

# Counting with Adaptive Auxiliary Learning

Yanda Meng, Joshua Bridge, Meng Wei, Yitian Zhao, Yihong Qiao, Xiaoyun Yang, Xiaowei Huang, and Yalin Zheng\*

**Abstract**—This paper proposes an adaptive auxiliary task learning based approach for object counting problems. Unlike existing auxiliary task learning based methods, we develop an attention-enhanced adaptively shared backbone network to enable both task-shared and task-tailored features learning in an end-to-end manner. The network seamlessly combines standard Convolution Neural Network (CNN) and Graph Convolution Network (GCN) for feature extraction and feature reasoning among different domains of tasks. Our approach gains enriched contextual information by iteratively and hierarchically fusing the features across different task branches of the adaptive CNN backbone. The whole framework pays special attention to the objects’ spatial locations and varied density levels, informed by object (or crowd) segmentation and density level segmentation auxiliary tasks. In particular, thanks to the proposed dilated contrastive density loss function, our network benefits from individual and regional context supervision in terms of pixel-independent and pixel-dependent feature learning mechanisms, along with strengthened robustness. Experiments on seven challenging multi-domain datasets demonstrate that our method achieves superior performance to the state-of-the-art auxiliary task learning based counting methods. Our code is made publicly available at: [https://github.com/smallmax00/Counting\\_With\\_Adaptive\\_Auxiliary](https://github.com/smallmax00/Counting_With_Adaptive_Auxiliary)

**Index Terms**—Objects Counting, GCN, Dilated Contrastive Density Loss, Adaptive Auxiliary Task

## I. INTRODUCTION

OBJECT counting by inferring the number of objects in images or video contents is a crucial yet challenging computer vision task. This paper is primarily motivated to address crowd counting problems while it can be applied to other counting problems such as cell and vehicle counting. Due to the need for crowd gathering in many scenarios such as parades, concerts, and stadiums, a robust and accurate crowd counting model plays an essential role in multimedia applications for security alerts, public space design, crowd management, *etc.* [1].

Benefits from Convolutional Neural Network (CNN)’s excellent feature learning ability, the performance of crowd counting approaches has consistently been improved. Recent state-of-the-art approaches, such as [2], [3], [4], [5], [6],

Y. Meng, J. Bridge and Y. Zheng are with the Department of Eye and Vision Science, University of Liverpool, Liverpool, L7 8TX, United Kingdom.

M. Wei was with the Department of Eye and Vision Science, University of Liverpool, Liverpool, L7 8TX, United Kingdom.

Y. Zhao is with the Cixi Institute of Biomedical Engineering, Ningbo Institute of Materials Technology and Engineering, Chinese Academy of Science, Ningbo 315201, China.

Y. Qiao is with the China Science IntelliCloud Technology Co., Ltd, Shanghai, China.

X. Yang is with Remark AI UK Limited, London, SE1 9PD, United Kingdom.

X. Huang is with the Department of Computer Science, University of Liverpool, Liverpool, L69 3BX, United Kingdom.

Corresponding author: Yalin Zheng (yalin.zheng@liverpool.ac.uk).

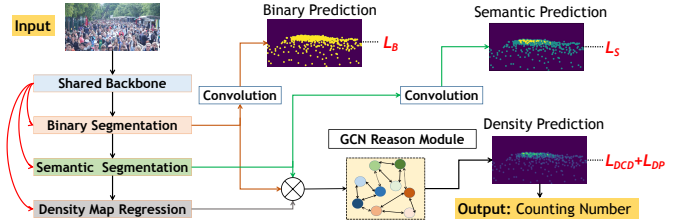


Fig. 1. Overview of the proposed network structure in the scene of crowd counting. An attention-enhanced adaptively shared backbone network is proposed to enable both task-shared and task-tailored features learning. A novel Graph Convolution Network (GCN) reasoning module is introduced to tackle issues of cross-domain features reasoning among three different tasks. A novel loss function  $L_{DCD}$  is proposed to take into account more adjacent pixels for regional density difference, which strengthens the network’s generalizability.

showed that a density map regression paradigm gains satisfying results. In these methods, given an input image, a CNN-based network is used to regress the corresponding density map; the summed pixel value of the density map gives the total counting numbers in that image. However, there is still much room to improve the counting performance due to challenging issues [1], such as significant scale changes, wide variations of density levels, and complex scene backgrounds. To solve these problems, some previous methods [7], [8], [9], [10], [11] relied on different types of information granularity in terms of ‘auxiliary task learning’. These methods applied a single shared backbone network structure to extract generalized features for all the tasks. Unfortunately, this strategy may lead to under-fitting as the generalizable representation often cannot effectively describe the comprehensive cross-domain features across different tasks at the same time [12], [13]. Intuitively, our motivation is that the backbone network should be able to yield both generic (or universal) representations shareable for all the tasks and specialized features tailored for individual tasks. To this end, we designed an attention-based adaptive shareable backbone network to enable task-shared and task-customized features learning in an end-to-end manner (Fig. 1 shows an overview of the network architecture). Note that, the term ‘auxiliary task learning’ is referred to as the feature learning of different density information granularity levels. Specifically, the crowd segmentation task and the density level segmentation task in Fig. 1 are the auxiliary tasks, and the density map regression task is the main task. We generated the ground truth of crowd segmentation and density level segmentation tasks from the density map regression task’s ground truth. Intuitively, the information from the ground truth of auxiliary tasks is not increased; however, the information is enhanced and specified through auxiliary tasks in terms of different density information granularity.

Instead of aiming to achieve the best performance for all

the tasks, our adaptive shareable backbone network primarily focused on optimizing the primary density map regression task, along with the multi-granularity information enhancement from the other auxiliary tasks. Our backbone network contained a multi-level information aggregation mechanism to iteratively and hierarchically fuse features learned from different stages and auxiliary branches to tackle large scale changes problems. We also applied two attention-based auxiliary task learning branches: 1.) crowd segmentation task to indicate spatial regions of interest to tackle complex background issues; and 2.) density level segmentation task to be aware of the varied density levels across the image, which can tackle the significant variations of density problems. An ablation study demonstrates that our adaptive auxiliary branches improve performance over a single-shared backbone based auxiliary task learning network with comparable model size.

Apart from the aforementioned components, we also studied how to reason and fuse features from different tasks for density map regression. The features extracted from crowd segmentation and density level segmentation branches belong to different feature domains with various granularity representations. Direct fusion (*e.g.* element-wise multiplication or channel-wise concatenation) among three task branches' outputs can result in domain conflicts [14]. Thus, further reasoning is necessary to improve the counting performance. To this end, we exploited the information reasoning nature of Graph Convolutional Networks (*GCN*). *GCN* has recently shown promising reasoning ability on many computer vision tasks, such as scene understanding [15], image segmentation [16], [17], *etc.*, but has been rarely studied in the crowd counting task domain. Our model projects a collection of spatial-aware density feature map's pixels with similar density levels to each graph vertex and exploits a *GCN* to reason about the relations among graph vertices. This is different from a recent work [14], which directly treated cross-domain feature maps as graph vertices and utilized a cascaded Graph Neural Network (*GNN*) to reason the cross-scale relationships. Our experiment results have proved that the proposed *GCN* reasoning module helps to improve the counting accuracy.

We also proposed a novel loss function to supervise the main task learning processes. Notably, for the supervision of density map regression, the widely adopted Least Absolute Error (*L1*) or Least Square Error (*L2*) loss in previous counting methods [18], [7], [19], [20] assumes pixel-wise independence, which supervises the predicted density map based on the individual pixels. However, it has two significant weaknesses. Firstly, the predicted density map tends to be over-smooth [8]; specifically, it may underestimate high-density level regions and overestimate low-density level regions. As a result, the model may primarily focus on achieving lower count errors rather than regressing high-quality density maps; thus, it cannot reflect the actual density levels. Secondly, without a large receptive field, individual pixel-wise loss functions may ignore the regional density level information during the training process [21]. Unbalanced low and high-level density distributions may introduce significant bias in the training process, thus weakening the network's robustness. To address these issues, we proposed a novel loss function for density map

regression, called Dilated Contrastive Density Loss ( $L_{DCD}$ ), where the density difference among dilated adjacent pixels is utilized to provide additional regional supervision. Ablation studies demonstrate that our proposed regional loss function can improve the counting performance of the pixel-wise loss supervised methods.

In summary, this work makes the following contributions: 1.) We addressed the feature learning issues of the backbone network of the auxiliary task in crowd counting challenges, by enabling task-shareable and task-specified feature learning simultaneously with a primary focus on the main task. 2.) We proposed crowd segmentation and density level segmentation as auxiliary tasks in crowd counting with additional spatial crowd location and density level information enhancement. Moreover, a *GCN* model was proposed to reason about the cross-domain feature relations between density map regression and other auxiliary tasks. 3.) We proposed a novel loss function tailored for density map regression, strengthening the network's generalizability and improving the counting accuracy. 4.) We conducted extensive experiments on seven well-known challenging counting benchmarks. Quantitative and qualitative results demonstrated that our model achieves state-of-the-art performance. Especially, to the best of our knowledge, we achieved the best counting performance among auxiliary tasks based counting methods on NWPU-Crowd [22] benchmark<sup>1</sup>, which is currently the largest crowd counting benchmark. Our model is robust and generalizable to indicate the wrong labeled or miss labeled object in the test datasets. Please refer to Section V-D for more details.

## II. RELATED WORK

In recent years density map regression-based counting methods with *CNN* achieved good performance. For example, *Boominathan et al.* [23] proposed a dual-column network to combine low-level and high-level features in different layers to estimate the count. However, because of the conflicts from optimization among different columns [24], these types of network structures have difficulty in attaining global minimization. Other works employed single column network structures and handled different scale challenges [25], [26], [27], [28], [29], [30] with adaptive modules, such as scaled spatial pyramid pooling [19], [31], [32] or Dilated kernels of filters [33], [34], [35], [36], [37]. They achieved promising counting performance along with architectural simplicity and training efficiency.

