



COMAL: compositional multi-scale feature enhanced learning for crowd counting

Fangbo Zhou¹ · Huailin Zhao¹ · Yani Zhang² · Qing Zhang² · Lanjun Liang¹ · Yaoyao Li¹ · Zuodong Duan³

Received: 23 January 2021 / Revised: 11 June 2021 / Accepted: 14 January 2022 /
Published online: 11 March 2022

© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2022

Abstract

Accurately modeling the crowd's head scale variations is an effective way to improve the counting accuracy of the crowd counting methods. Most counting networks apply a multi-branch network structure to obtain different scales of head features. Although they have achieved promising results, they do not perform very well on the extreme scale variation scene due to the limited scale representability. Meanwhile, these methods are prone to recognize background objects as foreground crowds in complex scenes due to the limited context and high-level semantic information. We propose a compositional multi-scale feature enhanced learning approach (COMAL) for crowd counting to handle the above limitations. COMAL enhances the multi-scale feature representations from three aspects: (1) The semantic enhanced module (SEM) is developed for embedding the high-level semantic information to the multi-scale features; (2) The diversity enhanced module (DEM) is proposed to enrich the variety of crowd features' different scales; (3) The context enhanced module (CEM) is designed for strengthening the multi-scale features with more context information. Based on the proposed COMAL, we develop a crowd counting network under the encoder-decoder framework and perform extensive experiments on ShanghaiTech, UCF_CC_50, and UCF-QNRF datasets. Qualitative and quantitative results demonstrate the effectiveness of the proposed COMAL.

Keywords Crowd counting · Crowd density estimation · Convolutional neural network · Multi-scale feature learning

✉ Huailin Zhao
zhao_huailin@yahoo.com

¹ School of Electrical and Electronic Engineering, Shanghai Institute of Technology, Shanghai, China

² School of Computer Science and Information Engineering, Shanghai Institute of Technology, Shanghai, China

³ Science and Technology on Electromechanical Dynamic Control Laboratory, School of Mechatronical Engineering, Beijing Institute of Technology, Beijing, China

1 Introduction

Crowd counting is to estimate the number and density distribution of people in an image or video frame. It is particularly prominent because of its special significance for public safety and management [6, 52, 53], especially during the COVID-19 pandemic, accurate crowd counting helps avoid gatherings of people. It has also attracted widespread attention of many scholars [5, 12, 23, 36, 38, 49, 58]. However, crowd counting is a very challenging due to the large scale variation of crowd head and complex backgrounds.

In recent years, with the renaissance of deep learning [28, 50, 51], convolutional neural network (CNN) based methods [8, 15], have achieved significant progress in crowd counting task [13, 18, 27, 55, 63]. They formulate the task as a regression problem [10, 22, 32, 39, 64], which designs the sophisticated network to establish the nonlinear relationship between the input crowd image and its corresponding crowd density map. Among them, efficiently modeling the scale variations of a crowd is a classical and hot research topic, and many researchers spare no effort to propose methods to handle it. For example, multi-column networks [37, 44, 66] are designed to model different scales of crowd head. However, they usually have complicated structures and need a long time to be optimized, requiring large computation resources to be implemented and not appropriate to real-world applications. The encoder-decoder frameworks are recently very popular in crowd counting tasks, and the sophisticated decoder is proposed. For instances, Zhao et al. [68] design the decoder with different auxiliary task branches for obtaining robust representations from the auxiliary tasks. Xie et al. [60] extract multi-scale features via the proposed decoder with the stacked dilated convolutional layers and the recurrent modules. Although they achieve promising performances, they may fail in the scene with the extreme crowd scale variations and complicated background stuff due to the limited scale representations and representational ability. Thus, modeling a crowd's scale variations in different scenes is still a challenging and unsolved problem for crowd counting.

To solve the challenges mentioned above in the crowd counting task, we aim to extract efficient multi-scale feature representations for crowd counting from three aspects: (1) extract high-level semantic features of crowd for enhancing the crowd-aware representations; (2) model continuous scale variations of crowd for multi-scale crowd counting; and (3) extract long-range dependency of pixel for obtaining context information. To this end, we propose a compositional multi-scale feature enhanced learning approach (COMAL). Specifically, for Semantic feature enhancement, the semantic enhanced module (SEM) is designed, which embeds the semantic information from high-level features to the multi-scale crowd features. The diversity enhanced module (DEM) enriches the varieties of feature representations via three diversity enhanced blocks in a cascade manner for scale diversity enhancement. For context enhancement, the context enhanced module is proposed to extract context information from spatial and channel dimensions via the neural attention mechanism. With the help of COMAL, the multi-scale features can own strong representational ability and abundant feature representations, which can handle the scale variation challenges from different crowd scenes. Based on the proposed COMAL, we design a counting network under the encoder-decoder framework. The COMAL is used as the decoder for final crowd density estimation. Extensive experiments are performed on commonly-used crowd counting benchmarks, and our network outperforms the other state-of-the-art methods. The visualization results further prove the effectiveness of the proposed COMAL.

To summarize, the main contributions of our paper are fourfold:

- We propose a semantic enhanced module (SEM) to embed the high-level semantic information into the multi-scale features, which can improve the crowd recognition performance on complex crowd scene.
- We develop a diversity enhanced module (DEM) to enrich the scale representations. It helps the counting network to handle the extreme scale variations case better.
- We design a context enhanced module (CEM) to strengthen the extracted multi-scale features with more context information. CEM can help the counting network recognize the foreground crowd and background stuff for the complex crowd scene.
- We combine the above three modules into a compositional learning approach, COMAL, and build an encoder-decoder network based on it for crowd counting. With the assistance of COMAL, the counting network outperforms the other state-of-the-art methods on commonly-used crowd counting benchmarks.

The rest of this paper is organized as follows. Section 2 demonstrates the related works of CNN-based crowd counting and multi-scale feature learning methods. In Section 3, we introduce the COMAL and its components in detail. We introduce the experiment details and model analysis in Section 4 and conclude our method in Section 5.

2 Related works

In this section, we review the CNN-based crowd counting methods and multi-scale feature representation learning methods.

2.1 CNN-based crowd counting

We first review the crowd counting method [7, 29, 34, 45, 46, 57, 62] and summarize them in Table 1. For example, Zhang et al. [66] proposed a Multi-column Convolutional Neural Network (MCNN) with different convolutional structures to solve the scale variations of crowd heads. Sam et al. [44] designed Switch-CNN, which trained a switch classifier to select the optimal CNN regressor for the specific scale density estimation. However, The limitation of Switch-CNN is that it chooses one of the results of different sub-networks rather than fusing them. Deb et al. [11] proposed an aggregated multi-column dilated convolution network for perspective-free counting. Although the above multi-column networks have achieved significant progress, they only consider limited crowd scale and doesn't perform well on continuous scale variation scene. To reduce the computational resources, Li et al. [30] proposed the CSRNet, which adopted the dilated convolutional layers to enlarge the receptive field of the network. However, the six successive dilated convolutional layers of CSRNet will cause a serious gridding effect [54], which can not efficiently extract crowd features. To solve this problem, our SEM adopted multiple parallel filters with different dilate rates for exploiting multi-scale features. Cao et al. [3] proposed a scale aggregation network (SANet), which applied the scale aggregation module to extract multi-scale features and the transposed convolutional layer to regress the final crowd density map. Besides, some neural attention based methods have also been applied to the crowd counting task [16, 19]. Guo et al. [19] explored a scale-aware attention fusion method with different dilated rates to obtain different visual granularities of the crowd's region of interest.

