Noname manuscript No.
(will be inserted by the editor)

Towards Trustworthy Crowd Counting by Distillation Hierarchical Mixture of Experts for Edge-based Cluster Computing

Jing-an Cheng [1†] · **Qilei Li** [2†] · **Alireza Souri** [3] ·
Xiang Lei [4] · **Chen Zhang** [1] · **Mingliang Gao** [1*]

Received: date / Accepted: date

Abstract Crowd counting plays a crucial role in analyzing and understanding crowd behavior. Existing models generally rely on large parameters to achieve high counting accuracy. This increases computational demands and limits deployment on mobile edge devices. On the other hand, lightweight networks often face difficulties in managing scale variation and show poor performance in complex crowd counting tasks because of their simplified design. To tackle these challenges, we propose a crowd counting model, termed Distillation Hierarchical Mixture of Experts (DHMoE). It is composed of two primary components. The first is a knowledge distillation training model. It transfers fine-grained knowledge from the pre-trained teacher model to the lightweight student model and improves counting accuracy. Second, to solve the problems of scale variation and complex environments, a hierarchical mixture of experts (HMoE) is proposed. The four stages of the student model are organized into four experts, where each network handles crowd features at a different scale. This approach effectively addresses scale variation and improves counting accuracy in diverse environments. Experimental results on four crowd and four vehicle datasets demonstrate that the proposed DHMoE achieves excellent counting accuracy while maintaining a lightweight design. The code is available at https://github.com/sdut-jacheng/DHMoE.

Keywords Crowd counting, Lightweight network, Mixture of experts (MoE), Knowledge distillation

🖂 Mingliang Gao
E-mail: mlgao@sdut.edu.cn
†: Authors contributed equally.

1 School of Electrical and Electronic Engineering, Shandong University of Technology, Zibo 255000, China.
2 School of Electronic Engineering and Computer Science, Queen Mary University of London, London, E1 4NS, United Kingdom.
3 Department of Computer Engineering, Haliç University, Istanbul 34394, Turkey.
4 Zhiyang Innovation Co., Ltd., Jinan 250101, China.
Jing-an Cheng, Mingliang Gao, and Chen Zhang are with the School of Electrical and Electronic Engineering, Shandong University of Technology, Zibo, 255000, China (e-mail: 23404020560@stumail.sdut.edu.cn, mlgao@sdut.edu.cn, 23404020559@stumail.sdut.edu.cn). Qilei Li is with the School of Electronic Engineering and Computer Science, Queen Mary University of London, London, United Kingdom (e-mail: q.li@qmul.ac.uk). Alireza Souri is with the Department of Computer Engineering, Haliç University, Istanbul 34394, Turkey (e-mail: alirezasouri@halic.edu.tr). Xiang Lei is with Zhiyang Innovation Co., Ltd., Jinan, 250101, China (e-mail: leixiang@zhiyang.com.cn)

1 Introduction

The task of crowd counting is to derive information about crowd density by quantifying the number of individuals in an image or video. It is crucial to various domains, *e.g.,* secure, efficient decision-making and management. As deep learning continues to advance rapidly, numerous high-performing models [62, 55, 5] for crowd counting have been developed. These models exhibit impressive accuracy and robustness, particularly in complex scenarios. However, they often depend on a large number of parameters and complex architectures, which demand substantial computational resources during inference. This significant computational burden presents a major limitation, significantly when deploying these models on edge devices or embedded systems with constrained processing capabilities.

To be specific, most state-of-the-art (SOTA) deep learning-based models often involve millions or billions of parameters [48]. They require greater depth and complexity to maintain accuracy in high-density crowd counting. As depicted in Fig. 1, the analysis reveals a fundamental trade-off in model design: achieving enhanced accuracy often requires an increase in parameters and FLOPs. Although these methods have achieved high counting accuracy, their substantial parameters and extensive FLOPs demands present significant challenges for real-time deployment on resource-constrained devices.

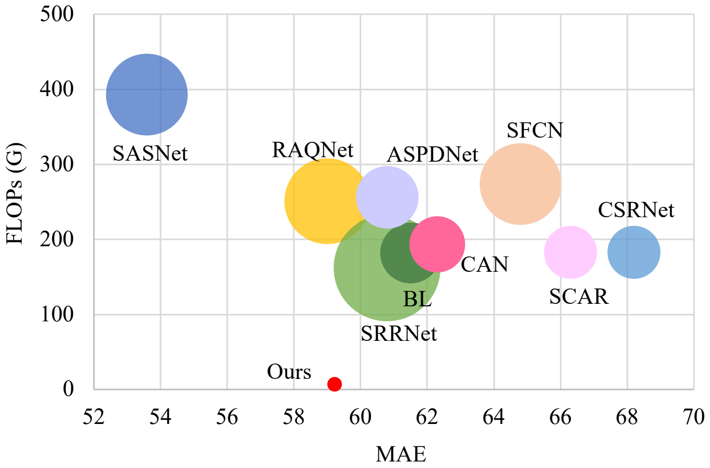

Fig. 1: Comparison of parameters, FLOPs, and mean absolute error (MAE) among SOTA counting models. Lower MAE indicates higher accuracy in counting. Higher FLOPs indicate that the model requires more computational resources. The bubble size in the figure reflects the parameters of each model, with larger bubbles signifying larger parameters. These SOTA models include: CSRNet [29], BL [38], SFCN [59], RAQNet [63], ASPDNet [12], CAN [35], SCAR [15], SASNet [18], SRRNet [18] and ours).

To achieve efficient crowd counting with reduced computational costs, researchers have recently focused on lightweight CNN models [21, 64, 20]. However, their accuracy often falls short of practical expectations [54]. Another line approach is to compress high-complexity crowd counting models through pruning and quantization, which reduces parameters and computational costs [48]. Nevertheless, this technique successfully decreases computational complexity, it can also cause a loss in accuracy when parameters are significantly reduced. Thus, the challenge remains in compressing models efficiently without sacrificing high counting accuracy.

In recent years, knowledge distillation (KD) has emerged as a widely recognized and effective technique for model compression [16]. It transfers knowledge from a large teacher model to a smaller student model and ensures that the smaller model maintains both computational efficiency and accuracy [24]. Liu *et al.* [33] developed a structured framework for knowledge transfer. It includes two key modules: intra-layer pattern transfer and inter-layer relation transfer. These components guide the acquisition of the necessary features and help the student model gain cross-layer knowledge. Although knowledge distillation has proven beneficial for model compression and maintaining accuracy, it encounters certain limitations in crowd counting tasks under varied conditions. Specifically, a single student model may struggle to achieve optimal counting performance in complex and high-density crowd situations [16].

In response to the aforementioned challenges, we present a counting method that integrates knowledge distillation with a Hierarchical mixture of experts (DHMoE). First, knowledge distillation transfers complex knowledge from a large teacher model to multiple lightweight expert models. This process allows the expert models to inherit the feature extraction and data processing capabilities from the teacher model. It helps the student models maintain high counting accuracy in various settings. Overall, the contributions of this work are summarized as follows.

1. We propose a distillation hierarchical mixture of experts (DHMoE) model for crowd counting. It can reduce parameters and computational costs while preserving high accuracy.
2. We propose the hierarchical mixture of experts (HMoE) combination with knowledge distillation to minimize the performance difference between the student and teacher models. It improves the generalization of the student model across varied scenarios and effectively addresses challenges related to scale variation.
3. Experiments on several widely-used crowd and vehicle counting datasets show that the DHMoE delivers superior performance, which confirms its effectiveness and robustness in various contexts.

