Scale Adaptive Enhance Network for Crowd Counting

Zirui Fan
School of Information Engineering
Wuhan University of Technology
Wuhan, China
wfan07659@gamil.com

Jun Ruan
School of Information Engineering
Wuhan University of Technology
Wuhan, China
ruanjun@whut.edu.cn

***Abstract*—Crowd counting is a fundamental computer vision task and plays a critical role in video structure analysis and potential down-stream applications, e.g., accident forecasting and urban traffic analysis. The main challenges of crowd counting lie in the scale variation caused by disorderly distributed "person-camera" distances, as well as the interference of complex backgrounds. To address these issues, we propose a scale adaptive enhance network (SAENet) based on the encoder-decoder U-Net architecture. We employ Res2Net as the encoder backbone for extracting multi-scale head information to relieve the scale variation problem. The decoder consists of two branches, i.e., Attention Estimation Network (AENet) to provide attention maps and Density Estimation Network (DENet) to generate density maps. In order to fully leverage the complementary concepts between AENet and DENet, we craft to propose two modules to enhance feature transfer: i) a lightweight plug-and-play interactive attention module (IA-block) is deployed to multiple levels of the decoder to refine the feature map; ii) we propose a global scale adaptive fusion strategy (GSAFS) to adaptively model diverse scale cues to obtain the weighted density map. Extensive experiments show that the proposed method outperforms the existing competitive method and establishes the state-of-the-art results on ShanghaiTech Part A and B, and UCF-QNRF. Our model can achieve 53.56 and 5.95 MAE in ShanghaiTech Part A and B, with obtains performance improvement of 6.0 % and 13.13%, respectively.**

Keywords—crowd counting, scale adaptive enhance network, interactive attention module, global scale adaptive fusion strategy

I. Introduction

With the increasing complexity of social phenomena, the research on crowd counting has become a newly-raised and challenge in the field of computer vision. The main challenges contain scale variation, complex background interference, occlusion and perspective distortion, and so on. Most common solution for the scale variation is to learn multi-scale feature information from images by a multi-column neural network with different convolutional subnetworks or a single-column network with different convolutional kernels. Because the scale feature in the image is continuously variable, the existing approaches merely model the scale information roughly, which can not effectively relieve the problem of discretization of scale information and learn robust scale information. To eliminate this issue, we propose Scale Adaptive Enhance Network (SAENet) based on encoder-decoder U-Net [1] architecture. Rather than adopt VGG [2] as the backbone, we integrate the structure of Res2Net [3] and employ the block-wise processing mechanism to extract the feature map of each layer continuously, which contributes to model diverse scale information to learn different head sizes. We leverage the U-Net structure to integrate multi-scale representation from Res2Net to different levels of the decoder. In order to avoid complex background interference, the decoder consists of two branches: one branch defined as Density Estimation Network (DENet) can generate density maps. Another branch (Attention Estimation Network (AENet)) is designed to provide attention maps, which is used to provide auxiliary concepts for DENet. Compared to other networks, AENet and DENet simply utilize 1 X 1 convolution network during the decoding procedure, which makes our network more fixable and lightweight. The multi-scale, complementary and multi-task information provided by SAENet contributes to playing an essential role in crowd counting.

However, the na ïve fusion of density map and attention map is too rough to take good advantage of the dual-branch features. In this paper, we craft to design a novel fusion mechanism from two aspects. On one hand, we propose a lightweight plug-and-play interactive attention block (IA-block) to connect dual branches based on the mechanism of information passing and attention, which enhances the robustness of the network via frequent interaction between the attention map and density map. IA-block is easy to deployed on each layer of the decoder to implement the progressive enhancement to make good use of complementary scale information to generate a more reasonable feature map. On the other hand, we propose a global scale adaptive fusion strategy (GSAFS) to achieve the adaptive fusion between attention maps and density maps. In the procedure of feature map extraction, GSAFS helps to leverage several auxiliary networks to reconstruct the multiple density maps, which contain different scales, semantic and multi-task information. We unify these information-rich density maps and achieve the adaptive fusion via learning weight factors. IA-block and GSAFS promote each other in SAENet to improve the capability of the whole network during the end-to-end training process. The contributions can be summarized as follows:

1) We propose a scale adaptive enhance network (SAENet) based on the encoder-decoder U-Net architecture, which consists of two complementary branches, DENet and AENet.