Table 1 Summarizations of crowd counting methods

Methods	Backbone	Multi-branch	Dilation convolution	Attention
MCNN [66]	-	✓		
CrowdNet [2]	VGG-16	✓		
Switch-CNN [44]	-	✓		
AMDCN [11]	-	✓	✓	
ic-CNN [42]	-	✓		
IG-CNN [43]	VGG-16	✓		
CSRNet [30]	VGG-16		✓	
SANet [3]	-	✓		
DADNet [19]	-	✓	✓	
SCAR [16]	VGG-16		✓	✓
TEDnet [26]	-	✓	✓	
SACANet [1]	VGG-16	✓	✓	✓
DACC [14]	VGG-16			
ASD [61]	VGG-16	✓		
CAN [33]	VGG-16		✓	
PCC Net [17]	VGG-16	✓		
COMAL(Ours)	VGG-16	✓	✓	✓

Gao et al. [16] proposed a space-/channel-wise attention regression network to exploit the context information of crowd scene for accurate crowd counting. The well-designed attention models effectively encode the large-range contextual information. We propose a compositional learning approach to enhance the multi-scale feature, which guides the counting network to learn robust representations for different crowd scenes.

2.2 Multi-scale feature representation learning

Scale variation is a common problem in different computer vision tasks [4, 9, 20, 31, 67]. Many multi-scale feature representation learning methods are proposed to solve it. Lin et al. [31] proposed a feature pyramid network (FPN), which fused high-level features and low-level features by element-wise summation for small object detection. Zhao et al. [67] proposed a pyramid scene parsing network (PSPNet) for aggregating context information at different scales. Inspired by the spatial pyramid pooling (SPP) [21], Chen et al. [9] proposed the Atrous Spatial Pyramid Pooling (ASPP) module to use four convolutions with different dilated rates. ASPP can effectively enlarge the network's receptive field and obtain multi-scale information, which prompt the network to achieve a new superior result on semantic segmentation task. He et al. [20] proposed the Adaptive Pyramid Network (APCNet), which used Adaptive Context Modules to leverage local and global representation to estimate an affinity weight for local regions. To obtain larger-scale information, Cao et al. [4] proposed a global context network (GCNet), which focuses on the connection between different image positions by establishing a long-range relationship between pixels. In this paper, we propose the DEM to enrich the multi-scale feature representations, and apply the proposed SEM and CEM to strengthen the feature representations.

3 Proposed method

In this section, we firstly introduce the overview of the counting network with the proposed COMAL. Then, SEM, DEM, and CEM are elaborated. Finally, we demonstrate the loss function and evaluation metrics we use.

3.1 Overview

The overview of the counting network we used in this paper is shown in Fig. 1. Following [2, 16, 30], we choose VGG-16 [48] as the feature encoder. However, in order to obtain semantic features, we use the first thirteen layers instead of the first ten convolutional layers. Then, the encoder features are fed to SEM, DEM and CEM sequentially to get the enhanced multi-scale crowd features. Finally, the extracted multi-scale features are processed by a single 1×1 convolutional layer and the bilinear interpolation operation to regress the final crowd density map. Each component of the counting network is demonstrated as follows.

3.2 Semantic enhanced module

We propose the SEM to generate the multi-scale crowd features with abundant semantic information for final crowd density estimation. The detailed structure of SEM is shown in Fig. 2. It has two paths: the low-level feature process path (LFP) and the high-level feature process path (HFP). The LFP is designed to extract multi-scale features, and the HFP aims to enhance the extracted feature with more high-level semantic information. Specifically, we use four convolutional layers with different dilated rates in a parallel way to extract various scale features. After that, the different scale features are combined with the concatenate operation. A 1×1 convolutional layer is applied to reduce the feature dimensions. For HFP, the high-level feature from VGG-16 is fed to 1×1 convolutional layer to reduce the feature

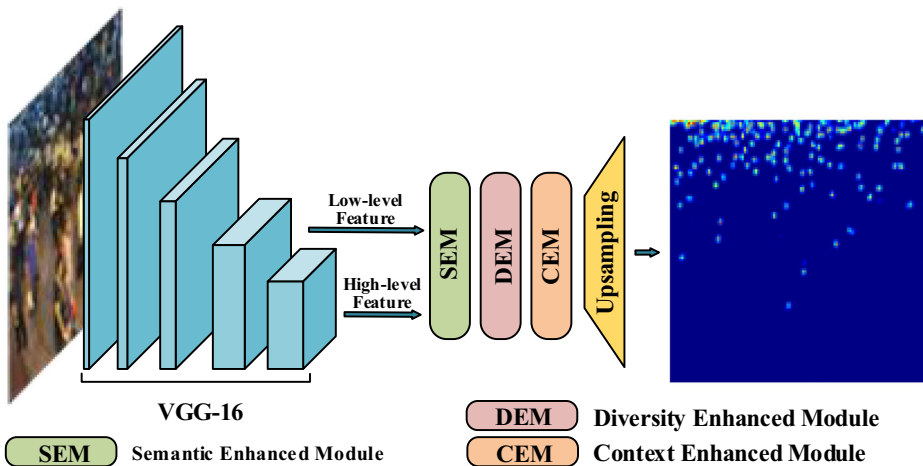


Fig. 1 Overview of the proposed counting network. Each input image is fed to the first 13 layers of VGG-16 to extract the crowd features. Then, the output of the first 10 layers of VGG-16 (Low-level feature) and the first 13 layers of VGG-16 (high-level feature) are sent to SEM, DEM and CEM to generate the enhanced multi-scale crowd features. Finally, the extracted multi-scale context features are processed by 1×1 convolutional layer and bilinear interpolation operation for final crowd density estimation

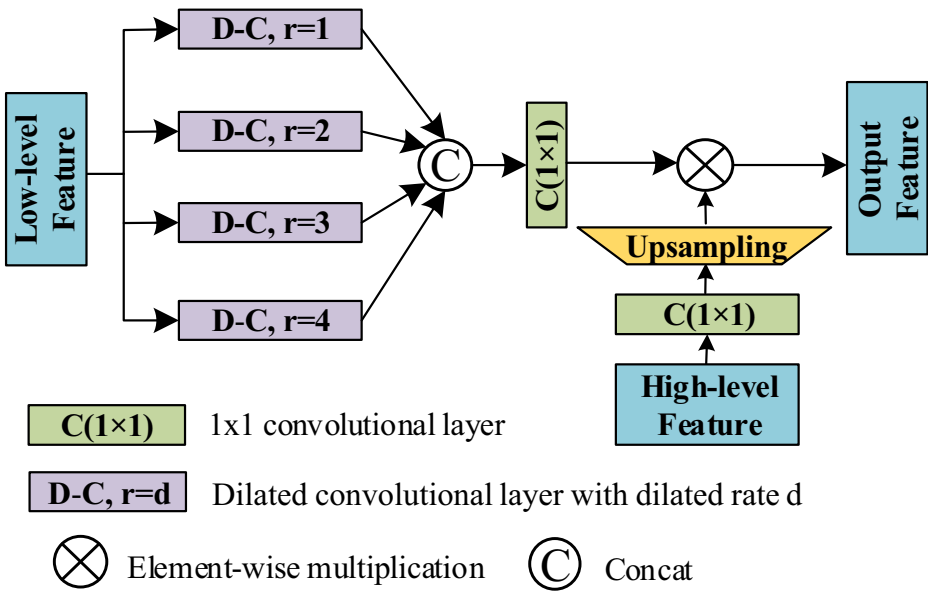


Fig. 2 Illustration of SEM. The low-level features are fed to different dilated convolutional layers to generate the initial multi-scale crowd features. The high-level features from the encoder are multiplied with the output of the dilated convolutional layers to modified the extracted multi-scale crowd features with more semantic information

dimension and processed by the bilinear interpolation operation to the same size as the low-level feature dimensions. Different from the previous approach [31], which directly uses the element-wise summation operation to fuse the upsampled high-level features and low-level features, we follow the design of Exfuse [65]. The output of HFP is multiplied with the output of LFP with element-wise multiplication operation to generate the initial multi-scale features, which prompt the network with more feature discriminability. More analysis can be seen in Section 4.3.