### A. Attention-Based Counting

The visual attention mechanism was applied among several works [38], [21], [39], [40], [41], [42], [36] in the crowd counting task, which helped the network focus on valuable information and addressed several challenges. For example, *Miao et al.* [38] utilized a shallow feature based attention module to highlight the regions of crowd interest and filter out the noise in the background clutter. To tackle various density levels issues, *Jiang et al.* [21] employed an attention

<sup>1</sup><https://www.crowdbenchmark.com/nwpucrowd.html>

mask to refine the density map for adapting to different density levels. Furthermore, *Zhang et al.* [39] proposed the *Attention Neural Field* that incorporated non-local attention modules and conditional random fields to maintain multi-scale features and long-range dependencies, to handle large scale changes problems of the input crowd images. *Wan et al.* [43], [36] exploited the self-attention mechanism to adaptively generate density maps with different Gaussian kernel sizes, which is used as the ground truth to supervise the model. The aforementioned methods adopted the attention mechanism as a feature enhancement module to implicitly address the crowd counting task’s challenges, such as significant scale changes, wide variations of density levels, and complex scene backgrounds. Our model explicitly addressed those challenges through auxiliary tasks. On the other hand, our model adopted the attention mechanism to construct an adaptively shared backbone network, enabling the task-shared and task-specific features learning simultaneously.

### B. Auxiliary Tasks Based Counting

Recently, auxiliary task learning based counting methods [44], [45], [46], [47], [48], [49], [50], [51], [9], [52], [53] attracted researchers’ attention because of its ability to capture extra granularity information and contextual dependencies for the density map regression. Most of the methods utilized the potential of a model itself with auxiliary tasks, such as object detection, crowd segmentation, density level classification, *etc.*, to enhance the feature tuning for density map regression. For example, the task of patch-based density level classification [7], [54], [55], [9], [56], [57], [58] can enhance patch-level density level information, which helped to address the underestimation and the overestimation problems of density map regressions. However, it may be difficult to guide the pixel-wise density map regression via patch-wise density level classification because of the gap between pixel-level and patch-level feature learning. In contrast, our model proposed a density level segmentation auxiliary task, which can be regarded as the dense pixel-wise density level classification task. In this way, our model can enhance the pixel-wise density level information to the pixel-wise density map regression task, aiming to address the challenges of wide variations of density levels.

Moreover, because the background regions in complex scenes contain confusing objects or similar appearance, the crowd segmentation task, adopted by previous methods [59], [7], [14], [10], [60], can provide spatial location information for the crowd, which highlighted the foreground over the background and guided the network focus on the region of interest. Our model also adopted the crowd segmentation task because of its superiority in spatial location information enhancement. Similarly, the task of object (crowd) detection [8], [61], [62], [63], [64], [65] can enhance location information and alleviate local spatial inconsistency issues in the density map.

### C. Learn to Count with Different Supervisions

Instead of tackling the counting task through different learning frameworks or strategies, recent methods [66], [67],

[68], [69], [70], [71], [72], [73], [74], [75] paid attention at the way of supervisions. For example, *Sravya et al.* proposed a bin loss [68] to enable the data-distribution aware optimization, which helped to address the domain variation challenges from different crowd data source. *Song et al.* [69] studied the counting problem in a different way, where a combination of *Euclidean* loss and *Cross Entropy* loss was used for point locations learning, instead of density map regression. Along the same line, *Bayesian* loss was proposed by [71] to provide more reliable supervisions at each annotated point. Differently, *Wan et al.* [70] studied the combination of pixel-wise loss and point-wise loss, which investigated the density map representation through an unbalanced optimal transport problem. [72] proposed a novel loss function to address the spatial annotation noise during training, where a weighted MSE term and a pixel-wise correlation term were involved. Recently, [73] proposed distribution matching loss to tackle the weakened generalizability of the Gaussian smoothed density map. Moreover, *Wang et al.* [74] treated the counting with density map as a classification problem, where a Cross-Entropy loss was used to classify each patch into certain intervals.

The aforementioned methods introduced different loss functions to supervise the model, such as points location, bounding box, matching, ranking, classification, *etc.* However, the mainstream counting methods still relies on pixel-wise supervision with the density map ground truth [1], such as *L1* or *L2* loss functions. In this work, we proposed a Dilated Contrastive Density Loss ( $L_{DCD}$ ) to improve the pixel-wise loss’s receptive field and increase the regional supervision.

## III. METHODOLOGY

### A. Ground Truth Generation

Following [76], given a set of  $N$  images  $\{I_i\}_{i=1}^N$  with corresponding point annotations  $\{P_i\}_{i=1}^N$ , the ground truth of the density map  $\{D_i\}_{i=1}^N$  is generated by filtering the points with a normalized Gaussian kernel. The total object count number  $T_i$  of image  $I_i$  can be attained by summing all pixel values of the density map  $D_i$ .

The ground truth mask of the crowd segmentation task is generated from the density map ground truth. Given a set of  $N$  density maps  $\{D_i\}_{i=1}^N$ , the value for each pixel in the mask  $\{B_i\}_{i=1}^N$  is set to 1 if the pixel value in the density map is larger than the zero and 0 otherwise.

The ground truth mask used by the density level segmentation task is also generated from the density map. For pixel  $p$  in input image  $i$ , its density level class  $S_{p,i}$  is given as:

$$S_{p,i} = \min_{i=1,\dots,N} \left( \lfloor \frac{D_i(p) - \min(D_i)}{\max(D_i) - \min(D_i)} \times L + 1 \rfloor, L + 1 \right), \quad (1)$$

where  $L$  represents the overall levels of density; following previous patch-based density level classification methods [7], [9], we set  $L$  equal to 4 in our work.  $D_i$  is the pixel value in the  $i_{th}$  density map ground truth.

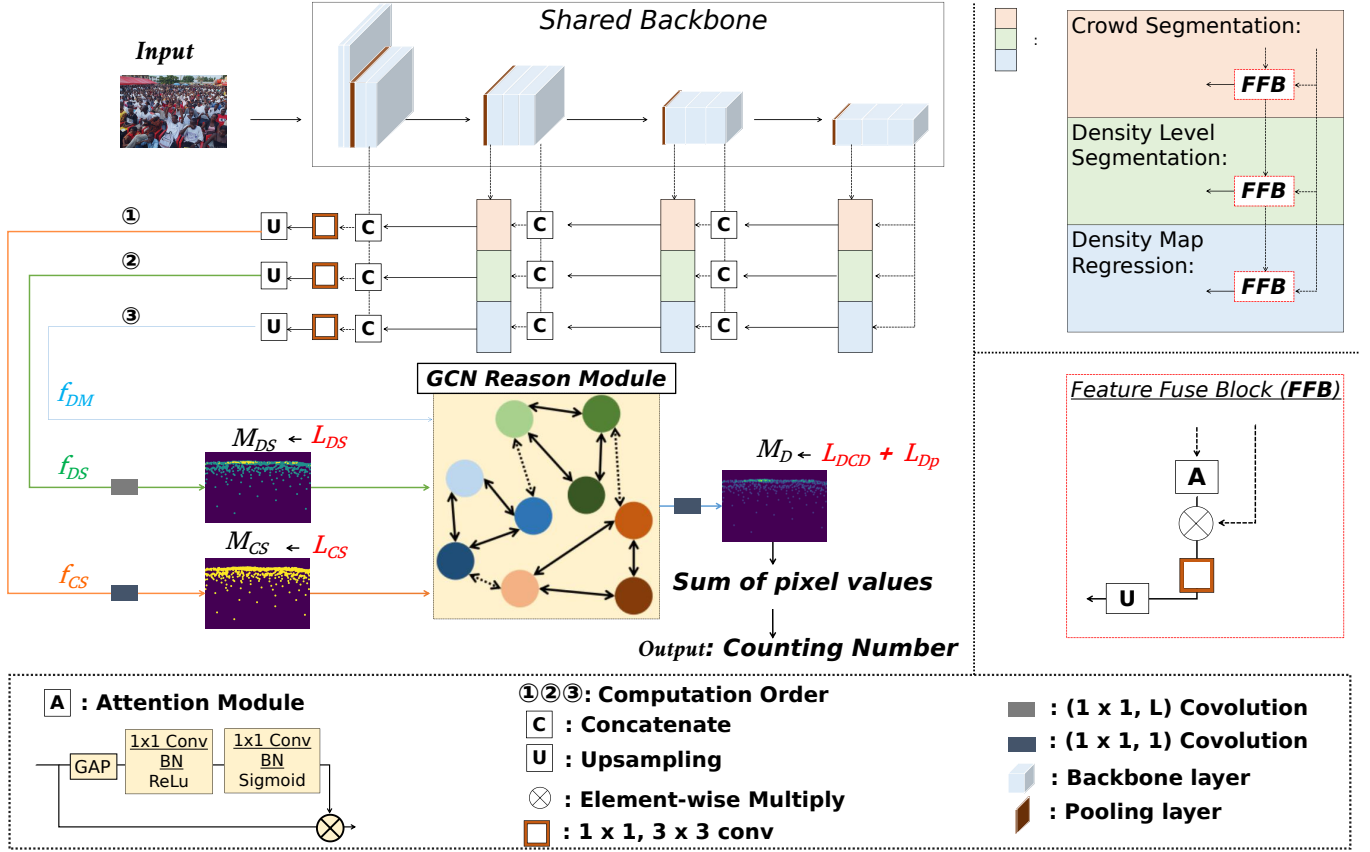


Fig. 2. Illustration of our proposed network. The adaptively shared backbone network has three outputs of  $f_{CS}$ ,  $f_{DS}$ ,  $f_{DM}$ , representing crowd segmentation, density level segmentation, and density map regression branches' output feature map, respectively. The order of their involvements indicates that the density map regression branch can benefit from the extra density level and crowd spatial supervision from the other two branches gradually.

### B. Task Adaptive Backbone Network

Instead of using a shared backbone network to extract generalizable features for different tasks, we proposed an attention-based task-adaptive shared backbone network to allow the model to extract discriminative features for the auxiliary tasks, thus helping to improve the performance of the main task. Fig. 2 shows the detailed structure of the proposed network, which consists of a shared backbone and three attention-based task-adaptive branches. To make a fair comparison with previous auxiliary task-based methods, such as [54], [14], [61], [9], *etc.*, the truncated VGG-16 [77] is used as the backbone network. However, it can be replaced by any other robust network structure; we have reported the counting performance with other powerful network backbones in TABLE. II. The shared backbone adopts the first 13 layers of the VGG-16 to extract multi-level features. To exploit the global contextual dependencies, we proposed a Feature Fuse Block (FFB), which aggregated and fused the outputs from posterior layers back to the preceding layers hierarchically and iteratively, with up-sampling, concatenation and convolution operations. This provides improvements in extracting the full spectrum of semantic and spatial information across different stages and resolutions. The up-sampling is performed by using a bilinear interpolation algorithm. The convolution operation aims to reduce and match the corresponding feature map channel size between different stages.

