Vehicle Counting Network with Attention-based Mask Refinement and Spatial-awareness Block Loss

Ji Zhang, Jian-Jun Qiao, Xiao Wu*, Wei Li
Southwest Jiaotong University
Chengdu, China
{jizhang901,qjjai56}@gmail.com,{wuxiaohk,liwei}@swjtu.edu.cn

ABSTRACT

Vehicle counting aims to calculate the number of vehicles in congested traffic scenes. Although object detection and crowd counting have made tremendous progress with the development of deep learning, vehicle counting remains a challenging task, due to scale variations, viewpoint changes, inconsistent location distributions, diverse visual appearances and severe occlusions. In this paper, a well-designed Vehicle Counting Network (VCNet) is novelly proposed to overcome the problem of scale variation and inconsistent spatial distribution in congested traffic scenes. Specifically, VCNet is composed of two major components: (i) To capture multi-scale vehicles across different types and camera viewpoints, an effective multi-scale density map estimation structure is designed by building an attention-based mask refinement module. The multi-branch structure with hybrid dilated convolution blocks is proposed to assign receptive fields to generate multi-scale density maps. To efficiently aggregate multi-scale density maps, the attention-based mask refinement is well-designed to highlight the vehicle regions, which enables each branch to suppress the scale interference from other branches. (ii) In order to capture the inconsistent spatial distributions, a spatial-awareness block loss (SBL) based on the region-weighted reward strategy is proposed to calculate the loss of different spatial regions including sparse, congested and occluded regions independently by dividing the density map into different regions. Extensive experiments conducted on three benchmark datasets, TRANCOS, VisDrone2019 Vehicle and CVCSet demonstrate that the proposed VCNet outperforms the state-of-the-art approaches in vehicle counting. Moreover, the proposed idea can be applicable for crowd counting, which produces competitive results on ShanghaiTech crowd counting dataset.

CCS CONCEPTS

• **Computing methodologies** → **Computer vision problems**.

*Xiao Wu is the corresponding author.

MM '21, October 20–24, 2021, Virtual Event, China

ACM ISBN 978-1-4503-8651-7/21/10…$15.00
https://doi.org/10.1145/3474085.3475246

KEYWORDS

Deep learning; Vehicle counting; Multi-scale density map; Attention-based mask refinement; Spatial-awareness block loss

ACM Reference Format:
Ji Zhang, Jian-Jun Qiao, Xiao Wu*, Wei Li. 2021. Vehicle Counting Network with Attention-based Mask Refinement and Spatial-awareness Block Loss. In *Proceedings of the 29th ACM International Conference on Multimedia (MM '21), October 20–24, 2021, Virtual Event, China.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3474085.3475246

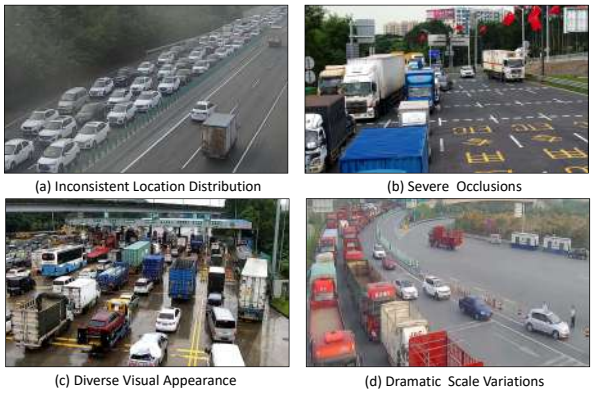

Figure 1: Vehicle counting in congested traffic scenes remains a challenging task.

1 INTRODUCTION

The goal of vehicle counting is to calculate the number of vehicles in congested traffic scenes, which has a wide variety of real-world applications, including traffic surveillance, congestion monitoring, urban planning, intelligent transportation system, smart city and so on. Although object detection has achieved tremendous progress, vehicle counting remains a challenging task due to inconsistent location distributions, diverse visual appearances, severe occlusions, dramatic scale variations and various sizes, which is illustrated in Fig. 1.

In recent years, crowd counting has been extensively studied (e.g., [1, 4, 5, 8, 19, 37, 39, 50, 52]), which is to count the number of people in an image, ranging from tens to thousands. With the rapid development of deep convolutional neural networks, promising performance has been achieved by leveraging the supervision information of center positions provided by pedestrians' heads. Although crowd counting and vehicle counting have some properties in common, vehicle counting demonstrates three significantly different but challenging aspects compared with crowd counting.

- **Severe scale variations**: One critical factor among all issues is the drastic scale variations in congested vehicle scenes. In crowd counting, the heads of pedestrians tend to have similar physical sizes, circle-like shapes and appearances. The scale variation is mainly caused by the distance between the cameras and people. Differently, the scale variation in vehicle counting is more severe than crowd counting. It is caused due to the perspectives, different types of vehicles as well as the distance between the cameras and vehicles.
- **Diverse visual appearance**: The visual appearances of human heads are more or less similar for crowd counting. On the contrary, large appearance variations exists in vehicle counting. Due to the viewpoint changes and the wide variety of vehicle types, the visual appearances of vehicles vary dramatically in terms of colors, shapes and sizes.
- **Inconsistent spatial distributions**: In general, dense location distribution is a common issue for both crowd counting and vehicle counting. However, in congested traffic scenes, the sparsity of vehicles varies more dramatically as the roads change. For example, there is a significant difference in the spatial distributions between the vehicles at the toll stations and those on the roads. Differences in vehicle types also lead to more inconsistent spatial distributions.

Therefore, existing methods for crowd counting cannot be directly applied to vehicle counting because of the significant scale and spatial differences between the crowd and dense vehicles. First, some CNN-based methods with fixed receptive fields usually only perceive the objects with a certain scale range, which makes them relatively less robust to severe scale variations of different vehicles. Second, the strategy of multi-branch density map generation has widely used to handle the scale variations. However, the scale ranges of vehicles among multiple branches tend to overlap each other. Simple accumulate or average multi-branch density maps will lead to count redundancy and produce a suboptimal solution. Moreover, the visual appearance differences of vehicles also enlarge the imbalance of spatial distributions, which makes existing methods biased in the learning process. In dense distribution area, the estimation of density map tends to be inaccurate, while in sparse distribution area, the isolated points are easily be ignored.