3.3 Diversity enhanced module

Although SEM generates multi-scale crowd features, the representation of crowd features are limited, which will hinder the performance of counting network in complex scene. To increase the diversity of crowd features, we design the DEM, which consists of three diversity enhanced blocks (DEB). The design philosophy of DEM comes from [56]. As shown in Fig. 3, each DEB has two branches. One branch with a single 3×3 convolutional layer and another branch with two stack 3×3 convolutional layers. All 3×3 convolutional layers have the half channel number of the input features and the output of two branches are fused with the element-wise summation. We place three DEBs in a cascade manner after SEM, as shown in Fig. 3 (b), and it is equivalent to eight branches with different receptive fields in parallel, as shown in Fig. 3 (c). Thus, it can generate the abundant crowd features for modeling continuous scale variations. The performance of different numbers of DEB can be seen in Section 4.3.3.

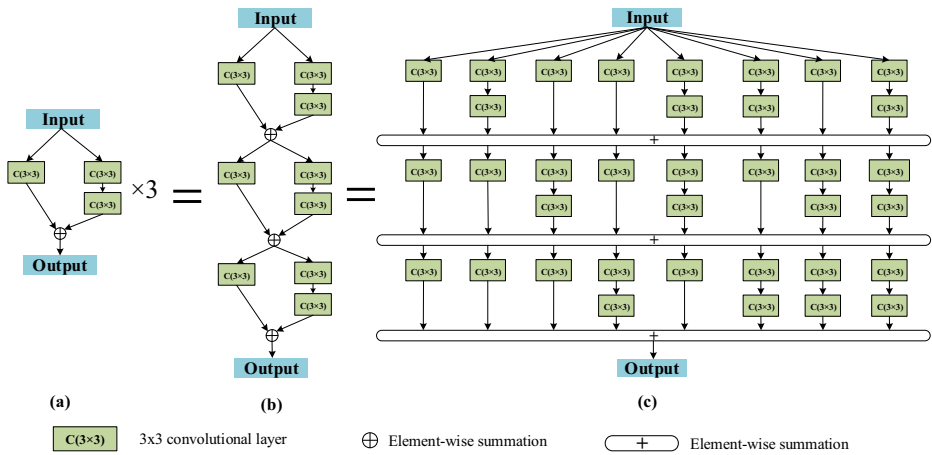


Fig. 3 Illustration of DEM. From left to right: (a) the structure of DEB, (b) the structure of three DEBs in a cascade manner (DEM), (c) the equivalent structure of (b)

3.4 Context enhanced module

To increase the discriminability of the proposed COMAL, we propose the CEM to exploit the context information from multi-scale crowd features. The detailed architecture of CEM is shown in Fig. 4. CEM includes two branches: position attention module (PAM) and channel attention module (CAM). The details of PAM and CAM are as follows.

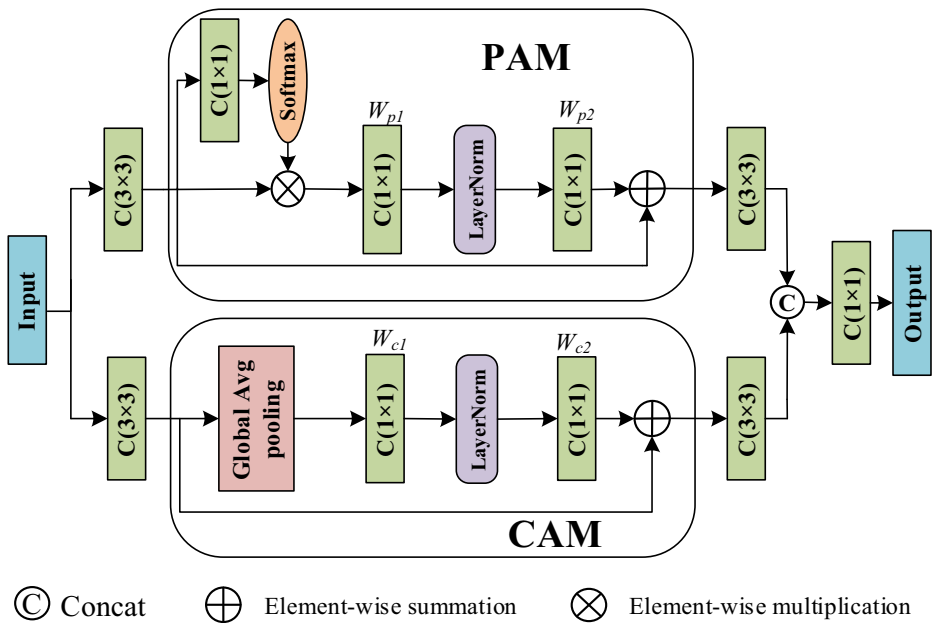


Fig. 4 Illustration of CEM. The PAM of CEM is designed for exploiting the context information from spatial dimension and the CAM of CEM is developed for acquiring the context information from channel dimension

3.4.1 Position attention module

The PAM encodes the context information by calculating the long-range pixel relationship. Its detailed structure is shown in Fig. 4. The input features are firstly processed by a 3×3 convolutional layer. After that, the processed features are fed into a 1×1 convolutional layer and the Softmax layer to get the position attention weight P_i^{att} , which can be formulated as follows:

$$P_i^{att} = \frac{\exp(P_i \cdot P_j)}{\sum_{j=1}^N \exp(P_i \cdot P_j)} P_i, \quad (1)$$

where $\{P_i | i \in \{1 \cdots N\}\}$ denotes the i -th position of input feature map, N is the number of positions in the feature map, which is equal to $H \times W$.

The position attention weight P_i^{att} is fed to the bottleneck structure which is constructed by two 1×1 convolutional layers. Specifically, we place the LayerNormalization (LN) at the middle of two 1×1 convolutional layers for better weight optimization. The output of bottleneck is fused with the input of PAM via the residual learning and the final position attention feature can be formulated as follows:

$$P_i^{final} = P + W_{p2} \text{ReLU} \left(\text{LN}(W_{p1}) \sum_{i=1}^N P_i^{att} P \right), \quad (2)$$

where P denotes the input feature of PAM. $\text{ReLU}(\cdot)$ and $\text{LN}(\cdot)$ denote the ReLU and LN layer, respectively. W_{p1} and W_{p2} represent the weight of two 1×1 convolutional layers, respectively.

3.4.2 Channel attention module

The structure of CAM is similar with PAM, which is shown in Fig. 4. Different from PAM, we apply the global average pooling layer to acquire the global context information and the final channel attention weight C_i^{final} can be defined as follows:

$$C_i^{final} = C + W_{c2} \text{ReLU}(\text{LN}(W_{c1} C_m)), \quad (3)$$

where C denotes the input feature of CAM. C_m represents the global average pooling feature. W_{c1} and W_{c2} denote the weight of two 1×1 convolutional layers, respectively.