With the aggregating process from low-level features to high-level features, the task-adaptive attention module is applied in three different task branches; details of the attention module are shown in the bottom left of Fig. 2. Each attention module consists of a global average pooling (GAP) layer to capture global context through different feature map channels, conducting an attention tensor to lead the emphasis of feature learning. Then, two blocks with a convolutional layer followed by a Batch Normalization (BN) [78] layer with *ReLU* and sigmoid as the activation functions are added. For the convolutional layer filter, the kernel size is  $1 \times 1$ . The element-wise multiplication is then performed between the outputs of the particular layer of the shared backbone and the task-specific attention module, which filters out the unrelated and redundant features from the backbone with respect to different auxiliary tasks and the main task. Therefore, the shared backbone can learn a generalizable representation, while the attention-based branches can extract task-specific features simultaneously in an end-to-end manner. The ablation study experiments proved that the attention-based adaptive backbone could boost the counting performance.

Apart from the aforementioned network structure component in three attention-based task-adaptive branches, we also introduced a cross-domain feature fusing operator in a particular order to focus on optimizing the primary density map regression task primarily. Specifically, the crowd seg-

mentation branch is applied to the shared backbone first to select the corresponding discriminative spatial features. Then, we applied the density level segmentation branch on the shared backbone and crowd segmentation branch, which can enhance the additional contextual density level information into the main task. At last, the main task of the density map regression branch is applied.

### C. Auxiliary Tasks

With three outputs from the task adaptive backbone network, we built two auxiliary tasks and a main task: crowd segmentation task, density level segmentation task, and density map regression task. We detail each of them subsequently.

**Crowd Segmentation.** We introduced crowd segmentation as one of the auxiliary tasks for two reasons. Firstly, the density map’s pixel value should be zero at non-crowd regions. However, the predicted density map can be noisy and inaccurate when the background is cluttered and complex. The crowd segmentation task provides a spatial focus to the density map regression process through zero out the non-crowd regions’ pixel values. Secondly, given the standard set-up of single density map regression, the pixels within a particular range of the point annotations should contribute more to the final counting results; however, the loss is dominated by the majority of less relevant pixels. To overcome this limitation, the crowd segmentation can provide additional information enhancement in terms of the spatial indicator with a standalone loss function.

Given an input image  $I_i \in \mathbb{R}^{3 \times H \times W}$ , we can get the output of the crowd segmentation branch in the backbone network,  $f_{CS} \in \mathbb{R}^{C \times H \times W}$ , where  $H$  and  $W$  represent the height and width of the feature map;  $C$  is the channel size. Then, we apply a convolution layer with filter parameters  $\theta_{CS} \in \mathbb{R}^{1 \times 1 \times 1}$ , followed by a sigmoid as the activation function. Through this operation, we can generate a probability map to calculate the crowd and background probability. The single channel crowd segmentation probability map  $M_{CS}$  is defined as:  $M_{CS} = \text{Sigmoid}(\theta_{CS}, f_{CS}) \in \mathbb{R}^{1 \times H \times W}$ .

**Density Level Segmentation.** Density map regression is a pixel-wise task, which focuses on low-level features learning but may ignore high-level contextual information during the training [31]. To address this issue, we perform density level segmentation as another auxiliary task. Compared with previous patch-based density level classification methods [7], [54], [55], [9], our proposed pixel-based density level segmentation can provide pixel-wise level density information and high-level semantic features at the same time. Upon the output of the density level segmentation branch of the backbone network  $f_{DS} \in \mathbb{R}^{C \times H \times W}$ , a convolution layer with filter parameters  $\theta_{DS} \in \mathbb{R}^{L \times 1 \times 1}$  and a softmax as activation function are applied. The prediction of density level segmentation branch  $M_{DS}$  is defined as:  $M_{DS} = \text{softmax}(\theta_{DS}, f_{DS}) \in \mathbb{R}^{L \times H \times W}$ , where  $L$  is the number of density levels.

### D. Density Map Regression

Intuitively, the different granularity features of density levels and spatial crowd locations need to be further reasoned to

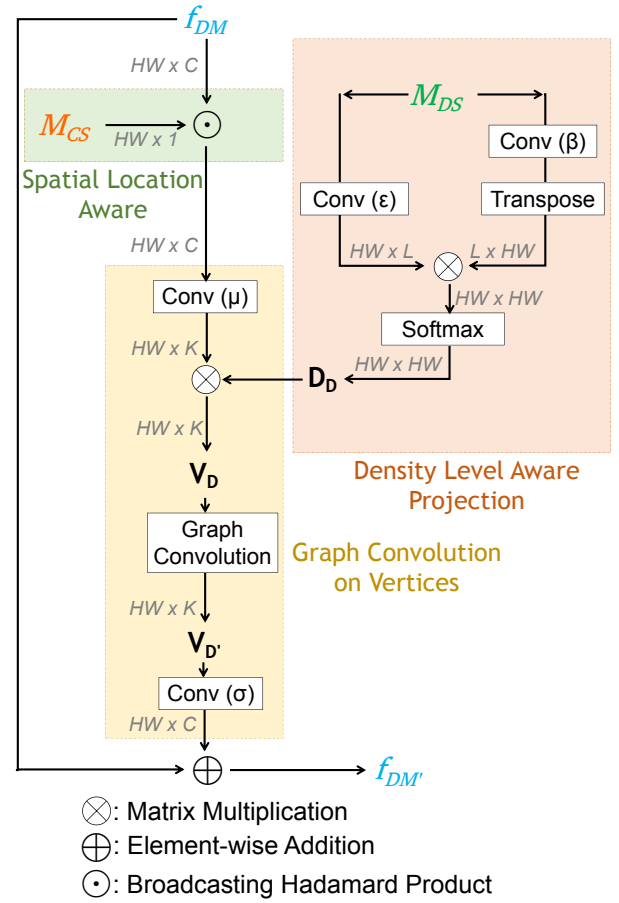


Fig. 3. Architecture of the proposed GCN reasoning module.  $f_{DM} \in \mathbb{R}^{C \times H \times W}$  is the feature map of the density map regression branch,  $C = 32$  is the channel size;  $M_{CS} \in \mathbb{R}^{1 \times H \times W}$  is the prediction of the crowd segmentation branch;  $M_{DS} \in \mathbb{R}^{L \times H \times W}$  is the prediction of density level segmentation branch,  $L = 4$  is the number of density levels;  $D_D \in \mathbb{R}^{HW \times HW}$  is the density level dependency matrix;  $V_D \in \mathbb{R}^{K \times HW}$  is the constructed vertex features and  $V_{D'} \in \mathbb{R}^{K \times HW}$  is the output vertex features after GCN,  $K = 16$  is the number of vertices.  $f_{DM'} \in \mathbb{R}^{C \times H \times W}$  is the output feature map after GCN reasoning.

fuse into the density map regression branch. To this end, with the predicted crowd segmentation output  $M_{CS}$  and density level segmentation output  $M_{DS}$  as the auxiliary information granularity, we input them along with the density map branch’s feature map  $f_{DM} \in \mathbb{R}^{C \times H \times W}$  into the GCN reasoning module to reason the relationship among themselves. Subsequently, the output feature map  $f_{DM'} \in \mathbb{R}^{C \times H \times W}$  of the GCN reasoning module is reduced into one-channel through  $1 \times 1$  convolution layer with a *ReLU* activation function.

### E. GCN Reasoning Module

The proposed GCN reasoning module structure is shown in Fig. 3. In detail, there are three primary modules: *Spatial Location Aware* module, *Density Level Aware Projection* module, *Graph Convolution on Vertices* module. Because of the nature of the crowd images, the density level varies across the image [19], which indicates that the pixel values of the density map should not just rely on their own pixel-wise features but also on different density level regions. To this end, our GCN reason

model projected a collection of spatial-aware density feature map’s pixels with similar density levels to each graph vertex and exploited a *GCN* to reason about the relations among graph vertices.

**Spatial Location Aware Module.** Before projecting the density map feature map  $f_{DM}$  into the graph vertices, we directly applied the broadcasting Hadamard Product operation between the crowd segmentation output  $M_{CS}$  and the density map regression branch’s feature map  $f_{DM}$ . There are two underlying reasons: (1)  $M_{CS}$  is a one-channel crowd segmentation map, with encoding the probability of the non-crowd regions’ pixel values approaching zero and crowd regions’ pixel values approaching one; one can serve as a filter to zero out the non-crowd region’s pixel value of the density map. (2) Direct broadcasting Hadamard Product can achieve crowd spatial awareness for every channel of the  $f_{DM}$  through zero out the non-crowd region’s pixel value. This can address the challenge of the complex scene backgrounds in the crowd images.

**Density Level Aware Projection Module.** As mentioned above, the pixel-wise density level information can help to address the challenges of the large variations of density levels in crowd images. However, direct broadcasting Hadamard product between the density map branch’s feature map  $f_{DM}$  and the density level output  $M_{DS}$  may result in domain conflicts [14]. We exploited the nature of *GCN* and projected the density level information into the graph vertices for further reasoning; one benefits the long-range relationship reasoning ability of *GCN* and the multi-granularity information enhancement from density level. Inspired by the non-local module [79], we encoded the long-range density level dependency among every pixel. Give the feature map  $M_{DS}$ , the density level dependency matrix  $D_D \in \mathbb{R}^{HW \times HW}$  is defined as:

$$D_D = \text{softmax}(\epsilon(M_{DS}) \otimes \beta^T(M_{DS})), \quad (2)$$

where  $\text{Conv } \beta$  and  $\text{Conv } \epsilon$  are two convolution layers with  $1 \times 1$  kernel size, respectively. The dependency matrix  $D_D$  can be regarded as a pixel-wise attention map, where pixels with similar density levels are assigned with larger weights. The dependency matrix itself can reflect the pixel-wise density level dependency. Besides, we projected it as a prior to the graph domain through matrix multiplication, which can enhance high-level contextual dependency simultaneously.