To explore the aforementioned problems, in this paper, a vehicle counting network (VCNet) is proposed to alleviate the problem of drastic scale variations and inconsistent spatial distributions in congested traffic scenes, in which a high-quality density map containing different scale information is generated and a loss function perceiving a variety of spatial distributions is adopted to boost the performance. The framework of the proposed method is illustrated in Fig. 2, in which the RGB images are imported into the network to generate the density maps so that the density estimation can be conducted. In order to cope with the problem of severe scale variations, an effective strategy of attention-refined multi-scale density map estimation is introduced to capture vehicles with multi-level scales, which integrates a multi-branch structure with hybrid dilated convolution blocks and an attention-based mask refinement module. By suppressing the information redundancy of different branches, the attention-based mask refinement is employed to highlight the vehicle regions corresponding to each scale. Furthermore, to deal with the inconsistent spatial distributions, a spatial-awareness block loss is employed to enhance the model's perception of different spatial distributions, by dividing the density map into blocks and establishing block-level constraints. Instead of using the L_2 loss to measure each block, an effective region-weighted reward strategy of the loss is proposed to balance the losses of different blocks. Finally, an ideal density map is precisely predicted by integrating multiple branches. The main contributions of this paper are listed as follows:

- An end-to-end Vehicle Counting Network (VCNet) is novelly proposed, which explores the characteristics of severe scale variations and inconsistent spatial distributions in congested traffic scenes, with the elaborated strategies of attention-refined multi-scale density map estimation and the spatial-awareness block loss.
- An effective Attention-refined Multi-scale Density map Estimation module (AMDE) is well-designed to handle severe scale variations among different vehicles, where a new multi-branch hybrid dilated convolution structure is employed to generate multi-level density maps and an attention-based mask refinement is used to suppress the information redundancy of different branches.
- Motivated by the region-weighted reward strategy, a novel Spatial-awareness Block Loss, named SBL, is introduced to capture the spatial distributions and establish block-level constraints to measure each block of the density map.
- A new Congested Vehicle Counting dataset covering diverse traffic scenes, namely CVCSet is collected, which consists of 3,885 images and 132,524 annotated vehicles.
- Extensive experiments conducted on three benchmark datasets, TRANCOS [9], VisDrone2019 Vehicle [1, 53] and CVCSet, demonstrate that the proposed VCNet achieves promising performance, which outperforms the state-of-the-art methods in vehicle counting.

This paper is organized as follows. Section 2 gives a brief overview of related works. Sections 3 elaborates the proposed vehicle counting network. The experimental setting and performance comparison are presented in Section 4. Finally, this paper is concluded with a summary.

2 RELATED WORK

2.1 Crowd Counting

The methods of crowd counting can be classified as detection-based [7, 8, 19, 42], regression-based [3, 31, 50] and CNN-based approaches [4, 5, 16, 36, 39, 52]. The detection-based methods adopt object detectors to identify and locate each people in an image, with which the crowd can be counted. The features for pedestrians can be traditional local features like SIFT, or the latest deep features [30]. Unfortunately, the detection-based methods are not competent for dense crowd scenes with hundreds to thousands of people [5]. The regression-based methods first map an image to a density map and then integrate it to obtain the final count, in which the integral of the density map over the entire image is the number of the objects in the image. Nevertheless, the regression-based methods lose the ability of location [14] and have difficulties in preserving the high-frequency variation of the density map [5]. Inspired by

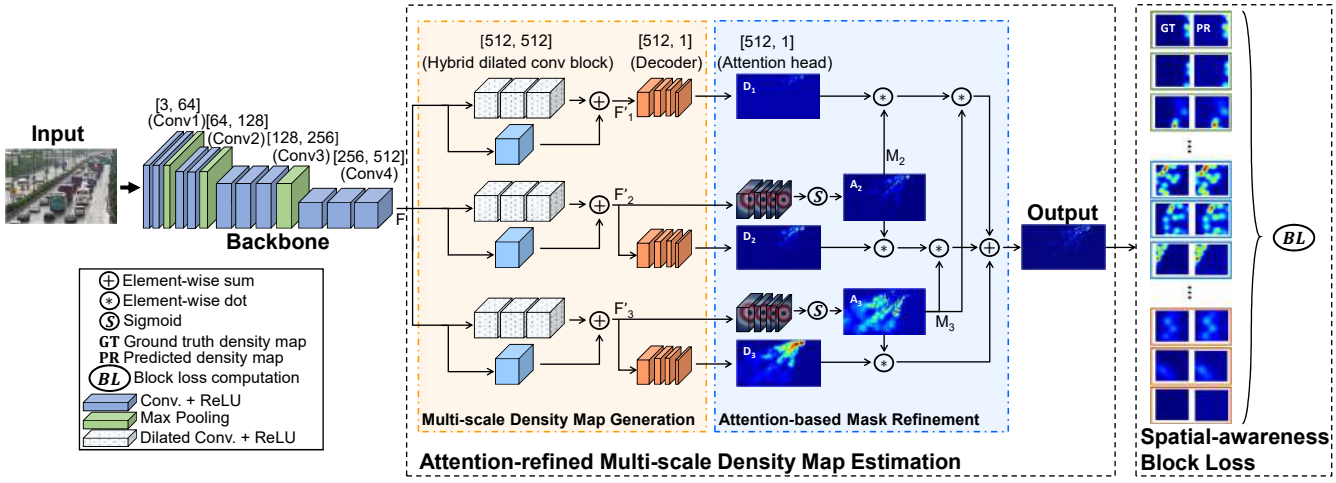

Figure 2: The framework of the proposed VCNet. Attention-refined multi-scale density maps estimation module is first utilized to generate high-quality density maps based on the multi-branch structure and attention-based mask refinement. The spatial-awareness block loss (SBL) accepts the high-quality density maps for loss computation.

the great success of CNN in vision tasks of image classification and recognition, CNN-based density estimate methods [52] for dense crowd counting are proposed and achieve great success. CNN-based methods tend to generate high-quality density map via the multi-column structure [52], multi-scale structure [37], or attention module [48, 49], etc. Therefore, high-precision crowd counting is achieved.

Overall, great progress is obtained in crowd counting with the latest methods. However, these methods are proposed for crowding counting, and vehicle counting of congested traffic scenes is a challenging task for them. Therefore, in this paper, the problem of density inconsistent vehicle counting is studied and a more suitable approach is proposed to address this problem.

2.2 Detection-based Vehicle Counting

With the progress of research on general object detection methods, more and more researchers turn their attention to special object detection. Most existing methods [12, 13, 20, 45] regard vehicle detection as a specific form of general object detection. In SINet [13], a context-aware ROI pooling associated with a multi-branch detection network is proposed to maintain contextual information by reducing the inter-class distance between features. In order to solve the problem of vehicle detection in aerial images, ClusDet [45] is proposed to unify object clustering and detection in an end-to-end framework. Although recent vehicle detection methods show good performance in terms of accuracy and speed, vehicle detection and counting in dense scenes is still a challenging problem. A Layout Proposal Network (LPN) [12] is proposed to count cars in UAV-based images. In order to deal with dense vehicles, a counting-driven attention network [21] is proposed to integrate counting information and vehicle detection into a whole framework. However, during training, detection-based vehicle counting methods often require bounding box information, which is difficult to label in dense scenes. Therefore, an estimated-based vehicle counting method is to be explored in this paper, which only requires point labeling information and has short training time as well as fast inference speed.