3.5 Ground-truth density map generation

Following [66], we use the Gaussian kernel to convolve the head annotation points and generate a crowd density map $F(x)$, which is defined as follows:

$$F(x) = \sum_{i=1}^N \delta(x - x_i) * G_\sigma(x) \quad (4)$$

where $G_\sigma(x)$ stands for the Gaussian kernel, x_i is the ground truth head location, x is a pixel position in the input image. We convolve $\delta(x - x_i)$ with a Gaussian kernel with parameter σ . For different datasets, σ is set as different values. For ShanghaiTech Part_B, UCF_CC_50, and UCF_QNRF, σ is set to 15. For ShanghaiTech Part_A, σ is equal to $\beta \bar{d}^i$, where \bar{d}^i represents the average distance of k nearest neighbors and β is set to 0.3.

3.6 Loss function and evaluation metrics

We use the L_2 loss to optimize the proposed COMAL. The loss function is defined as follows:

$$L(\Theta) = \frac{1}{2N} \sum_{i=1}^N \|F(X_i, \Theta) - F_i\|_2^2, \quad (5)$$

where N is the total number of training images. $F(X_i, \Theta)$ is the estimated density map generated by COMAL with parameters Θ . X_i represents the input image while F_i is the ground truth of the input image X_i .

The mean absolute error (MAE) and the mean square error (MSE) are chosen to evaluate the effectiveness of our method. The formulations are as follows:

$$MAE = \frac{1}{N} \sum_{i=1}^N |C_i^{pred} - C_i^{gt}|, \quad (6)$$

$$MSE = \sqrt{\frac{1}{N} \sum_{i=1}^N |C_i^{pred} - C_i^{gt}|^2}, \quad (7)$$

where N stands for the total number of the test images. C_i^{gt} and C_i^{pred} denote the ground truth number and the prediction number in the i -th image, respectively.

4 Experiments

In this section, we first describe the implementation details and experiment setup. Then, we introduce the commonly-used crowd counting datasets and compare our method with other state-of-the-art methods. Finally, we conduct ablation experiments to evaluate the effectiveness of each component from our method.

4.1 Implementation details

We apply the Adam to optimize our network. Following [16, 19, 40], the initial learning rate is set to 1×10^{-5} . And the learning rate decreased by 0.99 times every two epochs. The weight decay is set to 1×10^{-4} . To optimize the network better, we set a magnification factor to enlarge the value of the ground truth density map. The magnification factor is set to 100 for ShanghaiTech Part_A and UCF-QNRF, 200 for ShanghaiTech Part_B, and 10 for UCF_CC_50. All training images are cropped and resized to 576×768 . The experiments are conducted under the Pytorch framework with a single NVIDIA GTX 2080Ti GPU.

4.2 Datasets and comparisons

4.2.1 Datasets

We evaluate our method on three commonly-used crowd counting datasets. The details of each dataset are shown in Table 2.

ShanghaiTech [66] includes 1,198 images with 330,165 annotated people. It is divided two parts: Part_A and Part_B. Part_A contains 482 highly crowded images randomly grabbed from the Internet. Part_B contains 716 images taken on downtown Shanghai's bustling streets.

Table 2 Summarization of ShanghaiTech Part_A, ShanghaiTech Part_B, UCF_CC_50, and UCF-QNRF

Property	ShanghaiTech Part_A	ShanghaiTech Part_B	UCF_CC_50	UCF-QNRF
Year	2016	2016	2013	2018
Resolution	different	768 × 1024	different	different
Number of images	482	719	50	1535
Train/Test	300/182	400/316	-	1201/1334
Max	3139	578	4543	12865
Min	33	9	94	49
Avg	501.4	123.6	1279.5	815.4
Total	241677	88488	63974	1251642

UCF_CC_50 [24] dataset includes 50 images with 63,974 annotated heads. It is a very challenging dataset because the number of people in each image varies greatly.

UCF-QNRF [25] dataset contains 1,535 images with 1,251,642 annotated heads. It includes different crowd congested scenes and large variation crowd distributions, which is also challenge for current crowd counting methods.

4.2.2 ShanghaiTech

The comparison results on ShanghaiTech dataset are presented in Table 3. We can see that the proposed COMAL outperforms other state-of-the-art methods in terms of MSE metrics. Specifically, compared with CSRNet, our COMAL achieves lower 8.6 and 17.9 in terms of MAE and MSE metrics, which benefits from the proposed SEM that can avoid a serious gridding effect [54]. Compared with SCAR, our COMAL also performs better counting accuracy, which benefits from the proposed SEM and DEM. The qualitative results in Fig. 5 further prove the effectiveness of our method. We observe from the fourth column that our proposed DEM can capture continuous scale changes of the crowd.

Table 3 Comparison results of different methods on ShanghaiTech dataset

Method	Part_A		Part_B	
	MAE	MSE	MAE	MSE
MCNN[66]	110.2	173.2	26.4	41.3
Switch-CNN[44]	90.4	135	21.6	33.4
CSRNet[30]	68.2	115	10.6	16
PACNN[47]	66.3	106.4	8.9	13.5
SCAR[16]	66.3	114.1	9.5	15.2
TEDNet[26]	64.2	109.1	8.2	12.8
DADNet[19]	64.2	99.9	8.8	13.5
RAZ-Net[35]	65.1	106.7	8.4	14.1
2-DA-CNN[69]	64.6	106.6	8.9	13.9
SFCN [†] with Pre-GCC[40]	64.8	107.5	7.6	13.0
PSCC+DCL[41]	65.0	108.0	8.1	13.3
COMAL(Ours)	59.6	97.1	7.8	12.4

Bold fonts indicate the best results

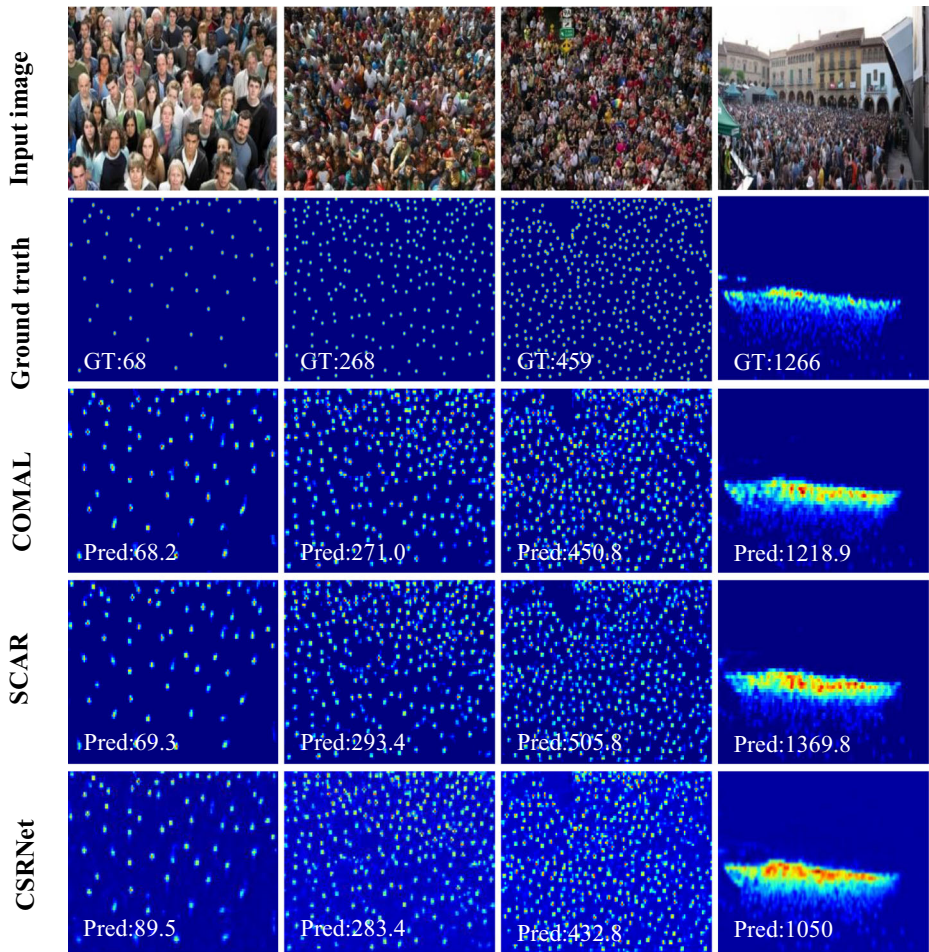


Fig. 5 Visualization results of different counting methods. From top to bottom, they are input images, ground truth, the results of COMAL, SCAR, and CSRNet, respectively