2 Related work

2.1 Knowledge Distillation

Knowledge distillation (KD) involves extracting knowledge from a large teacher model and transferring it to a smaller student model [16]. This approach fits well with lightweight networks [24]. By distilling knowledge from various aspects, such as model outputs, feature layers, and attention mechanisms, the latest distillation techniques have significantly enhanced the performance of student models. Tian *et al.* [52] proposed contrastive representation distillation. It aims to align the representations of the student model with those of the teacher model. KD promotes the development of lightweight models suitable for deployment on resource-constrained devices, which provides considerable advantages to computer vision. Jiao *et al.* [27] created TinyBERT, which is a streamlined and accelerated version of BERT. It delivers competitive results on several NLP benchmarks. KD shows particular effectiveness in the creation of efficient models for real-time applications within the computer vision domain. Heo *et al.* [1] thoroughly analyzed distillation methods to alleviate the computational burden of deep convolutional neural networks. Recent developments in knowledge distillation have demonstrated their usefulness in enhancing model performance and scalability. Thus, they are beneficial for the deployment of models in resource-constrained domains such as computer vision and NLP. This study applies the knowledge distillation framework to develop a lightweight neural network model. It seeks to reduce both the parameters and computational complexity while preserving the performance of the model.

2.2 Mixture of Experts

The mixture of experts (MoE) is a widely studied model architecture in the fields of deep learning and machine learning [2]. Its primary objective is to combine multiple expert networks to enhance both model performance and computational efficiency [39]. MoE has been successfully applied to various tasks and has demonstrated significant advantages in solving complex problems. Wang *et al.* [60] proposed the multi-gated mixture of experts architecture. It employs multiple gating networks that combine different experts with specific weight sets to achieve multi-task learning. Du *et al.* [9] developed a hierarchical mixture density expert architecture. It solves multi-scale problems by finding the optimal solution through the collaboration and competition of experts operating at various scales. Fedus *et al.* [10] simplified the routing algorithm using the MoE framework and developed an improved model that reduces both communication and computation costs. This approach successfully trained sparse models with up to a trillion parameters. Reisser *et al.* [43] proposed a federated mixture of experts, which addresses data heterogeneity by adaptively selecting and training user-specific ensemble members. The continuous evolution of MoE models shows their potential to handle intricate tasks across various scales and domains. This highlights their important role in improving model efficiency and scalability. This study presents an HMoE framework. It aims to enhance the capacity of the student model in handling multi-scale and complex environments.

2.3 Lightweight Networks

To simplify the network and improve computational efficiency, lightweight network models have drawn extensive attention from the research community [54]. Howard *et al.* [21] developed the mobilenetv1 by employing depthwise separable convolutions in place of regular convolutions to reduce the parameters of models. Sandler *et al.* [46] proposed mobilenetv2, which streamlines the model from mobilenetv1 by adding residual structures and pointwise convolutions. Zhang *et al.* [64] introduced shufflenetv1, which reduces flops with group convolutions and facilitates information exchange between different groups via channel shuffling. Ma *et al.* [36] re-optimized the network structure of shufflenetv1 by introducing four criteria and proposed the shufflenetv2. Furthermore, Han *et al.* [20] identified redundancy in feature maps extracted through convolution and developed the ghost module to mitigate feature redundancy. Inspired by the aforementioned studies, we propose Distillation Hierarchical Mixture of Experts. It employs knowledge distillation to enhance the counting accuracy of the student model. Moreover, an HMoE architecture is incorporated to effectively handle multi-scale and complex scenarios. Therefore, the model achieves greater efficiency, which allows it to be deployed in resource-constrained environments effectively.

3 Methodology

3.1 Overall Framework

The DHMoE framework is shown in Fig. 2. It consists of a teacher model, a student model, and their respective decoders. The teacher model is typically pre-trained and contains large parameters. Conversely, the student model has fewer parameters and operates under the guidance of the teacher model. Moreover, the framework incorporates an HMoE structure, which helps the student model acquire deep knowledge from the teacher model. The decoder reconstructs the predicted density map to the input size by applying deconvolution layers.

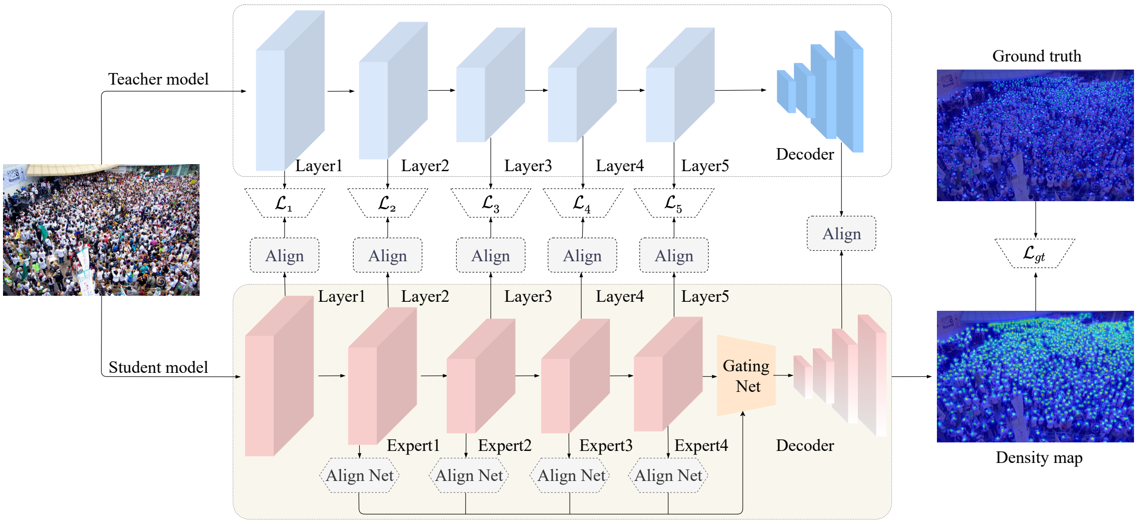

Fig. 2: The pipeline of the DHMoE for crowd counting. “Alig” refers to a 1 × 1 convolution layer that adjusts the channel dimensions between the student and teacher models at each stage. “Align Net” uses bilinear interpolation and 1 × 1 convolution to harmonize the feature map dimensions and channel numbers of the first three experts with those of Expert4 to ensure consistency.

3.2 Knowledge Distillation Framework

To maintain a balance between parameters and counting accuracy, a knowledge distillation framework is proposed with two models: a teacher model with large parameters and a lightweight student model. Both networks incorporate an encoder for feature extraction and a decoder that adjusts the density map size to match the ground truth. Fig. 2 illustrates the knowledge distillation framework, which involves feature-level knowledge transfer. Before knowledge transfer, hierarchical features must be extracted from both the teacher and student models. Specifically, the input image I is processed by both the teacher and student models to extract features. At each layer, the teacher model extracts features denoted as T_i, while the student model extracts features denoted as S_i, where i indicates the i-th layer.

During the training process of the student model, the teacher model remains in a frozen state to ensure that the knowledge of the teacher model is not affected by the updates of the parameter of the student model. To ensure precise guidance is given to the student model at each stage, DHMoE aligns the corresponding stages of both the teacher and student models. As the channel numbers differ at each stage, direct feature similarity computation is not feasible. Thus, 1×1 convolution layers are employed to align the channel dimensions and ensure they match the channels of the teacher model. Cosine similarity is applied to evaluate the similarity of features. The formula for this computation is as follows,

$$S_i' = \mathrm{Conv1} \times 1(S_i), \tag{1}$$

$$\mathrm{CS}(S_i', T_i) = \frac{S_i' \cdot T_i}{\|S_i'\| \|T_i\|}, \tag{2}$$

$$\mathcal{L}_i = 1 - \mathrm{CS}(S_i', T_i), \tag{3}$$

where $\mathrm{CS}(\cdot)$ denotes the Cosine Similarity. $\|S_i'\|$ and $\|T_i\|$ represent the norms of S_i' and T_i, respectively. $\mathcal{L}_i$ refers to the loss corresponding to each layer. The final distillation loss is formulated as,

$$\mathcal{L}_{KD} = \sum_{i=0}^{5} \mathcal{L}_i, \tag{4}$$

where $\mathcal{L}_i$ denotes the loss at the i-th layer. Through the reduction of the specified loss function, the DHMoE significantly enhances knowledge transfer and improves the performance of the student model.