2.3 Estimation-based Vehicle Counting

CSRNet [22] is proposed for crowd counting and achieves high-performance with the dilated convolution, which expands the receptive field and preserves the resolution. This method is evaluated on both the crowd counting datasets like ShanghaiTech and a vehicle counting dataset called TRANCOS. In ADCrowdNet, an attention-injective network is designed to detect crowd regions and compute congestion priors. A multi-scale deformable network is then utilized to generate high-quality density maps based on the detected crowd regions and congestion priors. In [1], the VisDrone2019 Vehicle dataset is reorganized for density estimation with the central point of the bounding box adopted as point supervision. The work in [1] addresses the problems of scale variations and isolated clusters in vehicle counting, by extracting long-range contextual information and attaining high scale sensitivity of isolated clusters.

Moreover, in TRANCOS dataset, most of the vehicles are similar in sizes, shapes and colors. Besides, the dense area and complex background are segmented from the image during the training and testing process [22]. The reorganized VisDrone2019 Vehicle [1] used for counting contains small vehicles and multi-scale vehicles, which is more challenging than TRANCOS. However, it has few vehicle occlusions and highly dense areas. In this paper, a new vehicle counting dataset is collected, which has highly dense distributions, severe occlusions, inconsistent scales and various appearances of vehicles.

3 VEHICLE COUNTING NETWORK

3.1 Framework

To overcome the challenges induced by drastic scale variations and inconsistent spatial distributions, a novelly designed vehicle counting network (VCNet) is proposed for vehicle counting. It mainly

consists of two components, the attention-refined multi-scale density map estimation and the spatial-awareness block loss, which is illustrated in Fig. 2. A multi-branch structure with hybrid dilated convolution blocks is introduced, so that the receptive fields of VCNet can cover diverse vehicles with inconsistent scales, in which each branch corresponds to a specific range of scales. Through the structure of multi-branch hybrid dilated convolution blocks, an RGB image can be used as input to generate multi-scale density maps. The overlap of the receptive field of different branches often leads to the redundancy of information among the multi-scale density maps. An effective strategy of generating attention-refined mask is then introduced to perform the inner product with the density map of the corresponding branch. With the attention-based mask refinement, the density map of vehicles can be selectively assigned to each branch according to the corresponding scale ranges. A novel spatial-awareness block loss (SBL) is finally proposed to capture the inconsistent spatial distributions with the region-weighted reward strategy. The predicted density map is first divided into M blocks and the K_{th} *Root* of L_2 loss of each block is calculated separately. In this way, the contribution of each block to final loss is transformed, which captures special and complicated spatial information of the congested traffic scenes and enhances the spatial awareness of the network.

3.2 Density Map Generation

Given N training images $(X_i, P_i^{gt})_{i=1}^{N}$, with X_i being the i_{th} input image and P_i^{gt} being the annotated center points of vehicles in X_i. Following crowd counting, the ground-truth density map for each pixel $p \in X_i$ is generated as follows,

$$D^{gt}(p) = \sum_{P \in P_i^{gt}} \mathcal{N}^{gt}(p; \mu = P, \delta^2) \tag{1}$$

where $\mathcal{N}^{gt}(p; \mu = P, \delta^2)$ is a normalized Gaussian distribution with mean μ and covariance δ^2, and P denotes a single point annotation of P_i^{gt}. In the training stage, the goal is to learn a elaborated vehicle-counting network by using the density maps of vehicles. Similar to previous works [1, 10, 22], a simplified VGG16 network with the first 10 layers utilized for feature encoding, which acts as the backbone.

3.3 Attention-refined Multi-scale Density Map Estimation

Compared with the scale variations of the human heads, the changes of the vehicles are more serious due to different vehicle types and viewpoints. Severe scale variations pose a great challenge for vehicle counting. Most existing neural networks with specific receptive field tend to more accurately perceive objects within the corresponding size range, so it is difficult to deal with the problem of dramatic variations in the object sizes. Although the pooling operation is directly used to expand the receptive fields for high-level visual tasks, simple pooling for severe scale variations will lose critical information of smaller vehicles. Therefore, it is necessary to design a network with multiple receptive fields in vehicle counting task.

Motivated by the aforementioned idea, a multi-branch structure with hybrid dilated convolution blocks is designed to enable the network to more accurately identify vehicles of different scales and learn multi-scale features. Inspired by CSRNet [22], the dilated convolution blocks are used to construct multiple branches corresponding to different receptive fields, as shown in Fig. 2. Different from CSRNet, three dilated rates are first used to perform hybrid convolution operation to suppress the gridding issues caused by the same dilated rate [43]. A decoder structure composed of multi-layer convolution is then followed the block to predict density map of the corresponding scale. In order to reduce parameters, the weights of the decoder are shared by multiple branches. Finally, the hybrid dilated convolution blocks are connected in parallel to establish a complete multi-branch structure.

Specifically, the feature F from the *Conv4* of VGG16 is first fed to the structure of multi-branch hybrid dilated convolution blocks to generate multi-scale features. After a 1x1 convolution, the feature F is added to multi-scale features to obtain the features F_i' of the i_{th} branche. Finally, the features F_i' are directly used to generate the multi-scale density map D_i through the decoder.

The density map of the corresponding scale can be predicted by using the structure of multi-branch hybrid dilated convolution blocks, where the branch with larger dilation rates tends to perceive the large vehicles, and vice versa. Unfortunately, if the density maps from all branches are simply fused, the result may include repeated counts, which is mainly due to the overlap of the perception ranges between different branches in an unsupervised manner. The next question is how to eliminate the perceptual overlaps from different branches and enhance the vehicle information. Inspired by the attention mechanisms [44], an effective strategy of generating attention-refined mask is introduced to suppress redundant information among multi-scale density maps and guide each branch to be assigned in a specific perceptual range. Attention maps are generated to focus on the salient areas of the corresponding density maps, except for the branch with the smallest receptive field. Mask maps induced from attention maps are then applied to refine the density maps with smaller receptive fields. This is because the density maps from the branches with larger receptive fields contain more contextual information and have higher priority than those from the branches with smaller receptive fields. Finally, the refined multi-scale density maps of different branches are aggregated into a high-quality density map.