Besides, we conduct the further statistic analysis of the performance of the proposed COMAL on ShanghaiTech Part_A dataset. Specifically, as shown in Table 4, the ShanghaiTech Part_A dataset is divided into five crowd density levels. We compare the performance of COMAL, SCAR and CSRNet on the five crowd density levels, the comparison

Table 4 Summarization of five crowd density levels on ShanghaiTech Part_A

Density Level	Number of images	Max	Min	Avg	Total
1	36	65	191	133.5	4805
2	37	192	253	220.2	7926
3	36	254	379	307.1	11054
4	37	380	567	457.9	16485
5	36	578	2255	1007.7	38290

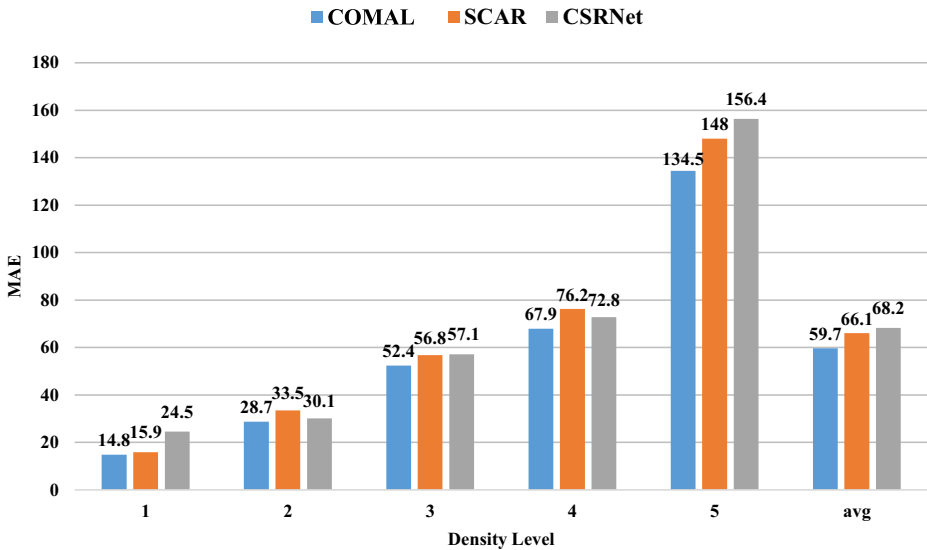


Fig. 6 Statistics analysis of SCAR [16], CSRNet [30] and our COMAL on different crowd density levels of ShanghaiTech Part_A dataset

details are shown in Fig. 6. We find that COMAL performs better than the other counting networks on all five crowd density levels, which demonstrates the effectiveness of the proposed method.

4.2.3 UCF_CC_50

Following some previous works, we perform five-fold cross validation to evaluate the performance of the proposed COMAL. The quantitative results on UCF_CC_50 are presented

Table 5 Comparison results of different methods on UCF_CC_50 dataset. “-” denotes the results are not provided by the original paper

Method	MAE	MSE
MCNN[66]	377.6	509.1
Switch-CNN[44]	318.1	439.2
CSRNet[30]	226.1	397.5
PACNN[47]	267.9	357.8
SCAR[16]	259	365.5
TEDNet[26]	249.4	354.5
DADNet[19]	285.5	389.7
RAZ-Net[35]	-	-
2-DA-CNN[69]	252.0	340.3
SFCN†with Pre-GCC[40]	214.2	318.2
PSCC+DCL[41]	-	-
COMAL(Ours)	231.9	333.7

Bold fonts indicate the best results

in Table 5. Compared with other state-of-the-art methods, SFCN[†] with Pre-GCC [40] uses synthetic data to expand the limited training images of UCF50 and achieves a better count performance. However we see that our COMAL achieves state-of-the-art results in methods without synthetic data pretraining, which further proves the superiority of our method. Although there are huge variation crowd distribution in this dataset, COMAL performs better 17.5 and 20.8 than the performance of TEDNet in terms of MAE and MSE metrics, which is a significant progress for crowd counting task.

4.2.4 UCF-QNRF

The performance of the proposed COMAL on UCF-QNRF is presented in Table 6. We can see that COMAL outperforms the other methods in methods without synthetic data pretraining, which further proves the superiority of our method. Compared with the performance of TEDNet, COMAL achieves lower 10.9 in terms of MAE, which further proves the effectiveness of our method. Without the help of synthetic data, our method still achieves similar performance on MAE metrics compared to Pre-GCC [40].

4.3 Ablation study

4.3.1 The effectiveness of different structures of COMAL

To evaluate the effectiveness of different structures, we design four variants of COMAL and conduct extensive experiments on the ShanghaiTech Part_A dataset. The details of the four variants are as follows.

The first model is the first 10 layers of VGG-16, which is denoted as **VGG-10**. The second model places the proposed SEM on the first model, which is represented as **VGG-10 + DEM**. Based on the second model, the third model changes the first 10 layers of VGG-16 into the first 13 layers of VGG-16, and is denoted as **VGG-13 + DEM**. The fourth model adds DEBs into the third model and is represented as **VGG-13 + SEM + DEM**.

Table 6 Comparison results of different methods on UCF-QNRF dataset

Method	MAE	MSE
MCNN[66]	-	-
Switch-CNN[44]	228	445
CSRNet[30]	120.3	208.5
PACNN[47]	-	-
SCAR[16]	-	-
TEDNet[26]	113	188
DADNet[19]	113.2	189.4
RAZ-Net[35]	116	195
2-DA-CNN[69]	-	-
SFCN [†] with Pre-GCC[40]	102.0	171.4
PSCC+DCL[41]	108	182
COMAL(Ours)	102.1	178.3

Bold fonts indicate the best results

Qualitative and quantitative results are displayed in Fig. 7 and Table 7. We can see that the counting performance is continually improved with the injection of the proposed