3.3 Hierarchical Mixture of Experts

Knowledge distillation approaches involve student model learning by emulating the feature distribution of the teacher model. However, simpler student models may struggle to capture the details in teacher models accurately. The challenge becomes particularly pronounced when performing crowd recognition in complex scenarios. Additionally, a single student model often generalizes excessively across various data scales. This over-generalization limits its ability to make accurate distinctions between sparse and dense regions of a crowd. As a result, the crowd counting accuracy decreases, especially in areas where crowd density varies significantly.

To address these issues outlined above, an HMoE architecture is integrated into the student model, as shown in Fig. 3. The last four stages of the student model are assigned to an expert network. Specifically, the first three expert networks incorporate two OSBottlenecks [66]. The architecture of the OSBottleneck is shown in Fig. 4. The fourth expert network consists solely of a 1×1 convolution layer. Each expert is responsible for processing distinct feature levels and handles information extraction at specific scales.

Specifically, given an input image I, the student model processes each stage through its expert networks and produces feature representations F_i at different levels. Then, a gating mechanism, which includes a fully connected layer and the Softmax function, generates dynamic weights W_i to control how the outputs of the expert networks are combined. It allows for the effective integration of diverse feature representations. Direct multiplication is not possible because of the mismatch in

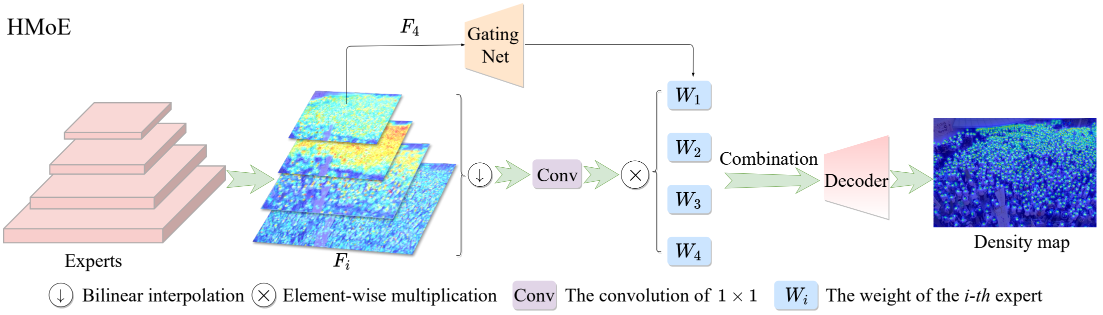

Fig. 3: The framework of the HMoE model. Features from each expert are subjected to bilinear interpolation and 1 × 1 convolution for channel adjustment. Afterwards, they are combined based on their assigned weights to form the final feature map.

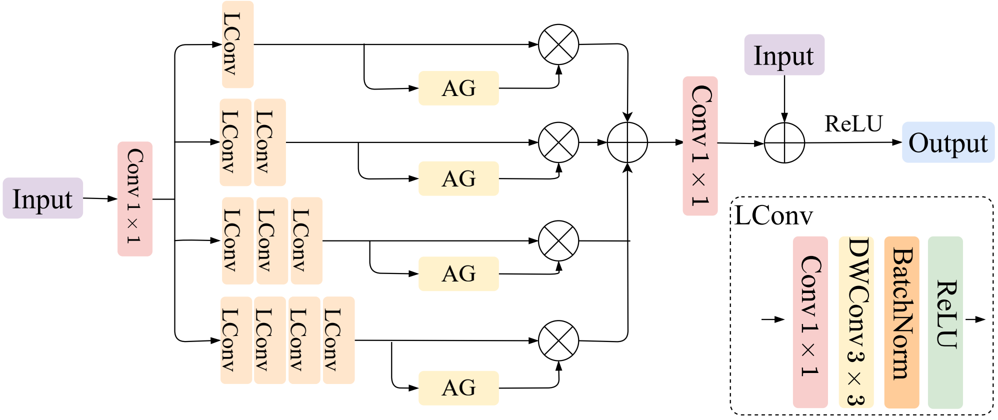

Fig. 4: The architecture of the OSBottleneck. AG stands for aggregation gate, LConv represents lite 3 × 3 convolution, and DW refers to depth-wise.

the dimensions and channels of the feature maps among experts. Therefore, bilinear interpolation and a 1 × 1 convolution are used to align the sizes and channel configurations of all feature maps with those of expert4. This process can be expressed as,

$$W_i = \mathrm{Softmax}(\mathrm{Linear}(F_4)), \tag{5}$$

$$F' = \sum_{i=1}^{4} W_i \times F_i, \tag{6}$$

where F' indicates the final fusion feature output from the HMoE, while $\mathrm{Linear}(\cdot)$ denotes the fully connected layer. F_4 represents the global information gathered from the output of the last expert network. It is formulated as,

$$F_4 = \frac{1}{H \times W} \sum_{h=1}^{H} \sum_{w=1}^{W} f_{h,w}, \tag{7}$$

where H and W correspond to the height and width of the feature map, respectively. $f_{h,w}$ denotes the eigenvalue at the (h, w) coordinate within the feature map.

Within the HMoE structure, each expert is assigned to process a specific subset of features. This design prevents the student model from introducing unnecessary complexity in each task. Additionally, the gating mechanism dynamically allocates computational resources. It ensures the student model acquires and integrates multi-scale features from the teacher model while maintaining a lightweight design. This approach significantly improves knowledge distillation efficiency and strengthens the generalization capability of the student model.

3.4 Ground Truth Generation

Ground truth density maps are generated as supervision through the conventional focal inverse distance transform map method [30]. This method provides a precise representation of crowd density by accounting for the distance between each pixel and the nearest annotated head. The density map generation is formulated as,

$$F_{\text{gt}} = \frac{1}{P(x,y)^{\alpha \times P(x,y)+\beta} + C}, \tag{8}$$

where α and β are empirically set to 0.02 and 0.75, respectively, based on prior studies [17, 61, 7]. The constant C is assigned a value of 1 to avoid division by zero and ensure numerical stability. $P(x, y)$ represents the Euclidean distance between the pixel at coordinates (x, y) and the nearest annotated object location (x', y').

3.5 Loss Function

The MSE loss function is employed to assess the pixel-level difference between the predicted map and the ground truth. It is formulated as,

$$\mathcal{L}_{gt} = \frac{1}{K} \sum_{i=1}^{K} \|F(I_i) - G_i\|_2^2\,, \tag{9}$$

where K stands for the batch size, $F(I_i)$ represents the predicted density map, and G_i denotes the corresponding ground truth density map.

4 Experimental results and analysis

4.1 Implementation Details

In the training stage, for the teacher model, we selected `OSNet` × `1` [66], while the student model uses `OSNet` × `0.5`. The architecture of these two models are shown in Table 1. For data augmentation, the samples are randomly cropped to 256 × 256 and then flipped horizontally. The batch size and epochs are 16 and 3,000, respectively. Optimization is performed using the Adam algorithm [28], with a learning rate set to 1e-3 and a weight decay of 0.0005. All the experiments are conducted on the same hardware with PyTorch [41] on an RTX 3080Ti GPU.

Table 1: The architecture of the two types of OSNet

Stage	Output size	Output channels	
		`OSNet` × `0.5`	`OSNet` × `1`
Layer1	128×128	32	64
Layer2	32×32	128	256
Layer3	16×16	192	384
Layer4	16×16	256	512
Layer5	16×16	256	512

4.2 Datasets

ShanghaiTech [65] is a commonly utilized dataset for crowd counting research. It comprises 1,198 images with a total of 330,165 annotated head locations. It is divided into two sections: Part_A and Part_B. The Part_A dataset includes 300 images for training and 182 images for testing. These images are collected from the internet and typically feature dense crowd distributions. In contrast, the Part_B dataset comprises 400 training images and 316 testing images, all captured directly on the streets of Shanghai, and predominantly show sparse crowd distributions.