AENet plays the role to guide DENet to generate high precision density maps via attention map.

2) We design a lightweight Interaction Attention module (IA-block) for connecting AENet and DENet to achieve information passing and sharing. We deploy IA-Block to each level of the decoder, refining the feature map of AENet and DENet progressively.

3) Global Scale Adaptive Fusion Strategy is introduced to employ auxiliary reconstruction network to predict density map from diverse decoding layers and achieve adaptive fusion of the attention map and density map.

4) Extensive experiments demonstrate that the proposed crowd-counting algorithm achieves state-of-the-art results on diverse datasets. Our model can achieve 53.56 and 5.95 MAE in ShanghaiTech Part A and B, with obtains performance improvement of 6.0 % and 13.13%, respectively.

II. Related Works

Previously, [4] employed the detection algorithms to count the crowd, but the problem of occlusion influences the performance of the detection algorithm. Recently, the regression methods [5, 6] and pixel-wise regression of density map [7] have become the main-stream method for crowd counting. With the development of deep learning, the deep neural network based models improve the performance of counting by increasing the depth and branches of the network and design the convolutional layer with different parameters. [8] introduces scale pyramid and [9] learns different scales feature maps through multi-column networks. [10, 11] adds a classification module to divide several patches of the image according to density levels. [12] introduces the global-local context learning and [13] implements adversarial learning to optimize the local-global consistency of the density map. [14] advocates a single-column network and uses dilated convolution to improve the receptive field of the network. [15] adopts [16] to extract multi-scale feature information. [17] continuously iterative dual-branch networks. [18] introduces [19] to generate receptive fields of different scales. This method only enhances the receptive field by the rough design of network architecture and did not refine the feature information.

The current popular methods usually learn multi-source information and improve the efficiency of feature fusion to enhance network performance. [20, 21, 22] integrate perspective information into the network smooth the discrete scale signal to ease the scale mix-match problem. [23, 24, 25] uses the segmentation, depth and counting heterogeneous properties to normalize the feature map. [26, 27] relieves background noise by attention map. [28] constructs a plurality of decoded branches and jumping connections to promote fusion. [29] propose the cross-scale residual function and allows the complementary feature to flow from adjacent layers. [30] uses a message-passing mechanism to aggregate the feature maps of different subnetworks and build CRF according to complementary information. [31] introduces CRF to learn more representation information. [32] achieve local attention and global attention with the self-attention mechanism. [33] models multi-scale information and localization with GNN [34] and extracts the relationship between nodes. [35] selectively enhances the network features with the multi-level attention mechanism. These methods are effective, but they also make the network structure redundant and difficult to improve the efficiency of feature information fusion. For these problems, we introduce an efficient method for crowd counting as below.

III. Approach

The proposed crowd-counting framework is shown in Fig. 1. It consists of three parts: SAENet, IA-block and GSAFS. SAENet is the basic architecture, which consists of AENet branch and DENet branch. We propose IA-block to enhance information transfer between two branches, GSASF is designed to achieve adaptive fusion of density map and attention map. We will illustrate these details in the following sections.

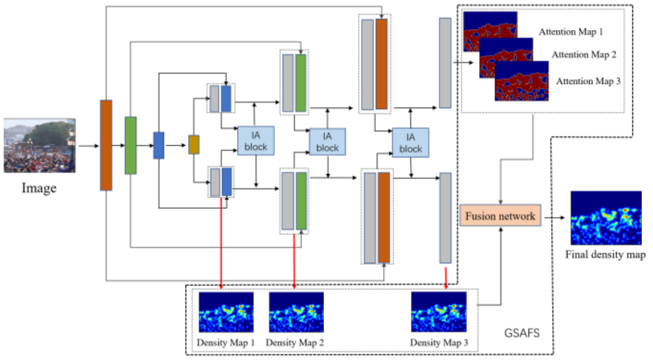

Fig. 1. The overall architecture of the proposed approach: SAENet, IA-block, GSAFS and black line represents the feature map generation path and red line denotes the density map generation path through Extraction network.