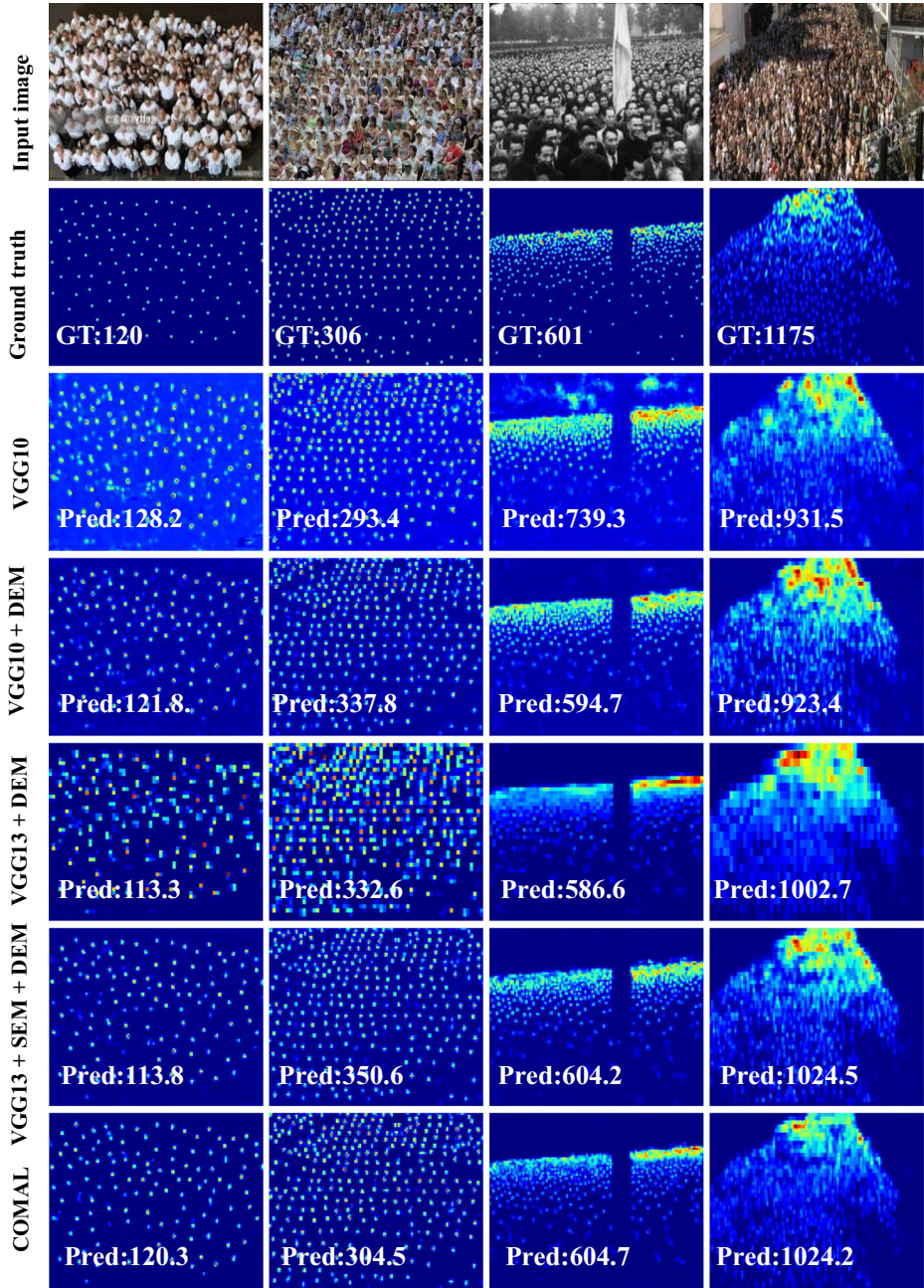


Fig. 7 Visualization results of COMAL with different components. From top to bottom, they are input images, ground truth, VGG-10, VGG-10 w/ DEM, VGG-13 w/DEM, VGG-13 w/SEM and DEM, and COMAL, respectively

Table 7 Comparison results of COMAL with different structures on ShanghaiTech Part_A dataset

Method	MAE	MSE
VGG-10(baseline)	74.2	122
VGG-10+DEM	67.4	113.7
VGG-13+DEM	66.4	106.9
VGG-13+SEM+DEM	66.2	105.7
VGG-13+SEM+DEM+CEM (COMAL)	59.6	97.1

Bold fonts indicate the best results

components into the counting model, and achieves the best results with all the proposed components, which proves the effectiveness of our method. Specifically, compared with the fourth model, COMAL performs better 6.6 and 8.6 in terms of MAE and MSE metrics, which demonstrates the importance of context information generated by CEM for final crowd counting.

4.3.2 The effectiveness of the components of COMAL

We design three different structures to verify the effectiveness of each component in COMAL. As shown in Table 8, $C(N_c)$ represents convolutional layer with N_c filters. From the first row and the last row of Table 8, we can see that the counting accuracy drops when we use the convolutional layer to replace the SEM, which demonstrates that the high-level semantic features are important for final crowd counting. Besides, compare the performance of the second row and the third row, we find that the method with DEM performs better than the method without DEM. This is contributed to the multi-scale features generated by the DEM. For the last two rows of Table 8, we can see that the performance of CEM outperforms the performance of CBAM [59] hugely, which further proves the effectiveness of our proposed CEM.

4.3.3 The number of DEB

We explore the effect of the number of DEB to the final counting accuracy. The comparison results are displayed in Table 9. We can see that with the increased number N_{DEB} of DEB, the counting performance of COMAL is improved, and COMAL achieves the best results when the N_{DEB} is equal to 3, which is benefited from the scale diversities provided by DEBs. However, when the N_{DEB} is larger than 3, the counting performance drops. The reason is that more DEBs increase the complexities of the network and hinder the optimization process of the counting network.

Table 8 Comparison results of the components of COMAL on ShanghaiTech Part_A dataset

Method	MAE	MSE	Params(MB)
VGG-13+C(512)+DEM+CEM	62.9	103.6	21.49
VGG-13+SEM+C(512,256,128,64)+CEM	61.5	106.7	21.53
VGG-13+SEM+DEM+CBAM [59]	65.4	109.7	21.47
VGG-13+SEM+DEM+CEM(COMAL)	59.6	97.1	21.49

Table 9 Comparison results of COMAL with different numbers of DEB on ShanghaiTech Part_A dataset

Bold fonts indicate the best results

Number	MAE	MSE
1	65.0	108.7
2	63.6	105.0
3	59.6	97.1
4	61.1	103.2

Table 10 Comparison results of different designs of CEM on ShanghaiTech Part_A dataset

Bold fonts indicate the best results

Method	MAE	MSE
COMAL w/CAM	63.4	101.2
COMAL w/PAM	60.2	100.2
COMAL w/CEM	59.6	97.1