**Graph Convolution on Vertices.** In this module, we learnt how to reason the region-based relationship in density map through *GCN* in graph domain. Firstly, we projected the spatial aware feature map of  $f_{DM}$  into graph domain with  $K$  vertices, and each vertex was represented by an embedding of shape  $H \times W$ . This is achieved by  $\text{Conv}(\mu)$ , which is a  $1 \times 1$  convolution layer. Furthermore, we projected the dependency matrix  $D_D$  to the graph domain through matrix multiplication, resulting in the vertex features  $V_D \in \mathbb{R}^{K \times HW}$ . The projection aggregated pixels with similar density levels to graph vertices, where each vertex represents a region in the crowd image. Formally,  $V_D$  is defined as:

$$V_D = D_D \otimes \mu(f_{DM} \odot M_{CS}), \quad (3)$$

where  $\otimes$  is matrix multiplication;  $\odot$  is broadcasting Hadamard product. With the constructed vertices, the long-range region-

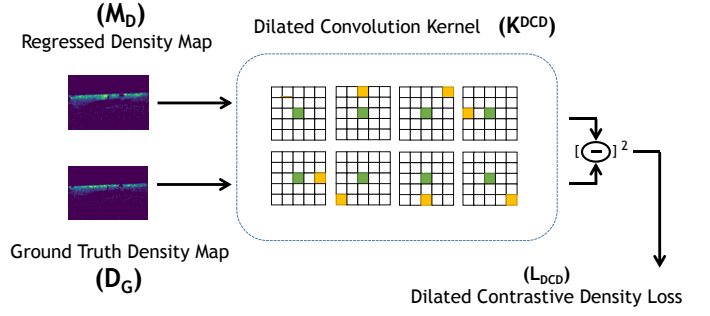


Fig. 4. Dilated Contrastive Density Loss ( $L_{DCD}$ ). There are eight dilated contrastive kernels with green, white, yellow blocks representing 1, 0, -1, respectively. The least-square error of two outputs from regression and ground truth is treated as the final  $L_{DCD}$ .

wise relationship is further reasoned in the graph domain through *GCN*. In detail, we reasoned over the region-wise relations by propagating information across vertices with a single layer *GCN*. Specifically, we fed the constructed vertex features  $V_D$  into a first-order approximation of spectral graph convolution [80], resulting the output vertex features  $V_{D'} \in \mathbb{R}^{K \times HW}$ . The  $V_{D'}$  is calculated as:

$$V_{D'} = \text{ReLU}\left((I - A) \otimes V_D \otimes W_D\right), \quad (4)$$

where  $I$  is the identity matrix;  $A \in \mathbb{R}^{HW \times HW}$  denotes the adjacent matrix that encodes the graph connectivity to learn;  $W_D \in \mathbb{R}^{K \times K}$  is the weights of the *GCN*. The adjacent matrix  $A$  is randomly initialized but can learn and update the edge weights from vertex features by gradient along the training process. The identity matrix  $I$  serves as a residual connection that alleviates the optimization difficulties. Based on the learned graph, the information propagation across all vertices leads to the finally reasoned relations between regions. After graph reasoning, a collection of pixels embedded within one vertex share the same context of features modeled by graph convolution. Then, we re-projected the vertex features in the graph domain to the original pixel grids. Given the reasoned vertices  $V_{D'}$ , we applied  $\text{Conv}(\sigma)$ , which is a  $1 \times 1$  convolution layer. Finally, we summed up the re-projected refined and the original density feature maps as the final feature map. The final pixel-wise density feature map  $f_{DM'}$  is thus computed by

$$f_{DM'} = f_{DM} + \sigma(V_{D'}). \quad (5)$$

## F. Loss Function

The whole network is end-to-end trainable, which includes four loss functions; the total loss function is defined as:

$$L_{total} = L_{CS} + L_{DS} + \gamma \cdot (L_{DP} + L_{DCD}), \quad (6)$$

where  $\gamma$  is empirically set as 2, which is a hyper-parameter to trade-off between auxiliary losses and main loss. Please note that, extensive experiments have been done to determine the weights of the losses for two auxiliary tasks, respectively. We found that there is no significant difference of counting

performance with respect to different weight values; thus, we set them both equal to 1 in the loss function. Binary cross-entropy ( $L_{CS}$ ) is used for crowd segmentation auxiliary task; categorical cross-entropy ( $L_{DS}$ ) is used for density level segmentation auxiliary task;  $L_2$  loss is used for pixel-wise density map regression supervision ( $L_{DP}$ ). However, pixel-wise  $L_2$  loss assumes pixel-wise independence, which results in over-smooth density map prediction [8] and the underlying bias from unbalanced low- and high-level density distributions of crowd images. To address the issue, we proposed Dilated Contrastive Density Loss ( $L_{DCD}$ ), where we take into account more adjacent pixels for regional density difference. In detail, we applied single layer convolution on the regressed density map  $M_D$  and the ground truth density map  $D_G$  respectively. The single layer convolution has eight filters; each filter contains dilated kernel with a fixed value (*e.g.* 1, 0, and -1). The least-square error of the calculated regional dilated contrastive values from the regressed and ground truth density map is the output of  $L_{DCD}$ . To this end, we define  $L_{DCD}$  as below:

$$L_{DCD} = \sum_i \|K_i^{DCD} \otimes M_D - K_i^{DCD} \otimes D_G\|_2^2, \quad (7)$$

where  $K_i^{DCD}$  is the  $i^{th}$  dilated contrastive convolution kernel,  $i \in [1, 8]$ . Details of the kernel are shown in Fig. 4, where a  $3 \times 3$  convolution layer with the dilated rate of 2 is applied; one gives a larger receptive field as  $5 \times 5$ . We perform extensive experiments to evaluate the effectiveness of the proposed  $L_{DCD}$  loss; quantitative results in *Ablation Study* (Section V-C) demonstrates that the proposed  $L_{DCD}$  loss can improve the counting accuracy not only for our model but also for previous single  $L_2$  loss based methods.

## IV. EXPERIMENTS

### A. Datasets

**ShanghaiTech** [18] consists of 1,198 images, containing a total amount of 330,165 people with head centre point annotations. This dataset is divided into two parts: **SHA** includes 482 images, in which crowds are mostly dense (33 to 3139 people); **SHB** includes 716 images, where crowds are sparser (9 to 578 people). Each part is divided into training and testing subset as specified in [18]. **UCF-QNRF** [82] is a large crowd dataset, consisting of 1,535 images with about 1.25 million annotations in total. The number of people in these images varies largely with a wide range from 49 to 12,865. As indicated by [82], For training, 1,201 images are used, the remaining 334 images form the test set. **JHU-Crowd** [83] is a recent challenging large-scale dataset that containing 4,372 images with 1.51 million annotations. The dataset includes several challenging scenes such as weather-based degradation and illumination variations *etc.*. This dataset is divided into 2,272 images for training, 500 images for validation, and 1,600 images for testing. **NWPU-Crowd** [22] is up to date the largest public crowd counting dataset, containing 5,109 images with over 2.13 million annotations. The dataset includes 3,109 training images, 500 validation images and 1,500 test images. Moreover, inspired by the potential of crowd counting,

we conducted experiments on commonly used cell counting dataset: **DCC** [84] with 100 images for training and 77 images for testing, and vehicle counting dataset: **Trancos** [85] with 403 images for training, 420 images for validation and 421 images for testing. These experiments further demonstrated our model’s robustness and applicability for different real-world applications.

Note that, for ShanghaiTech (**SHA**, **SHB**), **UCF-QNRF**, and **DCC** dataset, we use 10% of the given training images as the validation dataset.

### B. Implementation Details

To augment the dataset, we randomly cropped the input images, density maps, crowd segmentation masks, and density level segmentation masks with fixed size  $128 \times 128$  at a random location, then randomly horizontally flipped the image patches with the probability of 0.3. We trained our model for 400 epochs for all experiments, with a start learning rate of  $1e-4$  and a cosine decay schedule [86]. The batch size is set to 96. All the training processes are performed on a server with 8 TESLA V100 and 4 TESLA P100, and all the test experiments are conducted on a local workstation with a Geforce RTX 2080Ti. Five-fold cross-validation is used for fair comparison and hyper-parameters tuning in all settings.

### C. Evaluation Metrics

To evaluate the counting performance, we adopted Mean Absolute Error ( $MAE$ ) and Root Mean Squared Error ( $RMSE$ ). Since Mean Absolute Error ( $MAE$ ) and Root Mean Square Error ( $RMSE$ ) cannot measure the counted objects’ locations, Grid Average Mean absolute Error ( $GAME$ ) is used to indicate counting accuracy over local regions.  $GAME$  is defined as:

$$GAME(L) = \frac{1}{N} \sum_{n=1}^N \left( \sum_{l=1}^{4^L} |y_n^l - \hat{y}_n^l| \right), \quad (8)$$

where  $N$  is the total number of images,  $y_n^l$  and  $\hat{y}_n^l$  are the ground truth and estimated counts in the local region  $l$  of  $n^{th}$  image.  $4^L$  denotes the number of non-overlapping regions which cover the full image. When  $L$  equals to 0, the  $GAME$  is equivalent to  $MAE$ .

## V. RESULTS

### A. Counting Results

In this section, we present our experimental results on the crowd, cell, and vehicle counting tasks in comparison to other **auxiliary-task based** state-of-the-art crowd counting methods. These experiments further demonstrate our model’s robustness and applicability in multiple domain datasets. In the Discussion (Section V-D), we showed that our model could indicate some mislabeled or wrongly labeled point annotations from the ground truth of the test dataset. This highlights our approach’s generalizability and the potential issue of imperfect ground truth in object counting datasets.