A. SAENet

The current popular crowd counting models usually employ deep CNN as the backbone and implement some post-processing operations based on the encoded features to solve the problems. Considering that the scale grouping mechanism in Res2Net [3] may be effective for modeling multi-scale concepts, we integrate Res2Net [3] as the backbone, which is "earlier one step" to address the scale variation issue. To further learn more robust information, we propose a U-Net based dual-path encoder-decoder architecture [36]. The basic model is named scale adaptive enhance network (SAENet), which contains DENet to generate density map, and AENet to provide attention map to assist DENet to obtain fused density map. In the design of the network, we integrate a variety of shortcut structures and 1×1 convolution layers, which not only improves the representation capacity but also avoids the problem of vanishing gradient.

B. Interaction Attention Block

However, AENet is easy to suffer from deviation and it is hard to correct the final density map via the naïve fusion. The techniques in ANF and DSSinet is hard to directly apply to SAENet due to the particularity and complexity of connection modules. So in this paper, we propose a structural model to connect DENet and AENet to achieve decoder layer-wise mutual recalibration. Specifically, we design a lightweight plug-and-play interactive attention block (IA-block) to connect the dual-branch (refer to Fig. 2.). Firstly, we feed the feature map of AENet and DENet to IA-block to generate confidence map:

$$M = F_a \odot F_d \tag{1}$$

where $\odot$ denotes element-wise multiplication. F_a denotes the feature map from AENet and F_d denotes the feature map from DENet. M represents confidence map.

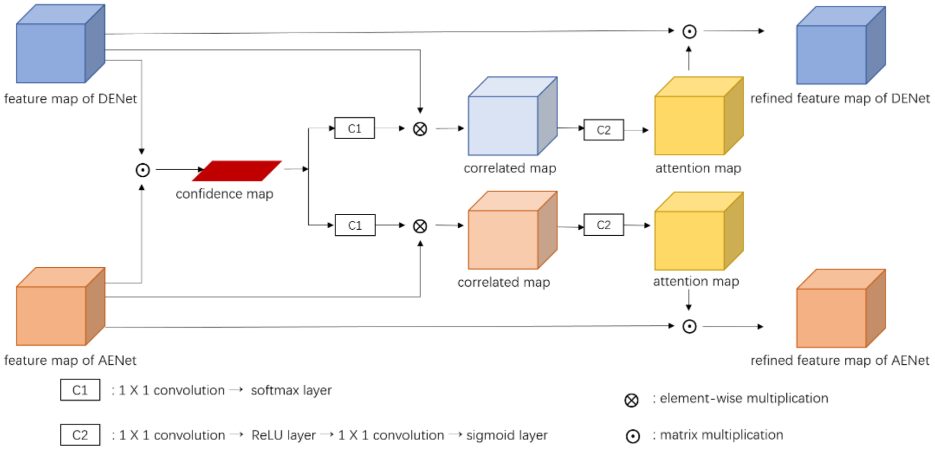
feature map of DENet
refined feature map of DENet
C1
C2
correlated map
attention map
confidence map
C1
C2
correlated map
attention map
feature map of AENet
refined feature map of AENet
C1 : 1 X 1 convolution → softmax layer
C2 : 1 X 1 convolution → ReLU layer → 1 X 1 convolution → sigmoid layer
⊗ : element-wise multiplication
⊙ : matrix multiplication

Fig. 2. The structure of proposed IA-block.