UCF_CC_50 [25] is a critical resource for evaluating crowd-counting techniques, which consists of 50 images with diverse resolutions and a total of 63,075 annotated individuals. The number of individuals in the images shows significant variation, with counts ranging from 94 to 4,543. This variation demonstrates the substantial diversity of the dataset.

UCF-QNRF [26] is a challenging dataset that includes a range of scenes, various viewpoints, different lighting conditions, and variations in density. It consists of 1,535 images, with 1,201 allocated for training and 334 for testing. The dataset is characterized by high-resolution images, with an average resolution of $2,013 \times 2,902$ pixels. It includes authentic outdoor settings from various global locations, with elements such as structures, vegetation, skies, and roadways. These diverse characteristics make the dataset a crucial resource for evaluating the robustness and generalization capabilities of crowd-counting models under varying real-world conditions.

NWPU-Crowd [58] comprises 5,109 images. It includes 2,133,238 annotated entities, with an average resolution of $2,191 \times 3,209$. This dataset is notable for its inclusion of negative samples, which improve the robustness of models during training. It stands out due to its greater variety in scales, densities, and background complexities compared to other datasets. Furthermore, the dataset includes negative samples that do not contain crowd scenes, which enhances its diversity. This feature increases its value for the evaluation of crowd-counting algorithms.

CARPK [22] is sourced from various regions in Hong Kong, such as parking lots and streets, and contains images of vehicles captured under diverse scenes and lighting conditions. It comprises 1,448 drone-view images, with 989 used for training and 459 for testing.

PUCPR+ [22] includes a wide range of video sequences and images from various driving scenarios and perspectives. It consists of 125 images with a total of 16,456 annotations, of which 100 images are used for training and 25 are set aside for testing.

Large-Vehicle [13] is an openly available dataset focused on large vehicle detection, recognition, and tracking research. It comprises 172 remote sensing images, each with an average resolution of 1,552×1,573 pixels.

Small-Vehicle [13], another dataset for remote sensing vehicle counting, comprises 280 high-resolution images containing a total of 148,838 small vehicles. It exhibits a more extensive range of scale variation compared to the large vehicle dataset.

4.3 Evaluation Metrics

To assess the accuracy and robustness of the counting task, the Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) are used as evaluation metrics. They are formulated as,

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}|s_i - \hat{s}_i|, \tag{10}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(s_i - \hat{s}_i)^2}, \tag{11}$$

where N denotes the total number of images, s_i represents the ground truth count, and $\hat{s}_i$ corresponds to the predicted count for $i-$th image.

4.4 Efficiency Evaluation

To assess the efficiency of the proposed DHMoE method, we conducted a comparison with SOTA methods on four criteria: parameters (Params), floating point operations (FLOPs), inference time (Time), and frames per second (FPS). All experiments used a unified output resolution of 576×768 and were performed on an RTX3080Ti GPU. The results are summarized in Table 2.

As shown in Table 2, DHMoE achieves lower parameter and computation overheads than other methods. It also shows shorter inference times and significantly higher FPS, which demonstrates its superior performance across all metrics.

Table 2: Comparison of efficiency across different methods. The best results are marked in **bold**.

Methods	Params (M)↓	FLOPs (G)↓	Time (ms)↓	FPS↑
RAQNet [63]	42.77	250.86	36.24	27.59
SRRNet [18]	66.14	162.09	29.62	33.76
BL [38]	21.50	182.19	15.59	64.14
CSRNet [29]	16.26	182.69	15.08	66.33
ASPDNet [12]	22.70	256.19	35.43	28.22
SCAR [15]	16.29	182.86	16.55	60.41
SFCN [59]	38.60	274.06	40.27	24.83
SASNet [50]	38.90	393.16	45.57	21.94
DHMoE (Ours)	**3.68**	**48.93**	**5.70**	**175.31**

4.5 Accuracy and Robustness Evaluation

To assess the counting performance of the proposed DHMoE, we compared it with SOTA methods across four crowd datasets, as summarized in Table 3. Furthermore, to explore the generalization capability of DHMoE, cross-domain experiments were performed, along with evaluations on four vehicle datasets. The results of these experiments are presented from Table 4 to Table 7.

Table 3: The results on the crowd datasets. The table categorizes "P(M)" as the parameters. It is structured into two main sections: heavyweight networks are described in the upper portion, while lightweight networks are outlined in the lower portion. In the latter section, the best results are highlighted in **bold**, and those ranked second are marked with underlines.

	Methods	Part_A		Part_B		UCF_CC_50		UCF_QNRF		NWPU		P(M)
		MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	
Heavyweight	CSRNet [29]	68.2	115.0	10.6	16.0	266.1	397.5	135.4	207.4	121.3	387.8	16.26
	SFCN [59]	64.8	107.5	7.6	13.0	214.2	318.2	102.0	171.4	105.7	424.1	38.60
	CAN [35]	62.3	100.0	7.8	12.2	212.2	243.7	107.0	183.0	106.3	386.5	18.10
	BL [38]	61.5	103.2	7.5	12.6	229.3	308.2	87.7	158.1	105.4	454.2	21.50
	SRRNet [18]	60.8	103.0	7.4	13.6	172.9	256.3	89.5	162.9	-	-	66.14
	RAQNet [63]	59.0	101.2	9.0	15.4	177.1	247.6	106.5	186.1	-	-	42.77
	DLPTNet [6]	58.4	95.0	9.3	15.6	-	-	121.0	225.8	103.3	421.9	110.90
	UEPNet [53]	54.6	91.2	6.4	10.9	165.2	275.9	81.1	131.7	-	-	26.12
	STNet [56]	52.9	83.6	6.3	10.3	162.0	230.4	87.9	166.4	-	-	15.56
	PET [32]	49.3	78.8	6.2	9.7	-	-	79.5	144.3	74.4	328.5	20.90
	APGCC [4]	48.8	76.7	5.6	8.7	154.8	205.5	80.1	136.6	71.7	284.4	18.68
Lightweight	MCNN [65]	110.2	173.2	26.4	41.3	377.6	509.1	277.0	426.0	232.5	714.6	0.13
	TDF-CNN [45]	97.5	145.1	20.7	32.8	354.7	491.4	-	-	-	-	0.13
	LCNet [37]	93.3	149.0	15.3	25.2	326.7	430.6	-	-	-	-	0.86
	MobileCount [57]	89.4	146.0	**9.0**	15.4	284.8	392.8	131.1	222.6	-	-	3.40
	C-CNN [47]	88.1	141.7	14.9	22.1	-	-	-	-	-	-	0.07
	1/4SAN+SKT [33]	78.0	126.6	11.9	19.8	-	-	157.5	257.7	-	-	0.06
	SANet [3]	75.3	122.2	10.5	17.9	258.4	334.9	152.6	547.0	190.6	491.4	0.91
	PCCNet [14]	73.5	124.0	11.0	19.0	240.0	315.5	148.7	247.3	-	-	0.55
	GAPNet [19]	67.1	110.4	9.8	**15.2**	202.8	246.9	**118.5**	**217.2**	174.1	514.7	2.85
	DHMoE (Ours)	**59.2**	**96.1**	11.0	19.6	**112.4**	**197.8**	132.1	253.6	**118.0**	**481.1**	3.68

The results of the experiments conducted on four crowd datasets are summarized in Table 3. The upper section of the table lists the heavyweight networks, which generally achieve higher accuracy but require greater computational resources. The lower section contains the lightweight networks, which have fewer parameters but typically offer lower accuracy compared to the heavyweight networks.