In VCNet, an attention head, including multi-layer convolution and an activation function, is utilized to accept the multi-scale features $\{F_i', i \in \{2, ..., n\}\}$ to generate the corresponding attention maps A_i. Similar to the decoder, the weights of the attention head are shared among multiple branches. Futhermore, the mask maps can be calculated as follows:

$$M_i = 1 - A_i, i \in 2, ..., n \tag{2}$$

The final predicted density map P is calculated by multiplying multi-scale density maps D_i, attention maps A_i and mask maps M_i, which can be formulated as follows,

$$D = \sum_{i=2}^{N} \prod_{j>i}^{N} M_j \otimes A_i \otimes D_i + \prod_{j>1}^{N} M_j \otimes D_1 \tag{3}$$

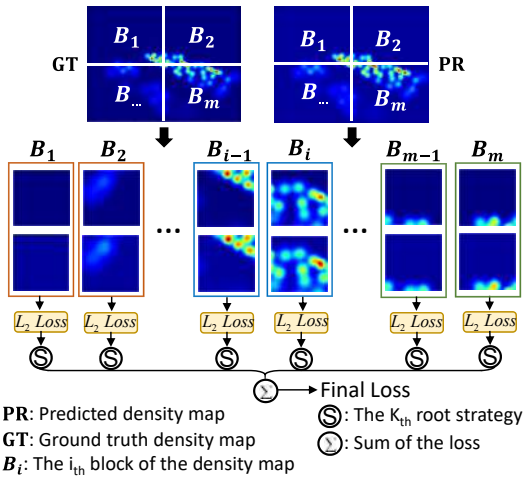

Figure 3: The description of the proposed spatial-awareness loss (SBL). Both the ground-truth density map and the predicted density map are first divided into M blocks and the losses of different blocks are calculated by L_2 loss function. The losses of M blocks are then processed by the K_{th} root strategy to balance the impact of different blocks on the final loss. The final loss is generated by summing the losses of M blocks.

where $\otimes$ denotes dot multiplication and $\sum$ refers to element-wise addition.

3.4 Spatial-awareness Block Loss

Although the attention-refined multi-scale density map estimation module can cope with the scale variation of dense vehicles, the spatial distributions in vehicle counting are more complex than the crowd due to the viewpoint changes and the wide variety of vehicle types. On one hand, there exist dense and sparse areas of vehicles in congested traffic scenes, and severe occlusions are pretty common. On the other hand, the commonly used L_2 loss has difficulty in learning spatial awareness of different regions [5]. There is still a gap between the inconsistent spatial distributions and the L_2 loss constrains. To address this problem, the most intuitive idea is to integrate block-level constraints by dividing the density map into different regions. However, for vehicle counting, the losses of congested and occluded regions usually tend to be large, while the losses of sparse areas are small. The contributions of different regions to total losses are unbalanced. In the dense distribution area, the estimation tends to be inaccurate, while in sparse distribution area, the isolated points are easily be ignored. Therefore, it is suboptimal to constrain the whole density map or simply divide it into some blocks.

This fact motivates us to explore the strategy to balance the losses of different blocks. Different from existing strategies, a spatial-awareness block loss (SBL) with the region-weighted reward strategy is proposed for congested vehicle counting scenes, as shown in Fig. 3. The main idea of SBL is not only to divide blocks but also to weight each block, which aims to control the loss contributions of different blocks. The predicted density map is divided into M blocks and the L_2 loss of each block is calculated independently. In order to balance the proportion of M blocks in the final loss, we plan to enlarge the proportion of the blocks with smaller loss and meanwhile reduce the impact of the blocks with larger loss, so that both dense and sparse areas can be perceived by the VCNet. K_{th} Root Strategy (KRS) is a good solution to achieve this operation, which satisfies our expectation. In addition, we also explore a couple of strategies, which will not be discussed in this paper due to space limitation. The weighted losses of M blocks are finally summed to obtain the final loss of the predicted density map. Specifically, the proposed SBL loss is formulated as follows,

$$SBL(D^{pr}, D^{gt}) = \frac{1}{2N}\sum_{i=1}^{N}\sum_{j=1}^{M}\sum_{p\in B_{ij}} S(\left\|D_{ij}^{pr}(p) - D_{ij}^{gt}(p)\right\|_2^2) \quad (4)$$

where N is the number of training images. M denotes the number of blocks, and B_{ij} denotes the j_{th} block in i_{th} density map. $D_{ij}^{pr}(p)$ and $D_{ij}^{gt}(p)$ denotes the counting results of the j_{th} blocks in i_{th} predicted density map and ground truth density map, respectively. For convenience, θ is denoted as,

$$\theta = \left\|D_{ij}^{pr}(p) - D_{ij}^{gt}(p)\right\|_2^2 \quad (5)$$

The K_{th} Root Strategy (KRS): The K_{th} Root Strategy (KRS) aims to change the loss distribution of the blocks, which is formulated as follows,

$$S_{KRS}(\theta) = \sqrt[K]{\left\|D_{ij}^{pr}(p) - D_{ij}^{gt}(p)\right\|_2^2}, K > 1 \quad (6)$$

With the K_{th} Root Strategy (KRS), the loss larger than 1 would be reduced. On the contrary, the loss smaller than 1 would be enlarged but still smaller than 1. As the value of K increases, the loss distribution will be changed to a larger extent.

Inspired by the step decay learning rate, the step decay block loss with a dynamic K value is proposed, which adopts the K_{th} Root Strategy (KRS) for weighting and can be defined as follows:

$$S_{KRS_{step}}(\theta) = \sqrt[K_{step}]{\left\|D_{ij}^{pr}(p) - D_{ij}^{gt}(p)\right\|_2^2} \quad (7)$$

where K_{step} defines the set of K values which is dynamically changed according to the number of training steps.

4 EXPERIMENTS

4.1 Implementation Details

In this paper, the truncated VGG-16 is adopted as the backbone of VCNet for end-to-end training. The first 10 convolution layers are pre-trained on ImageNet [18] and the other layers are initialized with a Gaussian distribution with 0.01 standard deviation. The dilation rates of the hybrid dilated convolution block in each branch are set to $\{1, 1, 1\}$, $\{1, 3, 5\}and\{3, 4, 5\}$, respectively. The Adam optimizer is utilized to train the model with a fixed learning rate of 1e-5 for vehicle counting and a fixed learning rate of le-6 for crowd counting. The number of training epochs is 200. The number of BL blocks is set to 16, which performs well in the benchmark datasets. The K_{th} Root Strategy of BL is adopted to get training loss for different models, in which the K_{step} are set to $\{10, 5, 10/3, 2.5, 2\}$ and decays every 40 epochs.

Figure 4: Examples of the newly collected CVCSet, which is a challenging dataset.

4.2 Datasets

TRANCOS [9]: TRANCOS is a public image dataset for vehicle counting with different traffic scenes. The number of images in TRANCOS is 1244, among which the training, validation and testing sets are 403, 420 and 421 images, respectively. The total number of annotated vehicles is 46796 [22]. Besides, the region of interest (ROI) of training and evaluation is provided, where the background and small vehicles far away are removed.

VisDrone2019 Vehicle [53]: VisDrone2019 is a public object detection dataset with 8599 images and 10 object categories of interest. Similar to [1], a subset containing the categories of car, van, truck, and bus is used for vehicle counting. To meet the practical scenarios as close as possible and increase the challenge of the task, severely occluded or truncated objects are also counted. Moreover, the cases having less than 10 annotated objects are filtered out. Finally, this dataset consists of 5025, 455 and 1242 training, validation and test samples, respectively. To generate the ground truth for vehicle counting, the original bounding box is changed to the center point.