Because crowd counting model is based on pixel-wise regression, we assume that establishing the relationship between each pixel from the confidence map can obtain more robust representations. So we follow the non-local method to build long-range dependencies for each pixel:

$$F_{ra} = S(f^{1\times1}(M)) \otimes Fa \qquad F_{rd} = S(f^{1\times1}(M)) \otimes F_d, \tag{2}$$

where S denotes the softmax function, $f^{1\times1}$ is 1×1 convolution, $\otimes$ denotes matrix multiplication, F_{ra} and F_{rd} denote correlated map from AENet and DENet, respectively. Aggregating the reasoable attention map from this correlation map with abundant channels can effectly rectify the feature map of SAENet. Inspired by channel attention models such as CBAM and SE-layer, we feed the correlation map directly into the convolutional neural network without pooling layer:

$$A_{Ia} = \sigma(f^{1\times1}(ReLU(f^{1\times1}(F_{ra})))) \tag{3}$$

$$A_{Id} = \sigma(f^{1\times1}(ReLU(f^{1\times1}(F_{ra})))) \tag{4}$$

where A_{Ia} and A_{Id} denotes attention map to guide the generation of the feature map of AENet and DENet. We leverage these attention map to refine the feature map of each branchs:

$$Fga = A_{Ia} \odot Fa, \qquad Fgd = A_{Id} \odot Fd, \tag{5}$$

where Fga and F_{gd} denotes the refined feature map from IA-block of AENet and DENet. The enhanced feature map will be passed to the next layer of decoder. In this paper, we deploy the lightweight IA-block to each layer of the decoder, which strengthens "layer-by-step" feature information transmitted in neural networks. It is worth noting that this module is plug-and-play, and can be deployed to many current competitive networks to improve the performance of crowd counting.

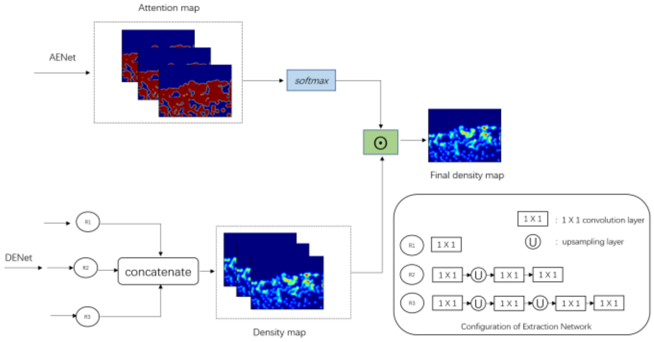

Fig. 3. The configurations and details of GSAFS.

C. Global Scale Adaptive Fusion Strategy

In this paper, we hope to fully explore these robust feature maps IA-block enhanced by rather than conducting a disposable fusion of density map and attention map. Considering the effectiveness of concatenation based on U-Net (See 4.4 section for details), we hypothesize that fusing density maps from different layers of the decoder that contain diverse scale cues would improve the model performance. Therefore, we design GSAFS to explore the adaptive fusion of density map and attention map. GSAFS consist of Extraction Network and Fusion Network, which is depicted in Fig. 3. We propose three auxiliary network, denoted as R1,R2,R3, which consists of several upsample layers and 1x1 convolution layers to extract the different layer of decoder feature map generated in DENet. AENet output triple-channel attention map to match these density maps:

$$AM_{fianl} = S(AM_1, AM_2, AM_3) \tag{6}$$

where AM_{fianl} denotes the final attention map of AENet, AM_1,AM_2,AM_3 denete the origi- nal output of AENet. In order to achieve adaptive scale fusion, we introduce a learnable parameter to learn the fusion ratio of attention map:

$$DM_{fianl} = S(DM_1,DM_2,DM_3), \quad DM = AM_{fianl} \odot DM_{final} \quad (7)$$

where DM_{fianl} denotes the final density map of DENet, DM_1,DM_2,DM_3 denote the original output of DENet, DM denotes the final result. During the optimization procedure of backpropagation, our GSAFS can realize scale adaptive fusion.