In Shanghai Part_A dataset, DHMoE ranks first MAE and RMSE among lightweight networks, with values of 59.2 and 96.1, respectively. Compared to GAPNet [19], another lightweight counting method that ranks second, DHMoE decreases the MAE and RMSE by 11.77% and 12.95%, respectively. In comparison to SRRNet [18], which also tackles the scale variation problem, the proposed DHMoE reduces the MAE and RMSE by 2.63% and 6.7%, respectively. This verifies the superior performance of DHMoE in handling scale variation efficiently.

In Shanghai Part_B dataset, DHMoE achieves an MAE of 11.0 and an RMSE of 19.6. It is ranked fourth among lightweight networks. Still, there is a considerable difference when compared to the heavyweight models, such as CSRNet [29], which also addresses the scale variation problem. Specifically, DHMoE falls behind CSRNet [29] by 3.77% in MAE and 22.5% in RMSE. However, DHMoE only uses 22.63% of the parameters required by CSRNet [29], which indicates its efficiency in reducing computational resources. This demonstrates that DHMoE effectively reduces computational costs while retaining competitive accuracy, which offers distinct advantages for applications in settings with limited resources.

In UCF_CC_50 dataset, DHMoE leads the lightweight networks with an MAE of 112.4 and RMSE of 197.8. Even when compared to heavyweight networks, DHMoE demonstrates certain advantages. For instance, compared to APGCC [4], the top-performing heavyweight network listed in Table 3, DHMoE reduces MAE and RMSE by 27.39% and 3.75%, respectively, while using only 19.7% of its parameters. It demonstrates the superiority of DHMoE in handling high-density scenarios and addressing scale variation challenges.

In UCF_QNRF dataset, DHMoE ranks third among lightweight networks with an MAE of 132.1 and an RMSE of 253.6. Although DHMoE shows some improvement in lightweight models, it still lags behind heavyweight models. For example, compared to CSRNet [29], which also addresses scale variation, DHMoE increases RMSE by 18.22% but reduces MAE by 2.44%. However, the parameters of DHMoE reduces by 77.37%.

In NWPU dataset, DHMoE reaches an MAE of 118.0 and an RMSE of 481.1, which ranks it as the top lightweight network. Compared to SANet [3], a lightweight model ranked third that also tackles scale variation, the performance of MAE and RMSE are improved by 38.09% and 2.1%, respectively. However, compared to heavyweight networks, there are certain limitations. For instance, when compared to the heavyweight network CAN [35], which also tackles multi-scale problems, DHMoE increases the MAE and RMSE by 9.92% and 19.66%, respectively. It indicates that lightweight networks still face challenges in dealing with complex scenarios.

Fig. 5 illustrates the qualitative results of DHMoE across crowd datasets. The first row represents the input images, the second shows the ground truth density maps, and the third provides the predicted density maps produced by DHMoE. The results indicate that the predictions from DHMoE align closely with the ground truth. This demonstrates its ability to model crowd distribution accurately and achieve high counting accuracy.

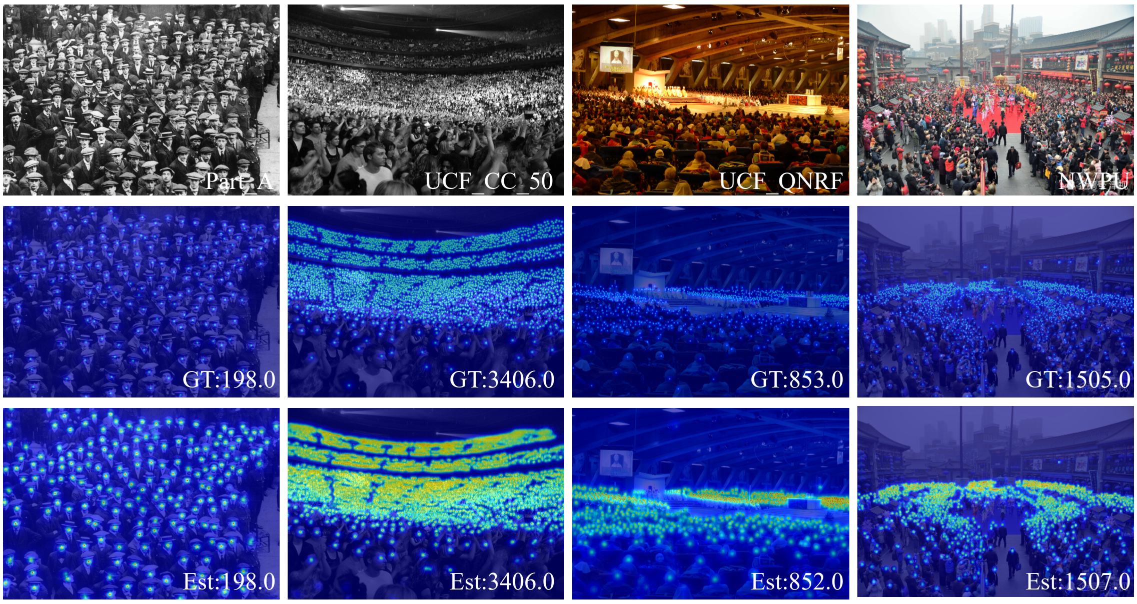

Fig. 5: Qualitative results from the crowd datasets. In the figure, “GT” represents Ground Truth, while “Est” indicates the estimated values.

Table 4: Comparison of various methods on the Large and Small vehicle datasets. The best results are highlighted in **bold**.

Method	Large vehicle		Small vehicle	
	MAE	RMSE	MAE	RMSE
SANet [3]	62.8	79.7	497.2	1276.7
CMTL [49]	61.0	78.3	490.5	1321.1
MCNN [65]	36.6	55.6	488.7	1317.4
SPN [8]	36.2	50.6	445.2	1252.9
CAN [35]	34.6	49.6	457.4	1260.4
CSRNet [29]	34.1	46.4	443.8	1252.2
SFCN [59]	33.9	49.7	440.7	1248.3
ASPDNet [12]	31.8	40.1	433.2	1238.6
SFANet [67]	29.0	47.0	435.3	1284.2
SRRNet [18]	18.3	31.2	122.8	419.7
DHMoE (Ours)	**11.8**	**23.8**	**104.1**	**417.9**

4.6 Generalization Analysis

To further evaluate the generalization ability of DHMoE, we conducted experimental analysis on four vehicle counting datasets: large vehicle, small vehicle, CARPK, and PUCPR+. Table 4 illustrates the performance comparison between DHMoE and SOTA methods for large and small vehicle datasets. The analysis demonstrates that DHMoE significantly outperforms the other methods listed. Specifically, DHMoE achieves the best MAE and RMSE on Large Vehicle dataset, with values of 11.8 and 23.8, respectively. Compared to SRRNet [18], the second-best model aimed at addressing scale variation, DHMoE improves the performance of MAE and RMSE by 35.5% and 23.7%. The small vehicle dataset primarily contains smaller vehicles. These vehicles often have smaller contours, lower heights, and shorter body lengths in images, which presents significant challenges for object counting. Despite the challenges, DHMoE still shows considerable performance benefits on the small vehicle dataset. DHMoE records an MAE of 104.1 and an RMSE of 417.9, which are superior to those of the other methods presented.

Table 5: Comparison results of different methods on the CARPK and PUCPR+ datasets. The best results are shown in **bold**, while the second-best are marked with underlines.