CVCSet: CVCSet is our newly collected vehicle counting dataset, which contains 3885 images from diverse scenarios, including highway roads, toll stations and service stations. The total number of annotated vehicles is 132524, and on average each image has around 34 vehicles. Compared to TRANCOS, CVCSet is a challenging dataset with inconsistent vehicle densities, imbalanced spatial distributions, severe occlusions, complex backgrounds, diverse visual and scale variations, which is illustrated in Fig. 4. Moreover, the dense and small vehicle targets are also reserved and annotated, which posts the challenge of small object counting in complex and dense background. The training, validation and testing sets of CVCSet contain 2313, 497 and 1075 images, respectively.

ShanghaiTech [50]: ShanghaiTech crowd counting dataset contains 1198 annotated images with a total amount of 330165 persons [22]. This dataset consists of two parts. Part A contains 482 images with highly congested scenes, which are randomly downloaded from the Internet. Part B includes 716 images with relatively sparse crowd scenes, which are taken from streets in Shanghai.

4.3 Evalution Metrics

Mean Absolute Error (MAE) and Mean Square Error (MSE) are commonly used performance metrics for crowd counting [22], which are defined as follows,

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|C_i^{pr} - C_i^{gt}| \tag{8}$$

Table 1: Performance comparison on TRANCOS.

Method	GAME 0	GAME 1	GAME 2	GAME 3
Hydra 3s [29]	10.99	13.75	16.69	19.32
AMDCN [6]	9.77	13.16	15.00	15.87
FCNN-skip [17]	4.61	8.39	11.08	16.10
LSC-CNN [32]	4.60	5.40	6.90	8.30
FCN-HA [51]	4.21	-	-	-
CSRNet [22]	3.56	5.49	8.57	15.04
DensityCNN-H [15]	3.17	4.78	6.30	8.26
KDMG [40]	3.13	4.79	6.20	8.68
HSRNet [54]	3.03	4.57	6.46	9.68
E2D [55]	2.88	4.81	7.77	12.47
PSDDN [27]	4.79	5.43	6.68	8.40
ADSCNet [2]	2.60	-	-	-
Shi et al. [35]	2.00	-	-	-
DADNet [10]	2.79	4.41	6.43	9.27
VCNet	**1.90**	**3.03**	**4.01**	**5.65**

Table 2: Performance comparison on VisDrone2019 Vehicle.

Method	GAME 0	GAME 1	GAME 2	GAME 3
VGG-16 [1]	21.4	-	-	-
MCNN [1]	14.9	-	-	-
CSRNet [22]	9.91	11.88	14.43	17.11
CAN [26]	9.5	11.21	13.27	16.18
SACANet [1]	8.6	-	-	-
VCNet	**3.20**	**4.24**	**5.27**	**6.39**

Table 3: Performance comparison on CVCSet.

Method	GAME 0	GAME 1	GAME 2	GAME 3
MCNN [52]	18.62	35.44	45.62	52.27
CSRNet [22]	9.31	11.60	14.60	17.90
CAN [26]	9.03	11.21	13.85	16.43
VCNet	**4.12**	**4.67**	**5.35**	**6.43**

$$MSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(C_i^{pr} - C_i^{gt})^2} \tag{9}$$

where C_i^{pr} and C_i^{gt} refer to the predicted and ground-truth numbers of the crowd, respectively. N is the total number of tested images. However, MAE metric often leads to mask mistaken estimations [9] as it ignores the locations where the estimation is done in the images.

Furthermore, Grid Average Mean Absolute Error (GAME) [9] combines the quantity and location to measure the performance, which is more convincing for vehicle counting evaluation. With GAME metric, the image is subdivided into 4^L non-overlapping regions, and MAE of each region is calculated, respectively. GAME is formulated as follows,

$$GAME(L) = \frac{1}{N}\sum_{i=1}^{N}(\sum_{j=1}^{4^L}|C_i^{j^{pr}} - C_i^{j^{gt}}|) \tag{10}$$

where $C_i^{j^{pr}}$ and and $C_i^{j^{gt}}$ indicates the predicted and ground-truth numbers of the crowd in the region j of the i_{th} image of testing set, respectively.

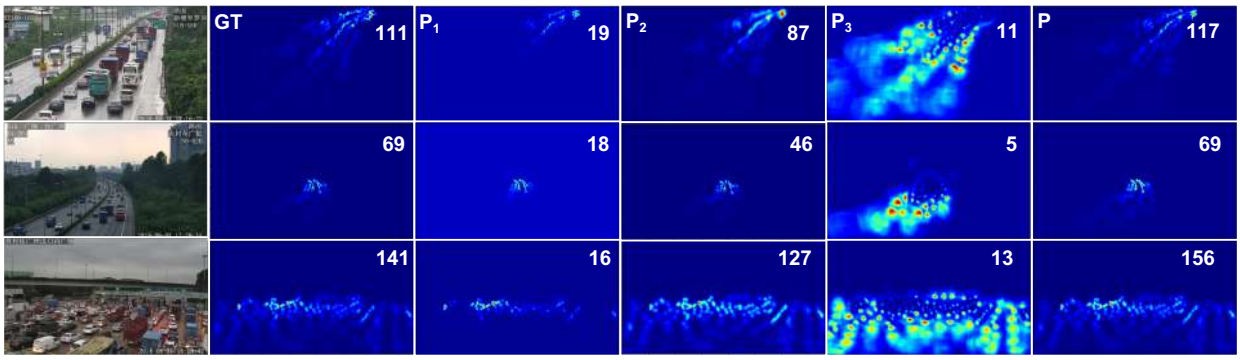

Figure 5: The estimated density maps from the attention-refined multi-scale density map estimation module. $\{P_i, i \in [1, 2, 3]\}$ **is the attention-refined multi-scale density maps.** P **refers to the final prediction of VCNet.**

4.4 Comparisons with SOTA methods

First, we compare the proposed method with the state-of-the-art approaches on three benchmark datasets, TRANCOS, VisDrone2019 Vehicle and CVCSet. Grid Average Mean Absolute Error (GAME) is employed to measure the performance, which combines the quantity and location to measure and is convincing in the evaluation of the vehicle counting task.

The performance comparison on TRANCOS dataset is listed in Table 1. The method in [29] uses a CNN-based multi-scale nonlinear regression model to generate the density map. CSRNet [22] adopts dilated convolution instead of a multi-column structure to enlarge the receptive field of the network. DADNet [10] deploys deformable convolution to promote the accuracy of object localization in the density map. Compared to DADNet, VCNet achieves huge improvements of 31.8%, 45.5%, 37.6% and 39.1% in GAME(0), GAME(1), GAME(2) and GAME(3), respectively. As discussed in [9], GAME metric becomes more sensitive to location information as the number of blocks L increases. The improvement of VCNet in GAME(3) metric fully shows that the density map predicted by VCNet not only has high counting accuracy but also has good locating accuracy. This is mainly due to the strong ability of spatial awareness that the block loss brings to VCNet.