D. Loss Function Setting

The loss setting in the proposed framework contains three parts. $Loss_1$ is designed to use l_2 loss to calculate the loss of the final density map and ground truth labels:

$$Loss_1 = \frac{1}{N}\sum_{k=1}^{N}\left\|D(X_1^k;\theta) - D^k\right\|_2^2 \quad (8)$$

where N denetes the number of the images, D^k denotes the gound truth of density map, $\mathrm{D}(X_1^k;\theta)$ denotes the output of DENet. X is the feature map, subscript denotes the number of channel and θ denotes the parameters of network. $Loss_2$ is used to calculate the pixel level loss of density map sets and ground truth labels:

$$Loss_2 = \frac{1}{N}\sum_{k=1}^{N}\left\|D(X_3^k;\theta) - D_3^k\right\|_2^2 \quad (9)$$

$Loss_3$ is calculated based on binary ground truth A_3^k, which comes from the D_3^k based on a pre-defined threshold, and the predicted normalized attention map is computed as follows:

$$Loss_3 = \frac{1}{N}\sum_{k=1}^{N}(A_3^k \log(P_3^k) + (1 - A_3^k)\log(1 - P_3^k)) \quad (10)$$

where P_3^k denotes the output of AENet. To sum up, the final loss function for optimization is formulated as:

$$L = Loss_1 + Loss_2 + \lambda\, Loss_3 \quad (11)$$

In this work, we assign the λ equal to 0.1.

IV. Experiments

A. Datasets and Evaluatoin

We use a fixed Gaussian kernel to blur each pixel in the image to generate a density map [9]. The ground truth density map is generated by convolution operations of annotation map and the Gauss Kernel . The ground truth attention map is genereted on the basic of ground truth density map. We follow [36] to set the piexl value as 1 where piexl value is not zero. We adopt Mean Absolute Error (MAE) and Root Mean Square Error (RMSE)to measure the counting accuracy. ShanghaitechA [9] contains 300 training pictures and 182 test pictures, and the images are relatively high density and noisy. ShanghaitechB [9] has 400 training pictures and 316 test pictures, where the images have relatively low density but not so dense. UCF-QNRF [37] contains 1535 pictures and 1251642 personal heads, including 1201 pictures and 334 pictures of the test set.

B. Implementation Detail

We employ the Imagenet dataset to pre-train Res2Net101 as the encoder of the model. Our model is trained end-to-end via the loss function L. The optimizer is Adam [38], and the learning rate is set to 0.0001. The learning rate drops to one-tenth of its value for every 150 epochs. In order to train and avoid overfitting, we use random clipping and horizontal folding to augment the data.

C. Comparison with State-of-the-art Method

TABLE I. Performance Comparison of State-of-the-Art Methods on Public Datasets

Method	***UCF_QNRF***		***ShanghaiTech A***		***ShanghaiTech B***	
	MAE↓	MSE↓	MAE↓	MSE↓	MAE↓	MSE↓
MCNN [9]	-	-	110.2	173.2	26.4	41.3
CSRNet [14]	-	-	68.2	115.0	10.6	16.0
RANet [32]	111.0	190.0	59.4	102.0	7.9	12.9
TEDNet [28]	113.0	188.0	64.2	109.1	8.2	12.8
RAZ-Net [25]	116.0	195.0	65.1	106.7	8.4	14.1
DSSINet [30]	99.1	159.2	60.6	94.0	6.9	10.3
PGCNet [24]	-	-	57.0	**86.0**	8.8	13.7
MBTTBF [29]	97.5	165.2	60.2	94.1	8.0	15.5
HA-CCN [35]	-	-	62.9	94.9	8.1	13.4
PADNet [39]	96.5	170.2	59.2	98.1	8.1	12.2
HyGNN [33]	100.8	185.3	60.2	94.5	7.5	12.7
ASNet [27]	91.59	159.7	57.78	90.13	-	-
Ours	**85.05**	**146.5**	**53.56**	89.22	**5.95**	**9.79**