**Crowd Counting Results.** We performed experiments to validate our model’s performance in five challenging crowd

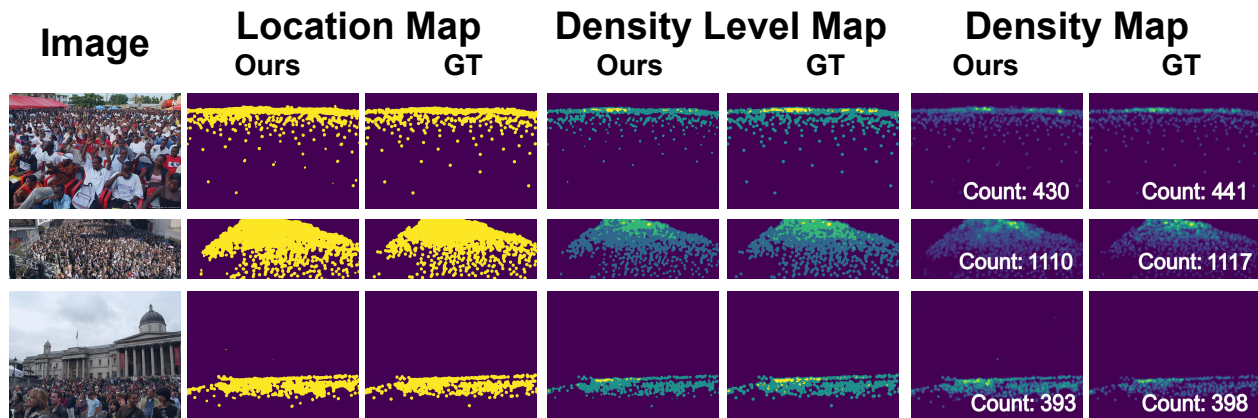


Fig. 5. Qualitative results of the density, crowd location and density level map in *SHA* test dataset. Our model can produce accurate density maps compared to the ground truth (*GT*), along with accurate auxiliary crowd segmentation and density level segmentation results.

TABLE I  
RESULTS ON FIVE CHALLENGING DATASETS FOR CROWD COUNTING, COMPARED WITH OTHER AUXILIARY TASK LEARNING BASED METHODS. OUR MODEL ACHIEVES A NEW STATE-OF-THE-ART WITHIN AUXILIARY LEARNING BASED COUNTING METHODS IN TERMS OF *MAE*.

Methods	<i>SHA</i>		<i>SHB</i>		<i>QNR</i>		<i>JHU-Crowd</i>		<i>NWPU-Crowd</i>	
	<i>MAE</i>	<i>RMSE</i>	<i>MAE</i>	<i>RMSE</i>	<i>MAE</i>	<i>RMSE</i>	<i>MAE</i>	<i>RMSE</i>	<i>MAE</i>	<i>RMSE</i>
<i>CP-CNN</i> [54]	73.6	106.4	20.1	30.1	-	-	-	-	-	-
<i>DecideNet</i> [8]	-	-	21.53	31.98	-	-	-	-	-	-
<i>CFF</i> [7]	65.2	109.4	7.2	12.2	93.8	146.5	83.6	400.7	80.8	364.1
<i>AT-CSRNet</i> [59]	-	-	8.11	13.53	-	-	-	-	-	-
<i>SHRGBD</i> [63]	70.3	111.0	8.8	15.3	113.3	177.6	107.9	446.7	103.0	478.1
<i>HA-CCN</i> [55]	62.9	94.9	8.1	12.7	118.1	180.4	-	-	-	-
<i>RAZ-Net</i> [62]	65.1	106.7	8.4	14.1	116	195	-	-	151.5	634.6
<i>HYGNN</i> [14]	60.2	94.5	7.5	12.7	100.8	185.3	-	-	-	-
<i>LSC-CNN</i> [61]	66.4	117.0	8.1	12.7	120.5	218.2	112.7	454.4	90.4	388.8
<i>ASCC</i> [21]	57.8	<b>90.1</b>	7.5	13.1	91.6	159.7	84.6	355.1	95.7	398.0
<i>UMRNet</i> [10]	62.6	103.3	7.2	11.5	86.3	153.1	-	-	-	-
<i>DAMNet</i> [9]	63.1	106.3	9.1	16.3	101.5	186.9	-	-	-	-
<i>MATT</i> [50]	59.5	97.3	<b>6.9</b>	<b>10.3</b>	-	-	-	-	-	-
<i>Ours</i>	<b>57.0</b>	98.6	7.1	12.3	<b>85.3</b>	<b>129.4</b>	<b>66.6</b>	<b>254.9</b>	<b>76.4</b>	<b>327.1</b>

counting datasets. Fig. 5 shows qualitative results; specifically, we presented the predictions from auxiliary task branches (crowd segmentation and density level segmentation masks) to demonstrate our model’s cohesion, along with the spatial location and density level variation’s contribution of auxiliary branches. To make a fair comparison, we only compared our model with previous **auxiliary task learning** based counting methods. TABLE. I shows that our method outperforms other methods in terms of *MAE* on all five datasets. In particular, our model outperforms the patch-based density level classification based method *HA-CCN* [55] by 14.7% via average *MAE*. Notably, the *JHU-Crowd* dataset [83] and *NWPU-Crowd* dataset [22] are recent public available datasets, which are more challenging due to large variations in scale, occlusion, and complex weather scenes. Specifically, *NWPU-Crowd* is current the largest crowd counting benchmark<sup>2</sup>. To the best of our knowledge, we achieved the best performance among other auxiliary task based methods. Except the auxiliary based

methods shown in TABLE. I, our method gains a superior reduction than single-task learning based methods as well, for example, scale-variation enhanced method *CACC* (100.1 *MAE*) [19] by 18.3% and dilated kernel-based method *CSR-Net* (85.9 *MAE*) [33] by 4.8% via *MAE*.

**Cell & Vehicle Counting Results.** We conducted experiments on cell (*DCC* [84]) and vehicle (*Trancos* [85]) counting datasets to show our model’s broad applicability and robustness. Fig. 7 shows the qualitative results, and Fig. III shows the quantitative results compared with the previous state-of-the-art methods. Due to the different scenes in the cell counting dataset, such as less occlusion, no scale variation, no complex background *etc.*, the contribution of some components of our model will be lessened because we design our model especially for crowd counting tasks; still, our model achieves comparable performance with previous methods. Furthermore, we presented local comparison performance through the *GAME* metric to indicate the model’s ability to recognize the objects’ locations. Fig. 6 shows the comparison results in terms of the *GAME* on the *Trancos* dataset. As illustrated, our method

<sup>2</sup><https://www.crowdbenchmark.com/nwpucrowd.html>



TABLE II  
RESULTS OF USING DIFFERENT BACKBONE NETWORKS ON FIVE CROWD COUNTING DATASETS.

Methods	SHA		SHB		QNRN		JHU-Crowd		NWPU-Crowd	
	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
VGG-16 [77]	<b>57.0</b>	98.6	7.1	12.3	<b>85.3</b>	129.4	<b>66.6</b>	254.9	<b>76.4</b>	<b>327.4</b>
VGG-19 [77]	59.7	99.8	8.4	13.2	87.8	144.0	73.7	320.1	79.9	360.0
ResNet-50 [81]	57.8	<b>96.6</b>	<b>7.0</b>	<b>11.7</b>	85.5	<b>128.7</b>	77.9	318.1	79.3	344.4
ResNet-101 [81]	61.1	100.8	9.1	14.5	93.3	147.9	69.7	<b>253.3</b>	81.4	361.5

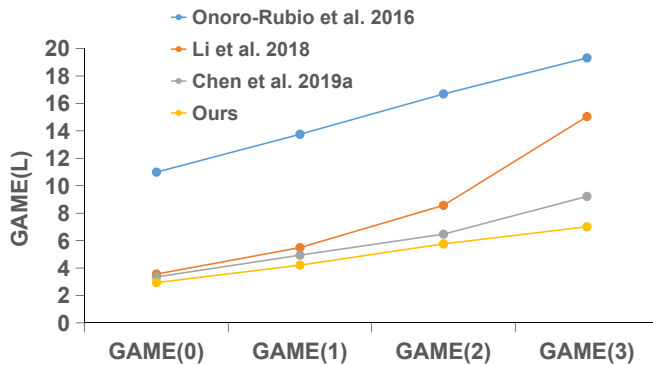


Fig. 6. Comparison of *GAME* performance on the *Trancos* dataset among the proposed approach and the state-of-the-arts, such as *Onoro-Rubio et al.* [2], *Li et al.* [33], *Chen et al.* [32]. Note that, a small range of increase among different *GAME* values indicates that our method counts and localizes overlapping vehicles more accurately.

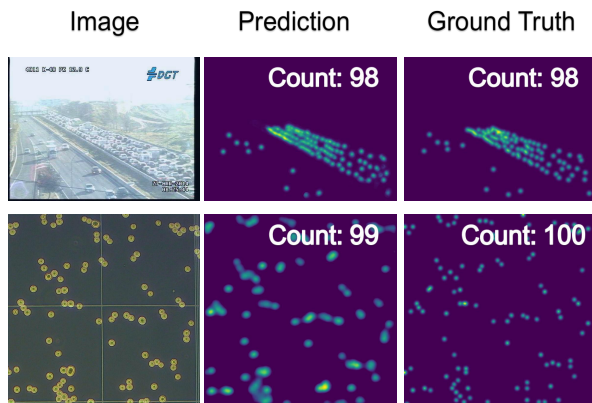


Fig. 7. Qualitative results on the *Trancos* (top) and *DCC* (bottom) dataset. Our model adapts well with scale variations and weather degradation challenges on the *Trancos* dataset. Further improvement is required for the cell dataset where individual cell locations are hard to distinguish, though the density levels and spatial distributions are clearly indicated.

localizes and counts overlapping vehicles more accurately.

### B. Auxiliary Task Results

In this section, we reported the performance of two auxiliary tasks. The commonly used segmentation metric Intersection over Union (*IoU*) is used to evaluate the auxiliary tasks' performance. In detail, we achieved average 88.7 % *IoU* for the crowd segmentation task and 81.0 % *IoU* for the density level segmentation task on the five crowd counting datasets.

TABLE III  
RESULTS ON CELL (*DCC*) COUNTING AND VEHICLE (*Trancos*) COUNTING DATASET. OUR MODEL ACHIEVES SUPERIOR PERFORMANCE TO THE PREVIOUS STATE-OF-THE-ART METHODS.

Methods	DCC	Trancos	
	MAE	MAE	RMSE
PPPD [84]	8.4	9.7	-
SAU-Net [3]	3.0	-	-
CSRNet [33]	-	3.5	5.1
CCF [7]	3.2	<b>2.0</b>	-
Ours	<b>2.9</b>	2.3	<b>4.8</b>

Fig. 5 shows examples of those tasks' predictions from our model.