The results on VisDrone2019 Vehicle dataset are listed in Table 2. We compare the proposed VCNet with five state-of-the-art methods [1, 22, 26, 52]. Among these methods, the official codes with the original settings are adopted to reproduce CSRNet and CAN on VisDrone2019 Vehicle dataset. The results of VGG-16, MCNN and SACANet in [1] are directly copied. Different from TRANCOS, the background noise information and small-scale objects in VisDrone2019 Vehicle dataset are retained, so that this dataset has more complex scene changes and serious scale variation for objects. Our VCNet outperforms other methods on all GAME metrics. Compared to the result from [1], VCNet decreases the errors by 62.8% for GAME(0). Although SACANet [1] has advantages in solving scale changes and isolated clusters in crowd counting tasks, it is not enough to deal with more serious scale variation in vehicle counting tasks. The proposed method integrates the multi-branch structure and an attention-based refinement method to deal with scale issues, which outperforms other methods on VisDrone2019 Vehicle dataset.

The results of different methods [22, 26, 52] on the CVCSet datasets are presented in Table 3. CVCSet includes complex scenes, severe occlusions, and extreme scale changes, which bring huge challenges for vehicle counting. It can be seen from Table 3 that VCNet achieves the best results on CVCSet among the listed methods. Compared to CAN, VCNet achieves improvements of 55.4%, 58.3%, 61.4% and 60.9% in GAME(0), GAME(1), GAME(2) and GAME(3), respectively. This demonstrates that VCNet is robust and effective for vehicle counting, which can deal with the problem of severe occlusions and scale variations.

4.5 Ablation Studies

In this section, CSRNet [22] is adopted as the baseline to perform ablation studies on TRANCOS and VisDrone2019 Vehicle datasets.

Effectiveness of attention-refined multi-scale density map estimation. To verify the effectiveness of the attention-refined multi-scale density map estimation module (AMDE), the networks of two settings are compared. (i) Baseline + MBS: the multi-branch structure with hybrid dilated convolution blocks is integrated into CSRNet and the multi-scale density maps are directly summed to generate the final density map. (ii) Baseline + AMDE: attention-refined multi-scale density map estimation module is used to replace the backend of the CSRNet. All two networks are trained using the MSE loss. As shown in Table 4, two networks have been improved to some extent. Specially, AMDE achieves 35.4%, 37.2%, 47.1% and 58.6% improvement compared to the baseline on TRANCOS, respectively, and similar results are obtained on VisDrone2019 Vehicle dataset. Moreover, it can be seen from Fig. 5: i) multi-scale density maps from the multi-branch structure tend to more accurately estimate the number of vehicles with different sizes, ii) the salient regions of multi-scale density maps are well segmented through the attention-refined mask maps. All of these explain the effectiveness of AMDE.

Effectiveness of Spatial-awareness block loss. To test the performance of our method, different strategies of the SBL loss are integrated into the baseline. Other settings of the baseline are reserved for a fair comparison. In this way, the effectiveness of SBL loss is verified. Table 4 shows the results of the baseline with or without SBL loss on TRANCOS dataset and VisDrone2019 Vehicle dataset, respectively. It can be seen that the precision of the baseline is greatly improved with SBL. Based on the original setting of the baseline, the performance of the baseline is improved with different

Table 4: The results of ablation studies on TRANCOS dataset and VisDrone2019 Vehicle dataset.

Method	TRANCOS				VisDrone2019 vehicle			
	GAME 0	GAME 1	GAME 2	GAME 3	GAME 0	GAME 1	GAME 2	GAME 3
Baseline [22]	3.56	5.49	8.57	15.04	9.91	11.88	14.43	17.11
Baseline + SBL ($KRS_{k=2}$)	2.21	3.19	4.18	5.93	5.26	6.42	7.71	9.01
Baseline + SBL ($KRS_{k=4}$)	2.12	3.15	4.14	5.79	5.14	6.33	7.63	8.93
Baseline + SBL (KRS_{step})	2.12	3.10	4.08	5.71	5.03	6.15	7.22	8.47
Baseline + MBS	2.29	3.35	4.52	6.35	4.51	6.26	8.35	10.34
Baseline + AMDE	2.18	3.27	4.43	6.01	3.63	5.92	7.74	9.54
VCNet (KRS_{step} + AMDE)	**1.90**	**3.03**	**4.01**	**5.65**	**3.20**	**4.24**	**5.27**	**6.39**

Table 5: Performance comparison with the detection-based methods on VisDrone2019 Vehicle dataset.

Method	MAE	MSE
FRCN [30]	24.77	32.72
SSD [25]	18.07	24.79
SDC [20]	15.58	20.09
FCOS [38]	15.33	22.39
MRCN [11]	6.69	11.99
RetinaNet [23]	6.56	10.92
GA-FPN [41]	7.25	12.31
SCRDet [46]	5.51	9.81
CODAN [21]	4.88	8.78
VCNet	**2.76**	**5.63**

Table 6: Performance comparison on crowd counting dataset ShanghaiTech.

Method	Part_A		Part_B	
	MAE	MSE	MAE	MSE
MCNN [52]	110.2	173.2	26.4	41.3
Switching-CNN [33]	90.4	135.0	21.6	33.4
CP-CNN [36]	73.6	106.4	20.1	30.1
CSRNet (Baseline) [22]	68.2	115.0	10.6	16.0
PACNN [34]	66.3	106.4	8.9	13.5
ADCrowdNet [24]	63.2	98.9	**7.6**	13.9
BL [28]	62.8	101.8	7.7	12.7
RPNet [47]	61.2	96.9	8.1	**11.6**
VCNet (AMDE)	64.5	101.4	8.8	14.7
VCNet (KRS_{step})	62.3	93.9	8.0	13.2
VCNet (KRS_{step} + AMDE)	**60.8**	**91.6**	7.7	13.1

settings of SBL. Specifically, the $SBL(KRS_{step})$ strategy achieves huge improvements to the baseline with percentages of 40.4%, 43.5%, 52.4% and 62.0% in GAME(0), GAME(1), GAME(2) and GAME(3) on TRANCOS. Similar results are reported on VisDrone2019 Vehicle. Moreover, $SBL(KRS_{k=2})$ and $SBL(KRS_{k=4})$ have also made good improvements on two datasets respectively. The experimental results illustrate the effectiveness of SBL. The experiment results provide strong evidence for the effectiveness of the proposed SBL.

4.6 Comparisons with Detection-based Methods

To illustrate the value of the regression-based method in the vehicle counting task, we compare the performance of VCNet with the detection-based vehicle counting methods. The same dataset as [21] is used to evaluate the performance, which is modified from the VisDrone2019 object detection dataset. Since VCNet only uses point annotations, MAE and RMSE are reported. The performance comparison on VisDrone2019 Vehicle dataset is listed in Table 5. The experimental results clearly show that VCNet outperforms all of the detection-based methods without using bounding box information. Comparing the second-ranked method, VCNet achieved 2.12 and 3.15 improvements in MAE and RMSE, respectively.