Model comparison: Table I reports the result of our model with the other 12 existing SOTA models on three datasets. It reveals that our model performs better than other methods on three datasets for comparison, which demonstrates the effectiveness of the proposed framework in crowd counting. In UCF-QNRF, the results of our model are better than all other models with lower error, which shows that our model is well suited for dense crowd scenarios. For example, our model obtains a 7.6% decrease in MAE and an 8.3% decrease in MSE compared to ASNet in UCF-QNRF. Furthermore, our framework achieves a 6.0% decline in MAE compared to the best model (PGCNet) in Shanghaitech A dataset, which shows that introducing additional complementary concepts can enhance model performance. In Shanghaitech B dataset, our model decline MAE 13.13% compared to the best model, DSSINet. It effectively shows that our model is also adapted to more sparse crowd scenarios. These above observations demonstrate that the effectiveness of the proposed method.

Qualitative Analysis: To better explore the performance of SAENet, we select several samples for qualitative analysis. As shown in Fig. 4., the dense level of the crowd increases from top to down, and the comparison of the estimation and ground truth prove that our model helps to accurately count different intensive levels of the crowd. We also obtain the following observations. i) We find that each point of our density map can correspond to objects in the sparse area in the image; ii) When the crowd is too dense to lead occlusion and scale variation, our model can reasonably estimate the dense crowd; iii) When the crowd scene is too complicated, our model can distinguish between counting area and background area; iv) The performance of SAENet is more advanced when compared with baseline, which also illustrates that the refinement of feature map and adaptive fusion is essential for crowd counting.

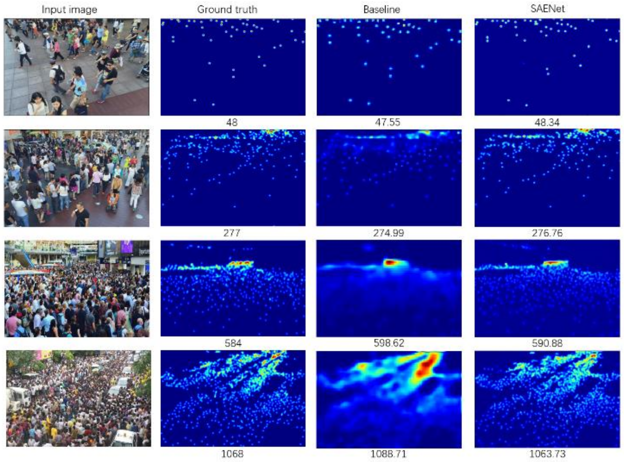

Fig. 4. Visually qualitative results on the Shanghaitech dataset.

D. Ablation Study

The effectiveness of U-Net and Res2Net: One of the main contributions of this work is to explore the influence of the network's basic backbone on crowd counting. In the experiment, we use the VGG16 network as the encoder, and the results decline 11.3 % and 24.5 % in two datasets, respectively. It reveals that Res2Net can better model feature information in crowd counting. Furthermore, we remove the U-Net structure in our model, and the experimental results show that it drops by 15.0 % and 23.0 % in two datasets. It indicates that the shortcut operation of the feature map can generate more robust features. The above experimental results also reveal that the difference of backbone is more obvious on the influence of the data results of the sparse crowd, and the U-Net structure is more obvious for the influence of the dense crowd.

TABLE II. TABLATION STUDY OF SAENET ON SHANGHAITECH PART A AND B DATASET

Model	ShanghaiTech A		ShanghaiTech B	
	MAE↓	MSE↓	MAE↓	MSE↓
Our VGG	59.61	102.60	7.41	12.76
Our w/o unet	61.60	100.72	7.32	12.23
Our w/o IA-block	56.68	93.21	6.42	10.34
Our w/o GSAFS	54.25	90.98	6.31	9.88
Our	53.56	89.22	5.95	9.79

The effectiveness of IA-block: In this work, we explore the significance of the IA-block on crowd counting. As shown in the table, when we don't deploy the IA-block model, MAE and MSE will decline 5.8 % and 7.9 %. This result not only shows that IA-block is effective but also indicates that complementary information sharing and passing is essential in this task.