### C. Ablation Study

We investigated the effect of each component in our proposed model. All ablation experiments were performed with the same settings detailed in the Implementation Details (Section IV-B).

**Ablation on Different Network Backbones** We evaluated the effectiveness of different backbone networks on the five crowd counting datasets. The counting performance is shown in TABLE. II with several different backbone networks. In general, *VGG*-based backbone networks achieved comparable counting performance, compared with the one of *ResNet*-based backbone networks in relatively large-scale datasets, such as *QNRN*, *JHU-Crowd* and *NWPU-Crowd*. While, *ResNet*-based backbone works better on small-scale counting datasets, such as *SHA* and *SHB*. We reported our model's performance with *VGG-16* backbone network in TABLE. I for a fair comparison with previous methods.

**Ablation on Auxiliary Tasks and Model Components.** In this section, we evaluated the effectiveness of the auxiliary tasks, adaptively shared backbone network, and *GCN*-enabled reasoning module, respectively. Please note that, in order to eliminate the performance improvement from a bigger model, we add feed-forward *CNN* blocks ( $3 \times 3$  convolution with Batch Normalization) into other ablation study models in TABLE. IV to maintain a similar model size as ours (18.8 million parameters). Firstly, we compared the single task density map regression network, in which we removed the *GCN* reasoning module, the auxiliary learning branches, and the adaptively shared backbone branches, to form a single column network structure (*Single Column*). Then we added two auxiliary branches separately and simultaneously after the single shared backbone's output to form an auxiliary learning mechanism

TABLE IV

ABLATION STUDY RESULTS ON NETWORK STRUCTURE COMPONENTS. EACH COMPONENT OF OUR NETWORK CONTRIBUTES TO THE FINAL PREDICTION.

Methods	SHA		JHU-Crowd	
	MAE	RMSE	MAE	RMSE
Single Column	71.3	122.3	99.3	391.0
w/ Crowd Seg	67.4	117.0	81.6	343.6
w/ Density Seg	68.1	119.9	86.1	360.0
w/ Both Auxiliary	65.2	115.2	77.3	311.7
w/ Adaptive Crowd Seg	61.3	104.6	75.7	300.9
w/ Adaptive Density Seg	63.8	108.1	76.9	307.8
w/ Both Adaptive Auxiliary	60.8	100.3	71.9	278.9
Ours	<b>57.0</b>	<b>98.6</b>	<b>66.6</b>	<b>254.9</b>

(w/ Crowd Seg, w/ Density Seg, w/ Both Auxiliary). To further improve the performance, we designed and added an adaptive backbone network to enable the task-shared and task-specific features being learned simultaneously (w/ Adaptive Crowd Seg, w/ Adaptive Density Seg, w/ Both Adaptive Auxiliary). Furthermore, we evaluated the proposed GCN reasoning module’s effectiveness, which can propagate region-based density level information across the image (Ours). The effect of each structural component is presented in Fig. IV. As illustrated, the proposed auxiliary task learning mechanism (w/ Both Auxiliary) is reduced by 14.3% over the single-task learning method (Single Column) via average MAE on two datasets, the task adaptive backbone (w/ Both Adaptive Auxiliary) reduces 6.8% over the single shared backbone (w/ Both Auxiliary), and the GCN reasoning module further reduces 6.7%. Qualitative comparison results of different modules’ effectiveness in terms of predicted density maps are shown in the Fig. 8, where the crowd segmentation auxiliary (w/ Adaptive Crowd Seg) can help the model to focus on the features in the region of interest and filter out the background (first and second rows). On the other hand, the density level segmentation auxiliary (w/ Adaptive Density Seg) can help to estimate more accurate density levels across the whole density map (second and third rows). We highlighted the different areas among those ablated models’ density map predictions with red bounding boxes for better visualization and comparison.

Moreover, in TABLE V, we further indirectly evaluate the auxiliary tasks’ effectiveness in this work. Specifically, for other ablation study models except for Ours, we maintained the same network structure as Ours to keep the same model size (18.8 million parameters) but switched off the two auxiliary tasks’ loss functions. In TABLE V, it proves that the supervision from multi-granularity information of auxiliary tasks contributes to the final counting performance in this work. Without  $L_{CS}$  and  $L_{DS}$  losses, the counting error increases by an average of 21.75 % on the SHA and the JHU-Crowd datasets via MAE.

**Ablation on Loss Function.** We performed experiments to evaluate the receptive field through different dilated rates in the proposed dilated contrastive density loss function  $L_{DCD}$ . In detail, we changed the dilated rate of the  $3 \times 3$  convolution layer into 1, 2, 3, 4, which resulted in the receptive field of the  $L_{DCD}$  being like 3, 5, 7, 9. TABLE VI shows the comparison results; when the dilated rate is 2, our model achieves the best

TABLE V

ABLATION STUDY RESULTS ON AUXILIARY TASKS. MAINTAINING THE SAME MODEL STRUCTURE (MODEL SIZE) AND TURNING OFF AUXILIARY TASKS’ LOSS FUNCTIONS CAN IMPLICITLY PROVE THAT THE AUXILIARY TASKS CONTRIBUTE TO THE FINAL COUNTING.

Methods	SHA		JHU-Crowd	
	MAE	RMSE	MAE	RMSE
w/o $L_{CS}$	64.4	107.7	78.7	310.5
w/o $L_{DS}$	62.0	104.8	74.9	302.2
w/o $L_{CS}$ and $L_{DS}$	67.1	115.2	93.0	377.5
Ours	<b>57.0</b>	<b>98.6</b>	<b>66.6</b>	<b>254.9</b>

TABLE VI

ABLATION STUDY RESULTS ON THE DILATED RATE OF THE PROPOSED LOSS FUNCTION  $L_{DCD}$ . WHEN THE DILATED RATE IS 2 AND THE CORRESPONDING RECEPTIVE FIELD IS 5, OUR MODEL CAN ACHIEVE THE BEST COUNTING PERFORMANCE ON THE SHA AND JHU-Crowd DATASETS.

Dilated Rate	SHA		JHU-Crowd	
	MAE	RMSE	MAE	RMSE
1	60.1	103.5	70.1	299.0
3	58.7	101.7	68.7	288.4
4	59.2	101.3	68.0	287.6
2 (Ours)	<b>57.0</b>	<b>98.6</b>	<b>66.6</b>	<b>254.9</b>

performance on SHA and JHU-Crowd datasets.

Furthermore, we conducted experiments to evaluate the effectiveness of the proposed dilated contrastive loss function, in which we removed the  $L_{DCD}$  and kept the rest of the network constant with the same trade-off hyper-parameters (Base in TABLE VII). Furthermore, we applied the proposed combined loss function (w/ contrastive in TABLE VII) into previous single  $L_2$  based methods. We re-implemented their network with their open-source code and used the same experimental setting as our method. Fig. VII shows the comparison results of our proposed combined loss function; as illustrated, with regional density difference supervision of  $L_{DCD}$ , our model attains a 3.5% reduction compared with single  $L_2$  loss function via average MAE on two datasets. Our proposed  $L_{DCD}$  also helps to reduce the original MCNN [18] by 6.4%, the CSRNet [33] by 2.7%, and the CACC [19] by 2.3% over average MAE on two datasets. Please note that, we did not compare with other loss functions that were proposed in recent crowd counting model [69], [71], [70], [73], [74], [72]. Because those methods are not pure density map regression based methods, it is unfair to compare.

TABLE VII

ABLATION STUDY RESULTS (MAE) ON OUR COMBINED LOSS (CONTRASTIVE AND  $L_2$  LOSS), COMPARED WITH SINGLE  $L_2$  LOSS (base). MOREOVER, WE APPLIED THE COMBINED LOSS FUNCTION TO OPTIMIZE PREVIOUS SINGLE  $L_2$  LOSS BASED METHODS TO DEMONSTRATE THAT THE COUNTING PERFORMANCE CAN BE IMPROVED WITH THE HELP OF REGIONAL DENSITY DIFFERENCE BASED LOSS FUNCTION  $L_{DCD}$ .

Methods	SHA		JHU-Crowd	
	Base	w/ contrastive	Base	w/ contrastive
MCNN [18]	110.2	<b>108.1</b>	188.9	<b>168.3</b>
CSRNet [33]	68.2	<b>65.9</b>	85.9	<b>84.1</b>
CACC [19]	62.3	<b>60.8</b>	100.1	<b>97.9</b>
Ours	59.5	<b>57.0</b>	70.8	<b>66.6</b>

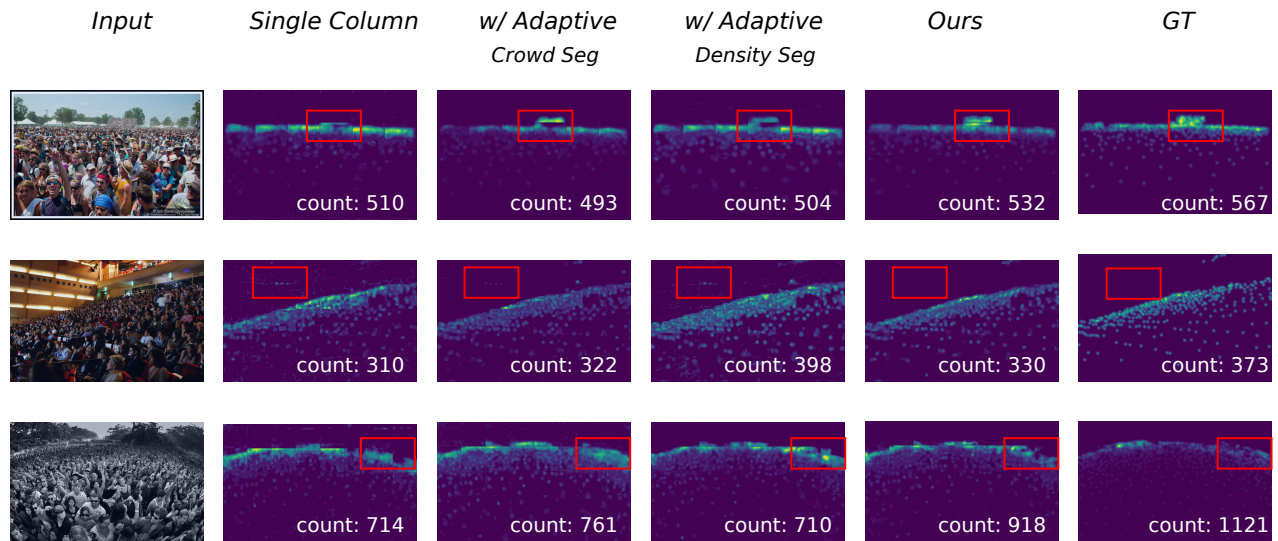


Fig. 8. The qualitative results of ablation studies about auxiliary tasks. The red bounding boxes are used for better visualization and comparison. *Ours* and *w/ Adaptive Crowd Seg* can know the crowd’s spatial regions (first and third rows), and filter out the background noise (second row). On the other hand, *Ours* and *w/ Adaptive Density Seg* can estimate more accurate density levels across the whole density maps (second and third rows).