4.7 Experiments on Crowd Counting Dataset

The proposed method is also tested on ShanghaiTech [52] crowd counting dataset. The comparison of VCNet and the state-of-the-art methods on ShanghaiTech is listed in Table 6. It can be seen that VCNet performs well in crowding counting and outperforms most of the state-of-the-art methods. Compared to the baseline model, VCNet improves MAE and MSE values from 68.2, 115.0 to 60.8, 91.6 on ShanghaiTech Part A, which is a huge improvement. The good performance of VCNet on two tasks of vehicle counting and crowd counting fully demonstrates the effectiveness of the proposed method.

5 CONCLUSION

In this paper, a novel network called Vehicle Counting Network (VCNet) is proposed for vehicle counting of congested traffic scenes. In order to generate a high-quality density map, the attention-refined multi-scale density map estimation module is designed to be aware of varying scales of vehicles and suppress overlap between multi-scale density maps. By adopting a more rational loss function called spatial-awareness block loss (SBL), VCNet can capture more spatial information of occlusions, dense and sparse distribution, etc. Extensive experiments conducted on four benchmark datasets demonstrate the effectiveness of VCNet. Compared to state-of-the-art methods, VCNet achieves the best performance with GAME metrics, especially for congested traffic scenes that exhibit large scale variation, diverse visual appearance, severe occlusions, or inconsistent spatial distributions.

ACKNOWLEDGMENTS

This work was supported in part by the National Natural Science Foundation of China (Grant No: 61772436, 62001400), Sichuan Science and Technology Program (Grant No. 2020YJ0207, 2021YJ0364), Foundation for Department of Transportation of Henan Province, China (2019J-2-2), and Grant of Institute of Applied Physics and Computational Mathematics, Beijing (Grant No. HXO2020-118).

REFERENCES

[1] Haoyue Bai, Song Wen, and S.-H. Gary Chan. 2019. Crowd Counting on Images with Scale Variation and Isolated Clusters. In *Proceedings of the IEEE International Conference on Computer Vision*. 18–27.

[2] Shuai Bai, Zhiqun He, Yu Qiao, Hanzhe Hu, Wei Wu, and Junjie Yan. 2020. Adaptive Dilated Network With Self-Correction Supervision for Counting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4594–4603.

[3] Antoni B. Chan and Nuno Vasconcelos. 2009. Bayesian Poisson regression for crowd counting. In *Proceedings of the IEEE International Conference on Computer Vision*. 545–551.

[4] Zhi-Qi Cheng, Jun-Xiu Li, Qi Dai, Xiao Wu, and Alexander G. Hauptmann. 2019. Improving the Learning of Multi-column Convolutional Neural Network for Crowd Counting. In *Proceedings of the ACM International Conference on Multimedia*. 1897–1906.

[5] Zhi-Qi Cheng, Jun-Xiu Li, Qi Dai, Xiao Wu, and Alexander G. Hauptmann. 2019. Learning Spatial Awareness to Improve Crowd Counting. In *Proceedings of the IEEE International Conference on Computer Vision*. 6152–6161.

[6] Diptodip Deb and Jonathan Ventura. 2018. An Aggregated Multicolumn Dilated Convolution Network for Perspective-Free Counting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 195–204.

[7] Piotr Dollár, Christian Wojek, Bernt Schiele, and Pietro Perona. 2012. Pedestrian Detection: An Evaluation of the State of the Art. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 34, 4 (2012), 743–761.

[8] Weina Ge and Robert T. Collins. 2009. Marked point processes for crowd counting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2913–2920.

[9] Ricardo Guerrero-Gómez-Olmedo, Beatriz Torre-Jiménez, Roberto Javier López-Sastre, Saturnino Maldonado-Bascón, and Daniel Oñoro-Rubio. 2015. Extremely Overlapping Vehicle Counting. In *Iberian Conference on Pattern Recognition and Image Analysis*. 423–431.

[10] Dan Guo, Kun Li, Zheng-Jun Zha, and Meng Wang. 2019. Dadnet: Dilated-attention-deformable convnet for crowd counting. In *Proceedings of the ACM International Conference on Multimedia*. 1823–1832.

[11] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. 2017. Mask r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision*. 2961–2969.

[12] Meng-Ru Hsieh, Yen-Liang Lin, and Winston H Hsu. 2017. Drone-based object counting by spatially regularized regional proposal network. In *Proceedings of the IEEE International Conference on Computer Vision*. 4145–4153.

[13] Xiaowei Hu, Xuemiao Xu, Yongjie Xiao, Hao Chen, Shengfeng He, Jing Qin, and Pheng-Ann Heng. 2018. SINet: A scale-insensitive convolutional neural network for fast vehicle detection. *IEEE transactions on intelligent transportation systems* 20, 3 (2018), 1010–1019.

[14] Xiaolong Jiang, Zehao Xiao, Baochang Zhang, Xiantong Zhen, Xianbin Cao, David S. Doermann, and Ling Shao. 2019. Crowd Counting and Density Estimation by Trellis Encoder-Decoder Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 6133–6142.

[15] Xiaoheng Jiang, Li Zhang, Tianzhu Zhang, Pei Lv, Bing Zhou, Yanwei Pang, Mingliang Xu, and Changsheng Xu. 2020. Density-Aware Multi-Task Learning for Crowd Counting. *IEEE Transactions on Multimedia* 23 (2020), 443–453.

[16] Di Kang and Antoni B. Chan. 2018. Crowd Counting by Adaptively Fusing Predictions from an Image Pyramid. In *British Machine Vision Conference*. 89.

[17] Di Kang, Zheng Ma, and Antoni B. Chan. 2019. Beyond Counting: Comparisons of Density Maps for Crowd Analysis Tasks - Counting, Detection, and Tracking. *IEEE Transactions on Circuits and Systems for Video Technology* 29, 5 (2019), 1408–1422.

[18] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2012. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems*. 1097–1105.

[19] Min Li, Zhaoxiang Zhang, Kaiqi Huang, and Tieniu Tan. 2008. Estimating the number of people in crowded scenes by MID based foreground segmentation and head-shoulder detection. In *International Conference on Pattern Recognition*. 1–4.

[20] Wei Li, Hongliang Li, Qingbo Wu, Xiaoyu Chen, and King Ngi Ngan. 2019. Simultaneously detecting and counting dense vehicles from drone images. *IEEE Transactions on Industrial Electronics* 66, 12 (2019), 9651–9662.

[21] Wei Li, Zhenting Wang, Xiao Wu, Ji Zhang, Qiang Peng, and Hongliang Li. 2020. CODAN: Counting-Driven Attention Network for Vehicle Detection in Congested Scenes. In *Proceedings of the ACM International Conference on Multimedia*. 73–82.

[22] Yuhong Li, Xiaofan Zhang, and Deming Chen. 2018. CSRNet: Dilated Convolutional Neural Networks for Understanding the Highly Congested Scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1091–1100.