The effectiveness of GSAFS: We also conduct experiments to verify the significance of GSAFS. The experimental results reveal that the errors will increase by 1.8 % and 6.0 % without GSAFS. It verifies the effectiveness of the global scale adaptive fusion strategy. Compared to the dense crowd, the GSAFS strategy is more obvious to the sparse crowd counting.

V. CONCLUSION

In this paper, we propose SAENet to perform the task of crowd counting. The encoder of SAENet integrates the Res2Net architecture to generate multi-channel different scales feature maps. We employ dual branches to extract complementary concepts. Furthermore, IA-block and GSAFS are introduced to achieve information transfer and adaptive fusion. We demonstrate the performance of SAENet and verify the core components through extensive experiments. We believe our method opens up new ideas for the field of crowd counting.

REFERENCES

[1] Olaf Ronneberger, Philipp Fischer and Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *MICCAI*, 2015.

[2] Karen Simonyan and Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. In arXiv preprint arXiv:1409.1556, 2014.

[3] Shang-Hua Gao, Ming-Ming Cheng, Kai Zhao, Xin-Yu Zhang, Ming-Hsuan Yang and Philip Torr. Res2Net: A New Multi-scale Backbone Architecture. In *IEEE Transac- tions on Pattern Analysis and Machine Intelligence*, 2021.

[4] B. Leibe, E. Seemann and B. Schiele. Pedestrian detection in crowded scenes. In CVPR, 2005.

[5] Antoni B. Chan, Zhang-Sheng John Liang and Nuno Vasconcelos. Privacy preserving crowd monitoring: Counting people without people models or tracking. In *CVPR*, 2008.

[6] Haroon Idrees, Imran Saleemi, Cody Seibert, and Mubarak Shah. Multi-source multi- scale counting in extremely dense crowd images. In *CVPR*, 2013.

[7] Victor Lempitsky and Andrew Zisserman. Learning to count objects in images. In *NeurIPS*, 2010.

[8] Daniel Oñoro-Rubio and Roberto J. López-Sastre. Towards Perspective-Free Object Counting with Deep Learning. In *ECCV*, 2016.

[9] Yingying Zhang, Desen Zhou, Siqin Chen, Shenghua Gao, and Yi Ma. Single-image crowd counting via multi-column convolutional neural network. In *CVPR*, 2016.

[10] Deepak Babu Sam, Neeraj N Sajjan and R. Venkatesh Babu. Divide and Grow: Cap- turing Huge Diversity in Crowd Images with Incrementally Growing CNN. In *CVPR*, 2018.

[11] DeepakBabuSam,ShivSurya,andRVenkateshBabu.Switchingconvolution alneural network for crowd counting. In *CVPR*, 2017.

[12] Vishwanath A. Sindagi and Vishal M. Patel. Generating High-Quality Crowd Density Maps Using Contextual Pyramid CNNs. In *CVPR*, 2017

[13] Jiang Liu, Chenqiang Gao, Deyu Meng and Alexander G. Hauptmann. DecideNet: Counting Varying Density Crowds Through Attention Guided Detection and Density Estimation. In *CVPR*, 2018.

[14] Yuhong Li, Xiaofan Zhang, and Deming Chen. Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes. In *CVPR*,2018.

[15] Xinkun Cao, Zhipeng Wang, Yanyun Zhao, and Fei Su. Scale aggregation network for accurate and efficient crowd counting. In *ECCV*, 2018.

[16] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke and Andrew Rabinovich. Going Deeper With Convolutions. In *CVPR*, 2015.

[17] Viresh Ranjan, Hieu Le and Minh Hoai. Iterative Crowd Counting. In *ECCV*, 2018.

[18] Weizhe Liu, Mathieu Salzmann and Pascal Fua. Context-Aware Crowd Counting. In CVPR, 2019.

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition. In ECCV, 2014.

[20] Miaojing Shi, Zhaohui Yang, Chao Xu and Qijun Chen. Revisiting Perspective Infor- mation for Efficient Crowd Counting. In *CVPR*, 2019.