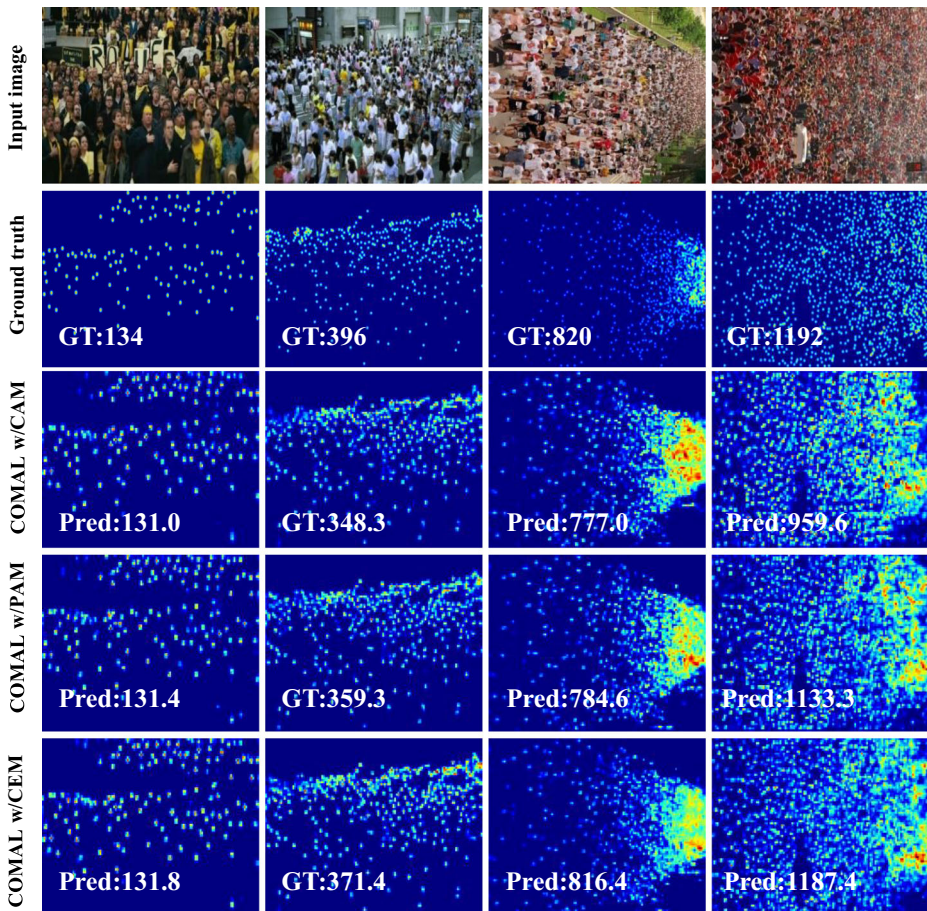


Fig. 8 Visualization results of COMAL with different attention modules. From top to bottom, they are input images, ground truth density maps, COMAL w/CA, COMAL w/PA, and COMAL, respectively

4.3.4 The design of CEM

To evaluate the rationality of CEM, we explore the performance of COMAL with only PAM (COMAL w/PAM) or CAM (COMAL w/CAM) on ShanghaiTech Part_A dataset. The quantitative results are shown in Table 10. We can see that the counting accuracy has continually improved with the help of CA and PA. The model achieves the best results when the model with CEM, which demonstrates that effectiveness of our method. The qualitative results in Fig. 8 further prove the importance of CEM to final counting accuracy.

5 Conclusions

In this paper, we propose the COMAL for multi-scale crowd counting. We use the first 13 layers of VGG-16 as the encoder to extract features, and adopt the proposed decoder to process the extracted features for final density estimation. COMAL is evaluated on three challenging crowd counting datasets and achieves superior results compared with other state-of-the-art methods. However, COMAL owns lots of network parameters which is not suitable for the devices with limited computation resources. Besides, we only model the image spatial context information and do not consider to extract temporal information of video. Thus, in future work, we can explore our COMAL to video crowd counting task in a lightweight design.

Acknowledgements This work is supported by Natural Science Foundation of Shanghai under Grant No. 19ZR1455300, and National Natural Science Foundation of China under Grant No. 61806126.

References

1. Bai H, Wen S, Gary Chan S-H (2019) Crowd counting on images with scale variation and isolated clusters. In: Proceedings of the IEEE Conference on Computer Vision, 0–0
2. Boominathan L, Kruthiventi SSS, Babu RV (2016) Crowdnet a deep convolutional network for dense crowd counting. In: Proceedings of the 24th ACM international conference on Multimedia, pp. 640–644
3. Cao X, Wang Z, Zhao Y, Su F (2018) Scale aggregation network for accurate and efficient crowd counting. In: Proceedings of the European Conference on Computer Vision, pp. 734–750
4. Cao Y, Xu J, Lin S, Wei F, Han H (2019) Gcnet: Non-local networks meet squeeze-excitation networks and beyond. In: Proceedings of the IEEE International Conference on Computer Vision Workshops, pp. 0–0
5. Cao J, Yang B, Nan W, Wang H, Cai Y (2020) Robust crowd counting based on refined density map. *Multimed Tools Appl* 79(3):2837–2853
6. Chan AB, Liang Z-SJ., Vasconcelos N (2008) Privacy preserving crowd monitoring: Counting people without people models or tracking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–7
7. Chen X, Bin Y, Sang N, Gao C (2019) Scale pyramid network for crowd counting. In: Winter Conference on Applications of Computer Vision, pp. 1941–1950
8. Chen C, Hui Q, Xie W, Wan S, Zhou Y, Pei Q (2021) Convolutional neural networks for forecasting flood process in internet-of-things enabled smart city. *Comput Netw* 107744:186
9. Chen L-C, Papandreou G, Schroff F, Adam H (2017) Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv: [1706.05587](https://arxiv.org/abs/1706.05587)
10. Cholakkal H, Sun G, Khan FS, Shao L (2019) Object counting and instance segmentation with image-level supervision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 12397–12405
11. Deb D, Ventura J (2018) An aggregated multicolumn dilated convolution network for perspective-free counting. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 195–204

12. Di K, Chan A (2018) Crowd counting by adaptively fusing predictions from an image pyramid. arXiv preprint arXiv:[1805.06115](https://arxiv.org/abs/1805.06115)
13. Gao G, Gao J, Liu Q, Wang Q, Wang Y (2020) Cnn-based density estimation and crowd counting: A survey. arXiv preprint arXiv:[2003.12783](https://arxiv.org/abs/2003.12783)
14. Gao J, Han T, Wang Q, Yuan Y (2019) Domain-adaptive crowd counting via inter-domain features segregation and gaussian-prior reconstruction. arXiv preprint arXiv:[1912.03677](https://arxiv.org/abs/1912.03677)
15. Gao Z, Li Y (2020) Shaohua Wan. Exploring deep learning for view-based 3d model retrieval. *ACM Trans Multimedia Comput Commun Appl* 16(1):1–21
16. Gao J, Qi W, Yuan Y (2019) Scar: Spatial-/channel-wise attention regression networks for crowd counting. *Neurocomputing* 363:1–8
17. Gao J, Wang Q, Li X (2019) Pcc net: Perspective crowd counting via spatial convolutional network. *IEEE Trans Circuits Syst Video Technol* 30(10):3486–3498
18. Gao J, Wang Q, Yuan Y (2019) Feature-aware adaptation and structured density alignment for crowd counting in video surveillance. arXiv preprint arXiv:[1912.03672](https://arxiv.org/abs/1912.03672)
19. Guo D, Li K, Zha Z-J, Wang M (2019) Dadnet: Dilated-attention-deformable convnet for crowd counting. In: *Proceedings of the 27th ACM International Conference on Multimedia*, pp. 1823–1832
20. He J, Deng Z, Zhou L, Wang Y, Yu Q (2019) Adaptive pyramid context network for semantic segmentation. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7519–7528
21. He K, Zhang X, Ren S, Sun J (2015) Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Trans Pattern Anal Mach Intell* 37(9):1904–1916
22. Hossain M, Hosseinzadeh M, Chanda O, Wang Y (2019) Crowd counting using scale-aware attention networks. In: *Winter Conference on Applications of Computer Vision*, pp. 1280–1288
23. Huiyuan Fu, Ma H, Xiao H (2014) Scene-adaptive accurate and fast vertical crowd counting via joint using depth and color information. *Multimed Tools Appl* 73(1):273–289
24. Idrees H, Saleemi I, Seibert C, Shah M (2013) Multi-source multi-scale counting in extremely dense crowd images. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2547–2554
25. Idrees H, Tayyab M, Athrey K, Zhang D, Al-Maadeed S, Rajpoot N, Shah M (2018) Composition loss for counting, density map estimation and localization in dense crowds. In: *Proceedings of the European Conference on Computer Vision*, pp. 532–546
26. Jiang X, Xiao Z, Zhang B, Zhen X, Cao X, Doermann D, Shao L (2019) Crowd counting and density estimation by trellis encoder-decoder networks. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6133–6142
27. Junyu Gao, Lin W, Zhao B, Wang D, Gao C, Wen J (2019) C³, 3 framework: An open-source pytorch code for crowd counting. arXiv preprint arXiv:[1907.02724](https://arxiv.org/abs/1907.02724)
28. Khare N, Devan P, Chowdhary CL, Bhattacharya S, Singh G, Singh S, Yoon B (2020) Smo-dnn: Spider monkey optimization and deep neural network hybrid classifier model for intrusion detection. *Electronics* 9(4):692
29. Laradji IH, Rostamzadeh N, Pinheiro PO, Vazquez D, Schmidt M (2018) Where are the blobs Counting by localization with point supervision. In: *Proceedings of the European Conference on Computer Vision*, pp. 547–562
30. Li Y, Zhang X, Chen D (2018) Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1091–1100
31. Lin Tsung-Yi, Dollár P, Girshick R, He K, Hariharan B, Belongie S (2017) Feature pyramid networks for object detection. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2117–2125
32. Liu J, Gao C, Meng D, Hauptmann AG (2018) Decidenet: Counting varying density crowds through attention guided detection and density estimation. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5197–5206
33. Liu W, Salzmann M, Fua P (2019) Context-aware crowd counting. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5099–5108
34. Liu Y, Shi M, Zhao Q, Wang X (2019) Point in, box out Beyond counting persons in crowds. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6469–6478
35. Liu C, Weng X, Mu Y (2019) Recurrent attentive zooming for joint crowd counting and precise localization. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1217–1226
36. Moustafa AN, Goma W (2020) Gate and common pathway detection in crowd scenes and anomaly detection using motion units and lstm predictive models. *Multimed Tools Appl* 79(29):20689–20728