[23] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision*. 2980–2988.

[24] Ning Liu, Yongchao Long, Changqing Zou, Qun Niu, Li Pan, and Hefeng Wu. 2019. ADCrowdNet: An Attention-Injective Deformable Convolutional Network for Crowd Understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3225–3234.

[25] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. 2016. Ssd: Single shot multibox detector. In *Proceedings of the European Conference on Computer Vision*. 21–37.

[26] Weizhe Liu, Mathieu Salzmann, and Pascal Fua. 2019. Context-Aware Crowd Counting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 5099–5108.

[27] Yuting Liu, Miaojing Shi, Qijun Zhao, and Xiaofang Wang. 2019. Point in, box out: Beyond counting persons in crowds. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 6469–6478.

[28] Zhiheng Ma, Xing Wei, Xiaopeng Hong, and Yihong Gong. 2019. Bayesian Loss for Crowd Count Estimation With Point Supervision. In *Proceedings of the IEEE International Conference on Computer Vision*. 6141–6150.

[29] Daniel Oñoro-Rubio and Roberto Javier López-Sastre. 2016. Towards Perspective-Free Object Counting with Deep Learning. In *Proceedings of the European Conference on Computer Vision*. 615–629.

[30] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *Advances in Neural Information Processing Systems*, Vol. 28. 91–99.

[31] David Ryan, Simon Denman, Clinton Fookes, and Sridha Sridharan. 2009. Crowd Counting Using Multiple Local Features. In *Digital Image Computing: Techniques and Applications*. 81–88.

[32] Deepak Babu Sam, Skand Vishwanath Peri, Mukuntha Narayanan Sundararaman, Amogh Kamath, and R Venkatesh Babu. 2019. Locate, size and count: Accurately resolving people in dense crowds via detection. *arXiv preprint arXiv:1906.07538* (2019).

[33] Deepak Babu Sam, Shiv Surya, and R. Venkatesh Babu. 2017. Switching Convolutional Neural Network for Crowd Counting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4031–4039.

[34] Miaojing Shi, Zhaohui Yang, Chao Xu, and Qijun Chen. 2019. Revisiting Perspective Information for Efficient Crowd Counting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 7279–7288.

[35] Zenglin Shi, Pascal Mettes, and Cees G. M. Snoek. 2019. Counting With Focus for Free. In *Proceedings of the IEEE International Conference on Computer Vision*. 4200–4209.

[36] Vishwanath A. Sindagi and Vishal M. Patel. 2017. Generating High-Quality Crowd Density Maps Using Contextual Pyramid CNNs. In *Proceedings of the IEEE International Conference on Computer Vision*. 1879–1888.

[37] Vishwanath A. Sindagi and Vishal M. Patel. 2019. Multi-Level Bottom-Top and Top-Bottom Feature Fusion for Crowd Counting. In *Proceedings of the IEEE International Conference on Computer Vision*. 1002–1012.

[38] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. 2019. Fcos: Fully convolutional one-stage object detection. In *Proceedings of the IEEE International Conference on Computer Vision*. 9627–9636.

[39] Elad Walach and Lior Wolf. 2016. Learning to Count with CNN Boosting. In *Proceedings of the European Conference on Computer Vision*. 660–676.

[40] Jia Wan, Qingzhong Wang, and Antoni B Chan. 2020. Kernel-based Density Map Generation for Dense Object Counting. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2020).

[41] Jiaqi Wang, Kai Chen, Shuo Yang, Chen Change Loy, and Dahua Lin. 2019. Region proposal by guided anchoring. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2965–2974.

[42] Meng Wang and Xiaogang Wang. 2011. Automatic adaptation of a generic pedestrian detector to a specific traffic scene. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3401–3408.

[43] Panqu Wang, Pengfei Chen, Ye Yuan, Ding Liu, Zehua Huang, Xiaodi Hou, and Garrison Cottrell. 2018. Understanding convolution for semantic segmentation. In *2018 IEEE Winter Conference on Applications of Computer Vision*. IEEE, 1451–1460.

[44] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. 2018. Cbam: Convolutional block attention module. In *Proceedings of the European Conference on Computer Vision*. 3–19.

[45] Fan Yang, Heng Fan, Peng Chu, Erik Blasch, and Haibin Ling. 2019. Clustered object detection in aerial images. In *Proceedings of the IEEE International Conference on Computer Vision*. 8311–8320.

[46] Xue Yang, Jirui Yang, Junchi Yan, Yue Zhang, Tengfei Zhang, Zhi Guo, Xian Sun, and Kun Fu. 2019. Scrdet: Towards more robust detection for small, cluttered and rotated objects. In *Proceedings of the IEEE International Conference on Computer Vision*. 8232–8241.

[47] Yifan Yang, Guorong Li, Zhe Wu, Li Su, Qingming Huang, and Nicu Sebe. 2020. Reverse Perspective Network for Perspective-Aware Object Counting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4373–4382.

[48] Anran Zhang, Jiayi Shen, Zehao Xiao, Fan Zhu, Xiantong Zhen, Xianbin Cao, and Ling Shao. 2019. Relational Attention Network for Crowd Counting. In *Proceedings of the IEEE International Conference on Computer Vision*. 6788–6797.

[49] Anran Zhang, Lei Yue, Jiayi Shen, Fan Zhu, Xiantong Zhen, Xianbin Cao, and Ling Shao. 2019. Attentional Neural Fields for Crowd Counting. In *Proceedings of the IEEE International Conference on Computer Vision*. 5714–5723.

[50] Cong Zhang, Hongsheng Li, Xiaogang Wang, and Xiaokang Yang. 2015. Cross-scene crowd counting via deep convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 833–841.

[51] Shanghang Zhang, Guanhang Wu, Joao P Costeira, and José MF Moura. 2017. Fcn-rlstm: Deep spatio-temporal neural networks for vehicle counting in city cameras. In *Proceedings of the IEEE International Conference on Computer Vision*. 3667–3676.

[52] Yingying Zhang, Desen Zhou, Siqin Chen, Shenghua Gao, and Yi Ma. 2016. Single-Image Crowd Counting via Multi-Column Convolutional Neural Network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 589–597.

[53] Pengfei Zhu, Longyin Wen, Xiao Bian, Haibin Ling, and Qinghua Hu. 2018. Vision meets drones: A challenge. *arXiv preprint arXiv:1804.07437* (2018).

[54] Zhikang Zou, Yifan Liu, Shuangjie Xu, Wei Wei, Shiping Wen, and Pan Zhou. 2020. Crowd Counting via Hierarchical Scale Recalibration Network. *arXiv preprint arXiv:2003.03545* (2020).

[55] Zhikang Zou, Huiliang Shao, Xiaoye Qu, Wei Wei, and Pan Zhou. 2019. Enhanced 3D convolutional networks for crowd counting. *arXiv preprint arXiv:1908.04121* (2019).