Method	CARPK		PUCPR+	
	MAE	RMSE	MAE	RMSE
Faster-RCNN [44]	103.5	110.6	156.8	200.6
YOLO [42]	102.9	110.0	156.7	200.5
One-look Regression [40]	59.5	66.8	21.9	36.7
LEP [51]	51.8	-	15.7	-
MCNN [65]	39.1	43.3	21.9	29.5
SSD [34]	37.3	42.3	119.2	132.2
LPN [23]	23.8	36.8	22.8	34.5
RetinaNet [31]	16.6	22.3	24.6	33.1
CSRNet [29]	11.5	13.3	8.7	29.5
SRRNet [18]	8.5	11.0	**2.0**	**2.8**
DHMoE (Ours)	**5.7**	**7.8**	2.4	3.3

Table 5 presents the experimental results of DHMoE on the CARPK and PUCPR+ datasets. In the CARPK dataset, DHMoE achieves the top performance with an MAE of 5.7 and an RMSE of 7.8. Compared to SFANet [67], which ranks third and also addresses the scale variation problem, DHMoE reduces MAE by 59.31% and RMSE by 49.36%. The results validate that the proposed DHMoE method effectively mitigates the challenges posed by scale variation and achieves an improvement in model accuracy.

In the PUCPR+ dataset, DHMoE is only behind SRRNet [18] which also addresses the scale variation challenge. The MAE and RMSE are decreased by 16.67% and 15.15%, respectively. However, DHMoE uses only 5.6% of the parameters that SRRNet [18] requires as shown in Table 2. When compared to CSRNet [29], the third-ranked model designed to handle scale variation, DHMoE decreases MAE and RMSE by 72.4% and 88.81%, respectively. Fig. 6 shows the qualitative outcomes of DHMoE on vehicle datasets. The results demonstrate a strong alignment between the predictions from DHMoE and the ground truth.

4.7 Statistical Analysis

The comparison of the baseline and the proposed DHMoE was conducted using the Part_A, PUCPR+, and Large vehicle datasets. The analysis included confidence interval evaluations of predicted values and ground truth, complemented by MAE boxplot visualizations. The confidence intervals are summarized in Table 6, and the boxplot visualizations in MAE are shown in Fig. 7.

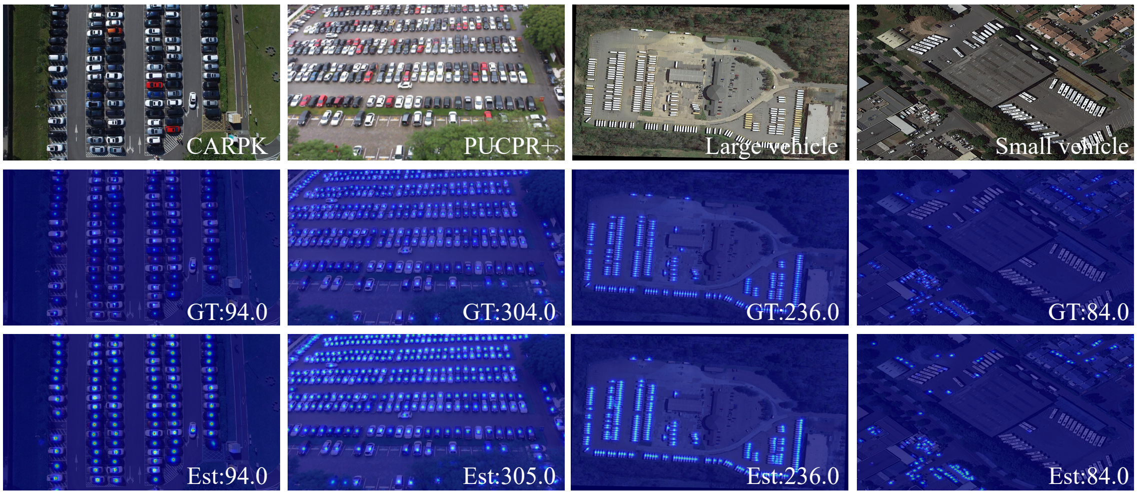

Fig. 6: The qualitative results on vehicle datasets. “GT” represents the Ground Truth, and “Est” refers to the corresponding estimated values.

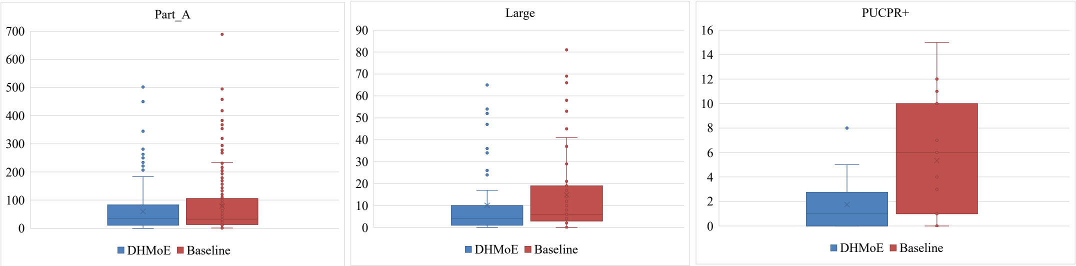

Fig. 7: Boxplots of the statistical results.

Table 6: Confidence Interval Comparison between Baseline and DHMoE

Confidence Interval	Part_A	Large vehicle	PUCPR+
Baseline	[62.52, 94.51]	[11.42, 23.00]	[3.58, 7.04]
DHMoE (Ours)	[48.98, 72.16]	[7.63, 17.80]	[1.04, 2.58]

On the Part_A and Large vehicle datasets, the confidence interval of the baseline method is [62.52, 94.51] for Part_A and [11.42, 23.00] for the Large vehicle. In comparison, the DHMoE achieves confidence intervals of [48.98, 72.16] and [7.63, 17.80] on the respective datasets. While some overlap exists in the confidence intervals of the two methods, DHMoE demonstrates a clear shift toward lower values and a narrower range. This reflects its superior stability, improved accuracy, and enhanced consistency in performance. Furthermore, the analysis of the boxplots for the Part_A and Large vehicle datasets reveals that the baseline method exhibits a wider error range and a higher number of outliers. In contrast, the DHMoE demonstrates a more concentrated

MAE distribution, with a significantly lower median and a marked reduction in outliers. These observations further validate that the DHMoE outperforms the baseline method in terms of error control and robustness.

On the PUCPR+ dataset, the confidence interval of the baseline method is [3.58, 7.04], while that of the DHMoE is [1.04, 2.58]. The complete non-overlap of the two confidence intervals indicates that the performance improvement of the proposed DHMoE on this dataset is statistically significant. Furthermore, the narrower confidence interval of the DHMoE demonstrates higher consistency and stability across multiple experiments. The boxplot analysis further highlights the significant differences in MAE distribution between the two methods. The baseline method exhibits greater variability and several high-value outliers. In contrast, the DHMoE shows a more concentrated MAE distribution within lower values, which underscores its superior accuracy and robustness. Based on confidence interval analysis and boxplot visualizations, the proposed DHMoE demonstrates significant advantages in both statistical significance and practical performance.

4.8 Cross-domain Analysis

Table 7: The experimental results of the cross-dataset evaluations are displayed; the best results are marked in **bold** and the secondary best results are marked in underline.

Methods	Part_A→QNRF		CARPK→ PUCPR+		PUCPR+→CARPK	
	MAE	RMSE	MAE	RMSE	MAE	RMSE
ASPDNet [12]	217.96	377.84	135.21	168.16	92.76	100.15
SASNet [50]	205.53	347.43	121.83	151.60	47.00	52.45
PSCGNet [11]	204.63	371.44	90.75	112.305	**33.34**	48.85
SRRNet [18]	167.47	**285.27**	115.89	144.08	48.10	57.30
DHMoE (Ours)	**164.59**	314.61	**84.33**	**103.09**	35.49	**40.54**

To assess the cross-domain capabilities of the proposed DHMoE, a cross-dataset evaluation was conducted, as shown in Table 7. Initially, the model was trained on the ShanghaiTech Part_A dataset and subsequently tested on the UCF-QNRF dataset. Next, the training was conducted on the CARPK dataset with subsequent evaluations on the PUCPR+ dataset. The process was reversed by training on the PUCPR+ dataset and testing on CARPK. For a thorough evaluation of DHMoE, it was benchmarked against popular methodologies including ASPDNet [12], SASNet [50], PSCGNet [11], and SRRNet [18]. As evident from Table 7, DHMoE achieves outstanding performance in terms of MAE and RMSE, which substantiates the exceptional generalization ability of the proposed approach.