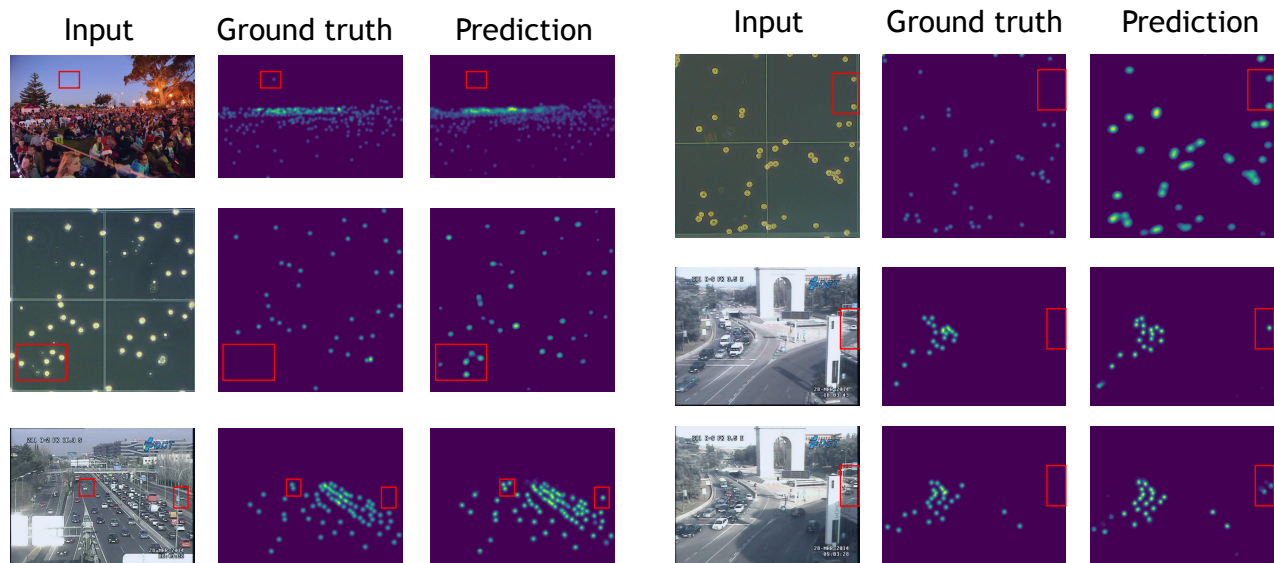


Fig. 9. Comparison of our predictions and the ground truth. Our predictions are robust when there are mislabeled or wrongly labeled point annotations in the ground truth of crowd counting, cell counting, and vehicle counting datasets, respectively. The red bounding boxes are used for better visualization and comparison.

#### D. Discussion: Comparison with Ground Truth

The underlying labeling errors (noisy ground truth) exist in most datasets due to the human annotators’ errors. However, a robust model can omit the noise ground truth during training and produce a more accurate prediction. This section showed that our model could indicate some mislabeled or wrongly labelled point annotations of the ground truth in the test dataset. This highlights the generalizability of our approach and the potential issue of the imperfect ground truth in object counting applications. Fig. 9 shows a wrongly labelled point annotation (top left) case of the crowd counting test dataset, and the other cases are mislabeled point annotation of vehicle

and cell counting test dataset. We highlighted the wrongly labelled or mislabeled area with red bounding boxes for better visualization and comparison.

## VI. CONCLUSION

We proposed a novel framework for auxiliary task learning based counting by employing an adaptively shared backbone, a *GCN* reasoning module and a novel dilated contrastive density loss function. Our model advocates task-shared and task-specified features to be learned simultaneously. The proposed method highlights that cross-domain reasoning in graph through *GCNs* using crowd segmentation and density level

segmentation can significantly improve feature learning in density map regression tasks. With our proposed loss function's regional density difference supervision, our model set a new state-of-the-art among auxiliary task learning based counting methods on seven challenging benchmarks.

## REFERENCES

- [1] G. Gao, J. Gao, Q. Liu, Q. Wang, and Y. Wang, "Cnn-based density estimation and crowd counting: A survey," *arXiv preprint arXiv:2003.12783*, 2020.
- [2] D. Onoro-Rubio and R. J. López-Sastre, "Towards perspective-free object counting with deep learning," in *European Conference on Computer Vision*. Springer, 2016, pp. 615–629.
- [3] Y. Guo, J. Stein, G. Wu, and A. Krishnamurthy, "Sau-net: A universal deep network for cell counting," in *Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, 2019, pp. 299–306.
- [4] M. K. K. Reddy, M. Rochan, Y. Lu, and Y. Wang, "Adacrowd: Unlabeled scene adaptation for crowd counting," *IEEE Transactions on Multimedia*, 2020.
- [5] Y. Meng, H. Zhang, Y. Zhao, X. Yang, X. Qian, X. Huang, and Y. Zheng, "Spatial uncertainty-aware semi-supervised crowd counting," *The IEEE International Conference on Computer Vision (ICCV)*, 2021.
- [6] L. Liu, J. Chen, H. Wu, T. Chen, G. Li, and L. Lin, "Efficient crowd counting via structured knowledge transfer," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 2645–2654.
- [7] Z. Shi, P. Mettes, and C. G. Snoek, "Counting with focus for free," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 4200–4209.
- [8] J. Liu, C. Gao, D. Meng, and A. G. Hauptmann, "Decidenet: Counting varying density crowds through attention guided detection and density estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5197–5206.
- [9] X. Jiang, L. Zhang, T. Zhang, P. Lv, B. Zhou, Y. Pang, M. Xu, and C. Xu, "Density-aware multi-task learning for crowd counting," *IEEE Transactions on Multimedia*, vol. 23, pp. 443–453, 2020.
- [10] D. Modolo, B. Shuai, R. R. Varior, and J. Tighe, "Understanding the impact of mistakes on background regions in crowd counting," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021, pp. 1650–1659.
- [11] J. Wan, W. Luo, B. Wu, A. B. Chan, and W. Liu, "Residual regression with semantic prior for crowd counting," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4036–4045.
- [12] S. Liu, E. Johns, and A. J. Davison, "End-to-end multi-task learning with attention," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1871–1880.
- [13] J. Gao, Y. Yuan, and Q. Wang, "Feature-aware adaptation and density alignment for crowd counting in video surveillance," *IEEE Transactions on Cybernetics*, 2020.
- [14] A. Luo, F. Yang, X. Li, D. Nie, Z. Jiao, S. Zhou, and H. Cheng, "Hybrid graph neural networks for crowd counting," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 11 693–11 700.
- [15] Y. Li and A. Gupta, "Beyond grids: Learning graph representations for visual recognition," in *Advances in Neural Information Processing Systems*, 2018, pp. 9225–9235.
- [16] Y. Meng, W. Meng, D. Gao, Y. Zhao, X. Yang, X. Huang, and Y. Zheng, "Regression of instance boundary by aggregated cnn and gcnn," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.
- [17] Y. Meng, M. Wei, D. Gao, Y. Zhao, X. Yang, X. Huang, and Y. Zheng, "Cnn-gcn aggregation enabled boundary regression for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2020.
- [18] Y. Zhang, D. Zhou, S. Chen, S. Gao, and Y. Ma, "Single-image crowd counting via multi-column convolutional neural network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 589–597.
- [19] W. Liu, M. Salzmann, and P. Fua, "Context-aware crowd counting," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5099–5108.
- [20] Y. Liu, Q. Wen, H. Chen, W. Liu, J. Qin, G. Han, and S. He, "Crowd counting via cross-stage refinement networks," *IEEE Transactions on Image Processing*, vol. 29, pp. 6800–6812, 2020.
- [21] X. Jiang, L. Zhang, M. Xu, T. Zhang, P. Lv, B. Zhou, X. Yang, and Y. Pang, "Attention scaling for crowd counting," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 4706–4715.
- [22] Q. Wang, J. Gao, W. Lin, and X. Li, "NWPU-Crowd: A large-scale benchmark for crowd counting and localization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.
- [23] L. Boominathan, S. S. Kruthiventi, and R. V. Babu, "Crowdnet: A deep convolutional network for dense crowd counting," in *Proceedings of the 24th ACM international conference on Multimedia*, 2016, pp. 640–644.
- [24] D. Babu Sam, N. N. Sajjan, R. Venkatesh Babu, and M. Srinivasan, "Divide and grow: Capturing huge diversity in crowd images with incrementally growing cnn," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 3618–3626.
- [25] Y. Yang, G. Li, Z. Wu, L. Su, Q. Huang, and N. Sebe, "Reverse perspective network for perspective-aware object counting," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 4374–4383.
- [26] Z. Ma, X. Wei, X. Hong, and Y. Gong, "Learning scales from points: A scale-aware probabilistic model for crowd counting," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 220–228.
- [27] M. Zhao, C. Zhang, J. Zhang, F. Porikli, B. Ni, and W. Zhang, "Scale-aware crowd counting via depth-embedded convolutional neural networks," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 10, pp. 3651–3662, 2019.
- [28] Y. Zhou, J. Yang, H. Li, T. Cao, and S.-Y. Kung, "Adversarial learning for multiscale crowd counting under complex scenes," *IEEE transactions on cybernetics*, 2020.
- [29] L. Liu, Z. Qiu, G. Li, S. Liu, W. Ouyang, and L. Lin, "Crowd counting with deep structured scale integration network," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 1774–1783.
- [30] L. Liu, H. Wang, G. Li, W. Ouyang, and L. Lin, "Crowd counting using deep recurrent spatial-aware network," in *IJCAI*, 2018.
- [31] Y. Tian, Y. Lei, J. Zhang, and J. Z. Wang, "Padnet: Pan-density crowd counting," *IEEE Transactions on Image Processing*, vol. 29, pp. 2714–2727, 2019.
- [32] X. Chen, Y. Bin, N. Sang, and C. Gao, "Scale pyramid network for crowd counting," in *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2019, pp. 1941–1950.
- [33] Y. Li, X. Zhang, and D. Chen, "Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 1091–1100.
- [34] Z. Yan, Y. Yuan, W. Zuo, X. Tan, Y. Wang, S. Wen, and E. Ding, "Perspective-guided convolution networks for crowd counting," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 952–961.
- [35] Z. Yan, R. Zhang, H. Zhang, Q. Zhang, and W. Zuo, "Crowd counting via perspective-guided fractional-dilation convolution," *IEEE Transactions on Multimedia*, 2021.
- [36] J. Wan, Q. Wang, and A. B. Chan, "Kernel-based density map generation for dense object counting," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.
- [37] S. Bai, Z. He, Y. Qiao, H. Hu, W. Wu, and J. Yan, "Adaptive dilated network with self-correction supervision for counting," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 4594–4603.
- [38] Y. Miao, Z. Lin, G. Ding, and J. Han, "Shallow feature based dense attention network for crowd counting," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 11 765–11 772.
- [39] A. Zhang, L. Yue, J. Shen, F. Zhu, X. Zhen, X. Cao, and L. Shao, "Attentional neural fields for crowd counting," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 5714–5723.
- [40] B. Chen, Z. Yan, K. Li, P. Li, B. Wang, W. Zuo, and L. Zhang, "Variational attention: Propagating domain-specific knowledge for multi-domain learning in crowd counting," *The IEEE International Conference on Computer Vision (ICCV)*, 2021.
- [41] H. Duan, S. Wang, and Y. Guan, "Sofa-net: Second-order and first-order attention network for crowd counting," *BMVC*, 2020.

