Hu, and X. Yang, “Crowd counting via adversarial cross-scale consistency pursuit,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 5245–5254.
ZUODONG DUAN received the B.S. and master’s degrees from the College of Information Science and Engineering, Shandong Agricultural University, Taian, China, in 2014 and 2016, respectively. He is currently pursuing the Ph.D. degree with the Beijing Institute of Technology. His research is about deep learning applications in computer vision and robotics, with specific interests in crowd analysis, vehicle fine-gained classification, and SLAM.

YUJUN XIE received the bachelor’s degree from the College of Information Science and Technology, Beijing University of Chemical Technology University, Beijing, China, in 2018. She is currently pursuing the master’s degree with the Beijing Institute of Technology. Her research is about deep learning applications in computer vision and pattern recognition, with specific interests in target tracking and crowd analysis.

JIAHAO DENG was born in Laizhou, Shandong, China, in 1958. He received the B.S. and M.S. degrees in fuze technology and the Ph.D. degree in mechatronic engineering from the Beijing Institute of Technology, Beijing, in 1982, 1985, and 1998, respectively. Since 1985, he has been a Teacher with the School of Mechatronical Engineering, Beijing Institute of Technology.