4.9 Ablation Study

To assess the efficacy of the DHMoE, ablation studies were conducted on the Shanghai Part_A and CARPK datasets, with the results detailed in Table 8. On the Shanghai Part_A dataset, the student model achieved an MAE of 77.13 and an RMSE of 130.81. Compared to the teacher model, these

Table 8: Ablation Analysis on the ShanghaiTech Part_A and CARPK Datasets. The best results are marked in **bold**.

Methods	Params (M)	FLOPs (G)	Part_A		CARPK	
			MAE	RMSE	MAE	RMSE
Teacher	4.61	7.70	60.24	99.74	6.78	9.37
Student	1.19	2.22	77.13	130.81	7.04	9.57
Student + HMOE	1.27	2.27	69.73	112.70	6.75	9.71
Student + KD	3.62	7.24	70.32	116.41	6.07	8.42
DHMoE (Ours)	3.68	7.25	**59.20**	**96.10**	**5.7**	**7.8**

represent improvements of 28.04% in MAE and 31.15% in RMSE, while only requiring 25.81% of the teacher model's parameters.

The proposed DHMoE, despite having 20.17% fewer parameters compared to the teacher model, demonstrates improved performance. Specifically, on the Shanghai Part_A dataset, DHMoE achieves enhancements of 1.73% in MAE and 3.65% in RMSE. On the CARPK dataset, these improvements are even more pronounced, with increases of 15.93% in MAE and 16.76% in RMSE. Compared to the student model, DHMoE significantly enhances performance on the Shanghai Part_A dataset, with MAE and RMSE improvements of 23.25% and 26.53%, respectively. On the CARPK dataset, the increases are 19.03% for MAE and 18.5% for RMSE. Although this required an additional 2.49M parameters, it resulted in significant performance enhancements for the student model. Furthermore, each component of DHMoE was independently verified through experiments, which confirmed that every element contributes to the performance of the student model positively.

The ablation studies confirm that the DHMoE delivers superior performance on both datasets. This proves that the proposed DHMoE effectively enhances the performance of the student model while maintaining a lightweight framework.

5 Conclusion

In this paper, we proposed DHMoE to address the limitations of lightweight networks in handling multi-scale variations and complex crowd counting tasks. DHMoE employs knowledge distillation to convey fine-grained knowledge from the teacher model to the student model. Additionally, the student model incorporates an HMoE structure, with four expert networks being integrated. Each expert network is designed to address the challenges posed by scale variations. In addition, it allows the student model to acquire the capability from the teacher model to manage complex scenarios, which enhances its performance and counting precision in diverse and complex settings. Experimental results prove that DHMoE demonstrates competitive counting performance across four different crowd datasets and four vehicle datasets, and it is lightweight model that is potential for application on mobile edge devices.

References

1. Bishop, C.M., Nasrabadi, N.M.: Pattern recognition and machine learning, vol. 4. Springer (2006)
2. Cai, W., Jiang, J., Wang, F., Tang, J., Kim, S., Huang, J.: A survey on mixture of experts. arXiv preprint arXiv:2407.06204 (2024)
3. Cao, X., Wang, Z., Zhao, Y., Su, F.: Scale aggregation network for accurate and efficient crowd counting. In: Proceedings of the European conference on computer vision (ECCV), pp. 734–750 (2018)
4. Chen, I., Chen, W.T., Liu, Y.W., Yang, M.H., Kuo, S.Y., et al.: Improving point-based crowd counting and localization based on auxiliary point guidance. arXiv preprint arXiv:2405.10589 (2024)
5. Chen, J., Gao, M., Guo, X., Zhai, W., Li, Q., Jeon, G.: Object counting in remote sensing via selective spatial-frequency pyramid network. Software: Practice and Experience **54**(9), 1754–1773 (2024)
6. Chen, J., Gao, M., Li, Q., Guo, X., Wang, J., Xing, X., et al.: Privacy-aware crowd counting by decentralized learning with parallel transformers. Internet of Things **26**, 101167 (2024)
7. Chen, J., Li, Q., Gao, M., Zhai, W., Jeon, G., Camacho, D.: Towards zero-shot object counting via deep spatial prior cross-modality fusion. Information Fusion p. 102537 (2024)
8. Chen, X., Bin, Y., Sang, N., Gao, C.: Scale pyramid network for crowd counting. In: 2019 IEEE winter conference on applications of computer vision (WACV), pp. 1941–1950. IEEE (2019)
9. Du, Z., Shi, M., Deng, J., Zafeiriou, S.: Redesigning multi-scale neural network for crowd counting. IEEE Transactions on Image Processing (2023)
10. Fedus, W., Zoph, B., Shazeer, N.: Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. Journal of Machine Learning Research **23**(120), 1–39 (2022)
11. Gao, G., Liu, Q., Hu, Z., Li, L., Wen, Q., Wang, Y.: Psgcnet: A pyramidal scale and global context guided network for dense object counting in remote-sensing images. IEEE Transactions on Geoscience and Remote Sensing **60**, 1–12 (2022)
12. Gao, G., Liu, Q., Wang, Y.: Counting from sky: A large-scale data set for remote sensing object counting and a benchmark method. IEEE Transactions on geoscience and remote sensing **59**(5), 3642–3655 (2020)
13. Gao, G., Liu, Q., Wang, Y.: Counting from sky: A large-scale data set for remote sensing object counting and a benchmark method. IEEE Transactions on Geoscience and Remote Sensing **59**(5), 3642–3655 (2020)
14. Gao, J., Wang, Q., Li, X.: Pcc net: Perspective crowd counting via spatial convolutional network. IEEE Transactions on Circuits and Systems for Video Technology **30**(10), 3486–3498 (2019)
15. Gao, J., Wang, Q., Yuan, Y.: Scar: Spatial-/channel-wise attention regression networks for crowd counting. Neurocomputing **363**, 1–8 (2019)
16. Gou, J., Yu, B., Maybank, S.J., Tao, D.: Knowledge distillation: A survey. International Journal of Computer Vision **129**(6), 1789–1819 (2021)
17. Guo, X., Gao, M., Pan, J., Shang, J., Souri, A., Ql, L., Bruno, A., et al.: Crowd counting via attention and multi-feature fused network. HUMAN-CENTRIC COMPUTING AND INFORMATION SCIENCES **13** (2023)