37. Onoro-Rubio D, López-Sastre R. J (2016) Towards perspective-free object counting with deep learning. In: Proceedings of the European Conference on Computer Vision, pp. 615–629
38. Pandey A, Pandey M, Singh N, mela AbhaTrivedi. (2020) Kumbh a case study for dense crowd counting and modeling. *Multimed Tools Appl* 79(25):17837–17858
39. Qi W, Gao J, Lin W, Yuan Y (2019) Learning from synthetic data for crowd counting in the wild. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 8198–8207
40. Qi W, Gao J, Lin W, Yuan Y (2021) Pixel-wise crowd understanding via synthetic data. *Int J Comput Vis* 129(1):225–245
41. Qi W, Lin W, Gao J, Li X (2020) Density-aware curriculum learning for crowd counting. *IEEE Trans Cybern.* <https://doi.org/10.1109/TCYB.2020.3033428>
42. Ranjan V, Le H, Hoai M (2018) Iterative crowd counting. In: Proceedings of the European Conference on Computer Vision, pp. 270–285
43. Sam DB, Sajjan NN, Babu RV, Srinivasan M (2018) Divide and grow: Capturing huge diversity in crowd images with incrementally growing cnn. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3618–3626
44. Sam DB, Surya S, Venkatesh Babu R (2017) Switching convolutional neural network for crowd counting. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4031–4039
45. Sam DB, Venkatesh BR (2018) Top-down feedback for crowd counting convolutional neural network. arXiv preprint arXiv: [1807.08881](https://arxiv.org/abs/1807.08881)
46. Shen Z, Xu Y, Ni B, Wang M, Hu J, Yang X (2018) Crowd counting via adversarial cross-scale consistency pursuit. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5245–5254
47. Shi M, Yang Z, Xu C, Chen Q (2019) Revisiting perspective information for efficient crowd counting. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7279–7288
48. Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv: [1409.1556](https://arxiv.org/abs/1409.1556)
49. Sindagi VA, Patel VM (2017) Generating high-quality crowd density maps using contextual pyramid cnns. In: Proceedings of the IEEE Conference on Computer Vision, pp. 1861–1870
50. Somayaji SRK, Alazab M, Manojm K, Bucchiarone A, Chowdhary CL, Gadekallu T (2020) A framework for prediction and storage of battery life in iot devices using dnn and blockchain. In: 2020 IEEE Globecom Workshops (GC Wkshps), pp. 1–6
51. Swarna Priya RM, Maddikunta PKR, Parimala M, Koppu S, Gadekallu TR, Chowdhary CL, Alazab M (2020) An effective feature engineering for dnn using hybrid pca-gwo for intrusion detection in iomt architecture. *Comput Commun* 160:139–149
52. Topkaya IS, Erdogan H, Porikli F (2014) Counting people by clustering person detector outputs. In: 2014 11th IEEE International Conference on Advanced Video and Signal Based Surveillance, pp. 313–318
53. Tripathi G, Singh K, Vishwakarma DK (2019) Convolutional neural networks for crowd behaviour analysis: a survey. *Vis Comput* 35(5):753–776
54. Wang P, Chen P, Yuan Y, Liu D, Huang Z, Hou X, Garrison Cottrell (2018) Understanding convolution for semantic segmentation. In: Winter Conference on Applications of Computer Vision, pp. 1451–1460. IEEE
55. Wang S, Lu Y, Zhou T, Di H, Lu L, Zhang L (2020) Sclnet: Spatial context learning network for congested crowd counting. *Neurocomputing* 404:227–239
56. Wang J, Wei Z, Zhang T, Zeng W (2016) Deeply-fused nets. arXiv preprint arXiv: [1605.07716](https://arxiv.org/abs/1605.07716)
57. Wang C, Zhang H, Yang L, Liu S, Cao X (2015) Deep people counting in extremely dense crowds. In: Proceedings of the 23rd ACM international conference on Multimedia, pp. 1299–1302
58. Wang S, Zhao H, Wang W, Di H, Shu X (2017) Improving deep crowd density estimation via pre-classification of density. In: Proceedings of the International Conference on Neural Information Processing, pp. 260–269
59. Woo S, Park J, Lee J-Y, Kweon IS (2018) Cbam: Convolutional block attention module. In: Proceedings of the European conference on computer vision (ECCV), pp. 3–19
60. Xie Y, Lu Y, Wang S (2020) Rsanet: Deep recurrent scale-aware network for crowd counting. In: Proceedings of the IEEE International Conference on Image Processing, pp. 1531–1535
61. Xingjiao Wu, Zheng Y, Ye H, Wenxin Hu, Yang J (2019) Liang He. Adaptive scenario discovery for crowd counting. In: Proceedings of the IEEE International Conference on Acoustics Speech and Signal Processing, pp. 2382–2386
62. Xu C, Qiu K, Fu J, Bai S, Xu Y, Bai X (2019) Learn to scale: Generating multipolar normalized density maps for crowd counting. In: Proceedings of the IEEE Conference on Computer Vision, pp. 8382–8390

63. Zhang C, Li H, Wang X, Yang X (2015) Cross-scene crowd counting via deep convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 833–841
64. Zhang Lu, Shi M, Chen Q (2018) Crowd counting via scale-adaptive convolutional neural network. In: Winter Conference on Applications of Computer Vision, pp. 1113–1121
65. Zhang Z, Zhang X, Peng C, Xue X, Sun J (2018) Exfuse: Enhancing feature fusion for semantic segmentation. In: Proceedings of the European Conference on Computer Vision, pp. 269–284
66. Zhang Y, Zhou D, Chen S, Gao S, Ma Y (2016) Single-image crowd counting via multi-column convolutional neural network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 589–597
67. Zhao H, Shi J, Qi X, Wang X, Jia J (2017) Pyramid scene parsing network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2881–2890
68. Zhao M, Zhang J, Zhang C, Zhang W (2019) Leveraging heterogeneous auxiliary tasks to assist crowd counting. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 12736–12745
69. Zou Z, Li C, Zheng Y, Shoukun Xu (2020) Two stages double attention convolutional neural network for crowd counting. *Multimed Tools Appl* 79(39):29145–29159

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.