[21] Zhaoyi Yan, Yuchen Yuan, Wangmeng Zuo, Xiao Tan, Yezhen Wang, Shilei Wen and Errui Ding. Perspective-Guided Convolution Networks for Crowd Counting. In *ICCV*, 2019.

[22] Yifan Yang, Guorong Li, Zhe Wu, Li Su, Qingming Huang, Nicu Sebe. Reverse Per- spective Network for Perspective-Aware Object Counting. In *CVPR*, 2020.

[23] Muming Zhao, Jian Zhang, Chongyang Zhang and Wenjun Zhang. Leveraging Hetero- geneous Auxiliary Tasks to Assist Crowd Counting. In *CVPR*, 2019.

[24] Zenglin Shi, Pascal Mettes and Cees G. M. Snoek. Counting with Focus for Free. In *ICCV*, 2019.

[25] Chenchen Liu, Xinyu Weng, and Yadong Mu. Recurrent attentive zooming for joint crowd counting and precise localization. In *CVPR*, 2019.

[26] Ning Liu, Yongchao Long, Changqing Zou, Qun Niu, Li Pan and Hefeng Wu. AD- CrowdNet: An Attention-injective Deformable Convolutional Network for Crowd Un- derstanding. In *CVPR*, 2019.

[27] Xiaoheng Jiang, Li Zhang, Mingliang Xu, Tianzhu Zhang, Pei Lv, Bing Zhou, Xin Yang, Yanwei Pang. Attention Scaling for Crowd Counting. In *CVPR*, 2020.

[28] Xiaolong Jiang, Zehao Xiao, Baochang Zhang, Xiantong Zhen, Xianbin Cao, David Doermann, Ling Shao. Crowd Counting and Density Estimation by Trellis Encoder- Decoder Networks. In *CVPR*, 2019.

[29] Vishwanath A. Sindagi and Vishal M. Patel. Multi-level bottom-top and top-bottom feature fusion for crowd counting. In *ICCV*, 2019.

[30] Lingbo Liu, Zhilin Qiu, Guanbin Li, Shufan Liu, Wanli Ouyang and Liang Lin. Crowd Counting with Deep Structured Scale Integration Network. In *ICCV*, 2019.

[31] Anran Zhang, Lei Yue, Jiayi Shen, Fan Zhu, Xiantong Zhen, Xianbin Cao, Ling Shao. Attentional Neural Fields for Crowd Counting. In *ICCV*, 2019.

[32] Anran Zhang, Jiayi Shen, Zehao Xiao, Fan Zhu, Xiantong Zhen, Xianbin Cao, Ling Shao. Relational Attention Network for Crowd Counting. In *ICCV*, 2019.

[33] Ao Luo, Fan Yang, Xin Li, Dong Nie, Zhicheng Jiao, Shangchen Zhou and Hong Cheng. Hybrid Graph Neural Networks for Crowd Counting. In *AAAI*, 2020.

[34] Keyulu Xu, Weihua Hu, Jure Leskovec and Stefanie Jegelka. HOW POWERFUL ARE GRAPH NEURAL NETWORKS?. In *ICLR*, 2019.

[35] Vishwanath A. Sindagi and Vishal M. Patel. PaDNet: HA-CCN: Hierarchical Attention-Based Crowd Counting Network. In *IEEE Transactions on Image Process- ing*, 2019.

[36] Shengqin Jiang, Xiaobo Lu, Yinjie Lei and Lingqiao Liu. Mask-Aware Networks for Crowd Counting. In *IEEE Transactions on Circuits and Systems for Video Technology*, 2020.

[37] HaroonIdrees,MuhmmadTayyab,KishanAthrey,DongZhang,SomayaAl-Maadeed, Nasir Rajpoot, and Mubarak Shah. Composition loss for counting, density map esti- mation and localization in dense crowds. In *ECCV*, 2018.

[38] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. In *arXiv preprint arXiv:1412.6980*, 2014.

[39] Yukun Tian, Yiming Lei, Junping Zhang and James Z. Wang. PaDNet: Pan-Density Crowd Counting. In *IEEE Transactions on Image Processing*, 2019.