- [42] Q. Wang, W. Lin, J. Gao, and X. Li, "Density-aware curriculum learning for crowd counting," *IEEE Transactions on Cybernetics*, 2020.
- [43] J. Wan and A. Chan, "Adaptive density map generation for crowd counting," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 1130–1139.
- [44] Y. Yang, G. Li, D. Du, Q. Huang, and N. Sebe, "Embedding perspective analysis into multi-column convolutional neural network for crowd counting," *IEEE Transactions on Image Processing*, vol. 30, pp. 1395–1407, 2020.
- [45] J. Cheng, H. Xiong, Z. Cao, and H. Lu, "Decoupled two-stage crowd counting and beyond," *IEEE Transactions on Image Processing*, vol. 30, pp. 2862–2875, 2021.
- [46] S. Abousamra, M. Hoai, D. Samaras, and C. Chen, "Localization in the crowd with topological constraints," in *AAAI*, 2021.
- [47] Q. Song, C. Wang, Y. Wang, Y. Tai, C. Wang, J. Li, J. Wu, and J. Ma, "To choose or to fuse? scale selection for crowd counting," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 3, 2021, pp. 2576–2583.
- [48] L. Liu, J. Chen, H. Wu, G. Li, C. Li, and L. Lin, "Cross-modal collaborative representation learning and a large-scale rgbt benchmark for crowd counting," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 4823–4833.
- [49] Q. Zhang, W. Lin, and A. B. Chan, "Cross-view cross-scene multi-view crowd counting," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 557–567.
- [50] Y. Lei, Y. Liu, P. Zhang, and L. Liu, "Towards using count-level weak supervision for crowd counting," *Pattern Recognition*, vol. 109, p. 107616, 2021.
- [51] J. Wan, N. S. Kumar, and A. B. Chan, "Fine-grained crowd counting," *IEEE transactions on image processing*, vol. 30, pp. 2114–2126, 2021.
- [52] W. Liu, M. Salzmann, and P. Fua, "Counting people by estimating people flows," *IEEE transactions on pattern analysis and machine intelligence*, 2021.
- [53] —, "Estimating people flows to better count them in crowded scenes," in *European Conference on Computer Vision*. Springer, 2020, pp. 723–740.
- [54] V. A. Sindagi and V. M. Patel, "Generating high-quality crowd density maps using contextual pyramid cnns," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 1861–1870.
- [55] V. Sindagi and V. Patel, "Ha-ccn: Hierarchical attention-based crowd counting network," *IEEE Transactions on Image Processing*, vol. 29, pp. 323–335, 2019.
- [56] T. Zhou, L. Zhang, D. Jiawei, X. Peng, Z. Fang, Z. Xiao, and H. Zhu, "Locality-aware crowd counting," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.
- [57] X. Liu, J. Yang, W. Ding, T. Wang, Z. Wang, and J. Xiong, "Adaptive mixture regression network with local counting map for crowd counting," in *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIV 16*. Springer, 2020, pp. 241–257.
- [58] H. Mo, W. Ren, Y. Xiong, X. Pan, Z. Zhou, X. Cao, and W. Wu, "Background noise filtering and distribution dividing for crowd counting," *IEEE Transactions on Image Processing*, vol. 29, pp. 8199–8212, 2020.
- [59] M. Zhao, J. Zhang, C. Zhang, and W. Zhang, "Leveraging heterogeneous auxiliary tasks to assist crowd counting," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 12736–12745.
- [60] Q. Wang, J. Gao, W. Lin, and Y. Yuan, "Pixel-wise crowd understanding via synthetic data," *International Journal of Computer Vision*, vol. 129, no. 1, pp. 225–245, 2021.
- [61] D. B. Sam, S. V. Peri, M. N. Sundararaman, A. Kamath, and V. B. Radhakrishnan, "Locate, size and count: Accurately resolving people in dense crowds via detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.
- [62] C. Liu, X. Weng, and Y. Mu, "Recurrent attentive zooming for joint crowd counting and precise localization," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2019, pp. 1217–1226.
- [63] D. Lian, J. Li, J. Zheng, W. Luo, and S. Gao, "Density map regression guided detection network for rgb-d crowd counting and localization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1821–1830.
- [64] Y. Wang, J. Hou, X. Hou, and L.-P. Chau, "A self-training approach for point-supervised object detection and counting in crowds," *IEEE Transactions on Image Processing*, vol. 30, pp. 2876–2887, 2021.
- [65] W. Ren, X. Wang, J. Tian, Y. Tang, and A. B. Chan, "Tracking-by-counting: Using network flows on crowd density maps for tracking multiple targets," *IEEE Transactions on Image Processing*, vol. 30, pp. 1439–1452, 2020.
- [66] X. Liu, J. Van De Weijer, and A. D. Bagdanov, "Exploiting unlabeled data in cnns by self-supervised learning to rank," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 8, pp. 1862–1878, 2019.
- [67] —, "Leveraging unlabeled data for crowd counting by learning to rank," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7661–7669.
- [68] V. S. Sravya, P. K. Mansi, B. Divij, R. Ganesh, and K. S. Ravi, "Wisdom of (binned) crowds: A bayesian stratification paradigm for crowd counting," in *Proceedings of the 2021 ACM Conference on Multimedia*. China: ACM, 2021.
- [69] Q. Song, C. Wang, Z. Jiang, Y. Wang, Y. Tai, C. Wang, J. Li, F. Huang, and Y. Wu, "Rethinking counting and localization in crowds: A purely point-based framework," *The IEEE International Conference on Computer Vision (ICCV)*, 2021.
- [70] J. Wan, Z. Liu, and A. B. Chan, "A generalized loss function for crowd counting and localization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 1974–1983.
- [71] Z. Ma, X. Wei, X. Hong, and Y. Gong, "Bayesian loss for crowd count estimation with point supervision," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 6142–6151.
- [72] J. Wan and A. Chan, "Modeling noisy annotations for crowd counting," *Advances in Neural Information Processing Systems*, vol. 33, 2020.
- [73] B. Wang, H. Liu, D. Samaras, and M. H. Nguyen, "Distribution matching for crowd counting," *Advances in Neural Information Processing Systems*, vol. 33, 2020.
- [74] C. Wang, Q. Song, B. Zhang, Y. Wang, Y. Tai, X. Hu, C. Wang, J. Li, J. Ma, and Y. Wu, "Uniformity in heterogeneity: Diving deep into count interval partition for crowd counting," *The IEEE International Conference on Computer Vision (ICCV)*, 2021.
- [75] Y.-J. Ma, H.-H. Shuai, and W.-H. Cheng, "Spatiotemporal dilated convolution with uncertain matching for video-based crowd estimation," *IEEE Transactions on Multimedia*, 2021.
- [76] V. Lempitsky and A. Zisserman, "Learning to count objects in images," in *Advances in Neural Information Processing Systems*, 2010, pp. 1324–1332.
- [77] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *International Conference on Learning Representations*, 2015.
- [78] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proceedings of the 32nd International Conference on Machine Learning*, 2015, pp. 448–456.
- [79] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7794–7803.
- [80] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," *ICLR*, 2017.
- [81] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
- [82] H. Idrees, M. Tayyab, K. Athrey, D. Zhang, S. Al-Maadeed, N. Rajpoot, and M. Shah, "Composition loss for counting, density map estimation and localization in dense crowds," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 532–546.
- [83] V. A. Sindagi, R. Yasarla, and V. M. Patel, "Pushing the frontiers of unconstrained crowd counting: New dataset and benchmark method," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 1221–1231.
- [84] M. Marsden, K. McGuinness, S. Little, C. E. Keogh, and N. E. O'Connor, "People, penguins and petri dishes: Adapting object counting models to new visual domains and object types without forgetting," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8070–8079.
- [85] R. Guerrero-Gómez-Olmedo, B. Torre-Jiménez, R. López-Sastre, S. Maldonado-Bascón, and D. Onoro-Rubio, "Extremely overlapping vehicle counting," in *Iberian Conference on Pattern Recognition and Image Analysis*. Springer, 2015, pp. 423–431.
- [86] I. Loshchilov and F. Hutter, "SGDR: Stochastic gradient descent with warm restarts," *ICLR*, 2017.