18. Guo, X., Gao, M., Zhai, W., Li, Q., Jeon, G.: Scale region recognition network for object counting in intelligent transportation system. IEEE Transactions on Intelligent Transportation Systems (2023)
19. Guo, X., Song, K., Gao, M., Zhai, W., Li, Q., Jeon, G.: Crowd counting in smart city via lightweight ghost attention pyramid network. Future Generation Computer Systems **147**, 328–338 (2023)
20. Han, K., Wang, Y., Tian, Q., Guo, J., Xu, C., Xu, C.: Ghostnet: More features from cheap operations. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 1580–1589 (2020)
21. Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., Adam, H.: Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861 (2017)
22. Hsieh, M.R., Lin, Y.L., Hsu, W.H.: Drone-based object counting by spatially regularized regional proposal network. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV),, pp. 4165–4173 (2017). DOI https://doi.org/10.1109/ICCV.2017.446
23. Hsieh, M.R., Lin, Y.L., Hsu, W.H.: Drone-based object counting by spatially regularized regional proposal network. In: Proceedings of the IEEE international conference on computer vision, pp. 4145–4153 (2017)
24. Huang, Z., Sinnott, R.O.: Improved knowledge distillation for crowd counting on iot devices. In: 2023 IEEE International Conference on Edge Computing and Communications (EDGE), pp. 207–214. IEEE (2023)
25. Idrees, H., Saleemi, I., Seibert, C., Shah, M.: 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 (2013)
26. Idrees, H., Tayyab, M., Athrey, K., Zhang, D., Al-Maadeed, S., Rajpoot, N., Shah, M.: Composition loss for counting, density map estimation and localization in dense crowds. In: Proceedings of the European conference on computer vision (ECCV), pp. 532–546 (2018)
27. Jiao, X., Yin, Y., Shang, L., Jiang, X., Chen, X., Li, L., Wang, F., Liu, Q.: Tinybert: Distilling bert for natural language understanding. arXiv preprint arXiv:1909.10351 (2019)
28. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
29. Li, Y., Zhang, X., Chen, D.: 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 (2018)
30. Liang, D., Xu, W., Zhu, Y., Zhou, Y.: Focal inverse distance transform maps for crowd localization. IEEE Transactions on Multimedia (2022)
31. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision, pp. 2980–2988 (2017)
32. Liu, C., Lu, H., Cao, Z., Liu, T.: Point-query quadtree for crowd counting, localization, and more. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 1676–1685 (2023)
33. Liu, L., Chen, J., Wu, H., Chen, T., Li, G., Lin, L.: Efficient crowd counting via structured knowledge transfer. In: Proceedings of the 28th ACM international conference on multimedia, pp. 2645–2654 (2020)
34. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., Berg, A.C.: Ssd: Single shot multibox detector. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam,

The Netherlands, October 11–14, 2016, Proceedings, Part I 14, pp. 21–37. Springer (2016)

35. Liu, W., Salzmann, M., Fua, P.: Context-aware crowd counting. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 5099–5108 (2019)
36. Ma, N., Zhang, X., Zheng, H.T., Sun, J.: Shufflenet v2: Practical guidelines for efficient cnn architecture design. In: Proceedings of the European conference on computer vision (ECCV), pp. 116–131 (2018)
37. Ma, X., Du, S., Liu, Y.: A lightweight neural network for crowd analysis of images with congested scenes. In: 2019 IEEE International Conference on Image Processing (ICIP), pp. 979–983. IEEE (2019)
38. Ma, Z., Wei, X., Hong, X., Gong, Y.: Bayesian loss for crowd count estimation with point supervision. In: Proceedings of the IEEE/CVF international conference on computer vision, pp. 6142–6151 (2019)
39. Masoudnia, S., Ebrahimpour, R.: Mixture of experts: a literature survey. Artificial Intelligence Review **42**, 275–293 (2014)
40. Mundhenk, T.N., Konjevod, G., Sakla, W.A., Boakye, K.: A large contextual dataset for classification, detection and counting of cars with deep learning. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14, pp. 785–800. Springer (2016)
41. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems **32** (2019)
42. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 779–788 (2016)
43. Reisser, M., Louizos, C., Gavves, E., Welling, M.: Federated mixture of experts. arXiv preprint arXiv:2107.06724 (2021)
44. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems **28** (2015)
45. Sam, D.B., Babu, R.V.: Top-down feedback for crowd counting convolutional neural network. In: Proceedings of the AAAI conference on artificial intelligence, vol. 32 (2018)
46. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.C.: Mobilenetv2: Inverted residuals and linear bottlenecks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4510–4520 (2018)
47. Shi, X., Li, X., Wu, C., Kong, S., Yang, J., He, L.: A real-time deep network for crowd counting. In: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2328–2332. IEEE (2020)
48. Shuvo, M.M.H., Islam, S.K., Cheng, J., Morshed, B.I.: Efficient acceleration of deep learning inference on resource-constrained edge devices: A review. Proceedings of the IEEE **111**(1), 42–91 (2022)
49. Sindagi, V.A., Patel, V.M.: Cnn-based cascaded multi-task learning of high-level prior and density estimation for crowd counting. In: 2017 14th IEEE international conference on advanced video and signal based surveillance (AVSS), pp. 1–6. IEEE (2017)
50. Song, Q., Wang, C., Wang, Y., Tai, Y., Wang, C., Li, J., Wu, J., Ma, J.: To choose or to fuse? scale selection for crowd counting. In: Proceedings of the AAAI conference on artificial intelligence, vol. 35, pp. 2576–2583 (2021)

51. Stahl, T., Pintea, S.L., Van Gemert, J.C.: Divide and count: Generic object counting by image divisions. IEEE Transactions on Image Processing **28**(2), 1035–1044 (2018)
52. Tian, Y., Krishnan, D., Isola, P.: Contrastive representation distillation. arXiv preprint arXiv:1910.10699 (2019)
53. Wang, C., Song, Q., Zhang, B., Wang, Y., Tai, Y., Hu, X., Wang, C., Li, J., Ma, J., Wu, Y.: Uniformity in heterogeneity: Diving deep into count interval partition for crowd counting. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 3234–3242 (2021)
54. Wang, C.H., Huang, K.Y., Yao, Y., Chen, J.C., Shuai, H.H., Cheng, W.H.: Lightweight deep learning: An overview. IEEE Consumer Electronics Magazine (2022)
55. Wang, J., Guo, X., Li, Q., Abdelmoniem, A.M., Gao, M.: Sdanet: scale-deformation awareness network for crowd counting. Journal of Electronic Imaging **33**(4), 043002–043002 (2024)
56. Wang, M., Cai, H., Han, X., Zhou, J., Gong, M.: Stnet: Scale tree network with multi-level auxiliator for crowd counting. IEEE Transactions on Multimedia (2022)
57. Wang, P., Gao, C., Wang, Y., Li, H., Gao, Y.: Mobilecount: An efficient encoder-decoder framework for real-time crowd counting. Neurocomputing **407**, 292–299 (2020)
58. Wang, Q., Gao, J., Lin, W., Li, X.: Nwpu-crowd: A large-scale benchmark for crowd counting and localization. IEEE transactions on pattern analysis and machine intelligence **43**(6), 2141–2149 (2020)
59. Wang, Q., Gao, J., Lin, W., Yuan, Y.: Pixel-wise crowd understanding via synthetic data. International Journal of Computer Vision **129**(1), 225–245 (2021)
60. Wang, S., Li, Y., Li, H., Zhu, T., Li, Z., Ou, W.: Multi-task learning with calibrated mixture of insightful experts. In: 2022 IEEE 38th International Conference on Data Engineering (ICDE), pp. 3307–3319. IEEE (2022)
61. Zhai, W., Gao, M., Guo, X., Li, Q.: Scale-context perceptive network for crowd counting and localization in smart city system. IEEE Internet of Things Journal pp. 1–1 (2023). DOI 10.1109/JIOT.2023.3268226
62. Zhai, W., Gao, M., Souri, A., Li, Q., Guo, X., Shang, J., Zou, G.: An attentive hierarchy convnet for crowd counting in smart city. Cluster Computing **26**(2), 1099–1111 (2023)
63. Zhai, W., Xing, X., Jeon, G.: Region-aware quantum network for crowd counting. IEEE Transactions on Consumer Electronics (2024)
64. Zhang, X., Zhou, X., Lin, M., Sun, J.: Shufflenet: An extremely efficient convolutional neural network for mobile devices. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 6848–6856 (2018)
65. Zhang, Y., Zhou, D., Chen, S., Gao, S., Ma, Y.: 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 (2016)
66. Zhou, K., Yang, Y., Cavallaro, A., Xiang, T.: Learning generalisable omni-scale representations for person re-identification. IEEE transactions on pattern analysis and machine intelligence **44**(9), 5056–5069 (2021)
67. Zhu, L., Zhao, Z., Lu, C., Lin, Y., Peng, Y., Yao, T.: Dual path multi-scale fusion networks with attention for crowd counting. arXiv preprint arXiv:1902.01115 (2019)