Transportation Object Counting with Graph-Based Adaptive Auxiliary Learning

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

***Abstract*—This paper proposes an adaptive auxiliary task learning-based approach for transport object counting problems such as humans and vehicles. These problems are essential in many real-world tasks such as video surveillance, traffic monitoring, public security, and urban planning, to aid intelligent transportation systems. 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 that are learned in an end-to-end manner. The network seamlessly combines a standard Convolution Neural Network (*CNN*) and a 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 features across different task branches of the adaptive *CNN* backbone. The whole framework pays special attention to 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, along with strengthened robustness. Experiments on six challenging multi-domain datasets demonstrate that our method achieves superior performance compared with state-of-the-art auxiliary task learning-based counting methods. Our code is publicly available [1].**

***Index Terms*—Object 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 human crowd counting problems whilst being applicable to other domains such as vehicle counting. Due to the occurrence of crowd gatherings 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, transportation management *etc.* [1].

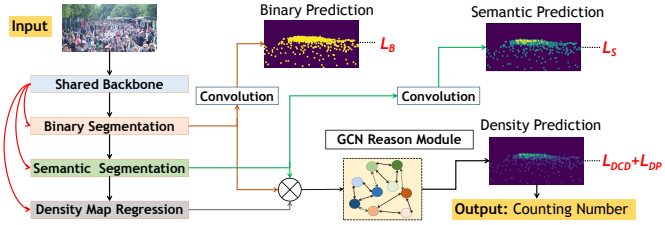

Fig. 1. Overview of the proposed network structure in the scene of crowd counting. An attention-enhanced adaptively shared backbone network is proposed to enable both task-shared and task-tailored features learning. A novel Graph Convolution Network (*GCN*) reasoning module is introduced to tackle issues of cross-granularity feature 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.

As a result of Convolutional Neural Network's (*CNN*)'s exceptional feature learning capability, the performance of crowd counting methods has been steadily enhanced. Recent state-of-the-art methods, such as [2], [3], have demonstrated that a density map regression paradigm yields satisfactory results. In these methods, given an input image, a *CNN*-based network is used to regress the corresponding density map; the sum of the pixel values in the density map represents the total number of counts in the image. There are a number of challenging issues [1] such as significant scale changes, wide variations in density levels, and complex scene backgrounds, however, there is still considerable room for counting performance improvement. Some previous methods [4], [5], [6], [7] rely on various types of information granularity in terms of 'auxiliary task learning' to address these issues. Using a single shared backbone network structure, these methods extract generalised features for all tasks. Unfortunately, this strategy may result in under-fitting, as the generalizable representation is frequently incapable of describing the comprehensive cross-granularity features across multiple tasks simultaneously [1]. Contrasting, our adaptive shared backbone network focuses on maximising the principal density map regression task and multi-granularity information augmentation from auxiliary tasks. Our backbone

Y. Meng, J. Bridge and M. Joddrell are with the Department of Eye and Vision Sciences, University of Liverpool, Liverpool, L7 8TX, United Kingdom. The work of Y. Meng is supported by a Remark AI UK Limited studentship. The work of J. Bridge is partially supported by an EPSRC studentship (No. 2110275). The work of M Joddrell is partially supported by a Huawei studentship.

Y. Zhao is with The Affiliated People's Hospital of Ningbo University, Ningbo, China. He is also with Cixi Institute of Biomedical Engineering, Ningbo Institute of Materials Technology and Engineering, Chinese Academy of Science, Ningbo, China. The work of Y. Zhao is partially supported by the Youth Innovation Promotion Association CAS (2021298) and the Ningbo major science and technology task project (2021Z054).

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.

Y. Zheng is with the Department of Eye and Vision Sciences, University of Liverpool, Liverpool, L7 8TX, United Kingdom. He is also with Liverpool Centre for Cardiovascular Science, University of Liverpool and Liverpool Heart & Chest Hospital, Liverpool, UK. Corresponding author: Yalin Zheng (yalin.zheng@liverpool.ac.uk).

[1]https://github.com/smallmax00/Counting_With_Adaptive_Auxiliary

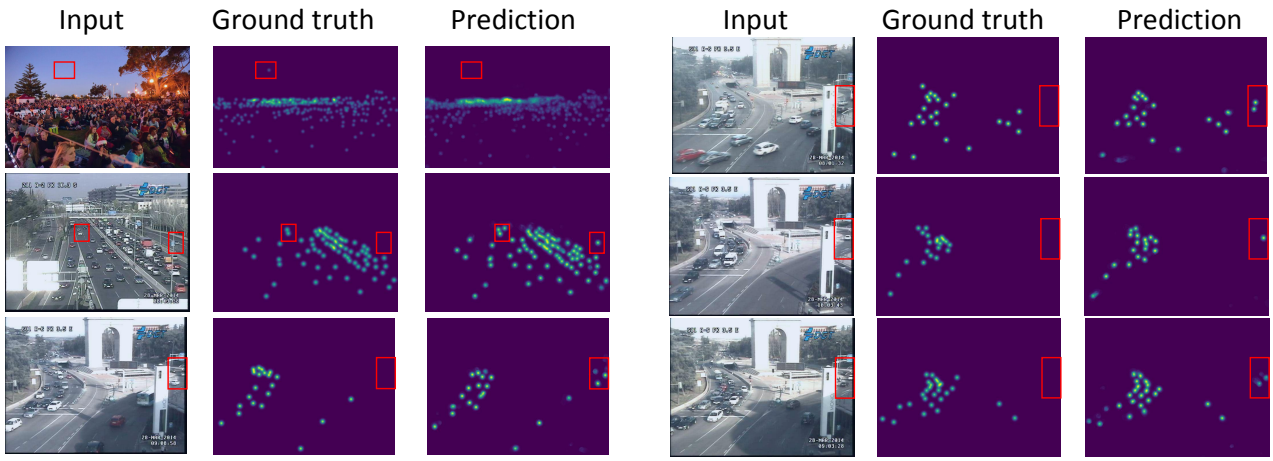

Fig. 2. Comparison of our predictions and the ground truth. Our predictions are robust enough even when there are mislabeled or incorrectly labeled point annotations in the ground truth of crowd counting and vehicle counting datasets. Our model can indicate more accurate object locations or counting numbers compared with the ground truth. The red bounding boxes are used for better visualisation and comparison.

network has a multi-level information aggregation mechanism to repeatedly and hierarchically combine features from distinct stages and auxiliary branches. 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 from the density map regression ground truth. Intuitively, no increase in information from the ground truth of auxiliary tasks is generated; however, the information is enhanced and specified through auxiliary tasks in terms of different density information granularity.

Given the auxiliary-task learning paradigm, we researched how to reason and fuse features from different tasks for density map regression. Crowd segmentation and density level segmentation feature domains have different granularity of representations. Direct fusion (element-wise multiplication or channel-wise concatenation) of three task branches' outputs might cause domain conflicts [8]. To improve counting accuracy, we exploited the nature of Graph Convolutional Networks (*GCN*) for information reasoning. *GCN* has showed promising reasoning ability on several computer vision problems, including scene interpretation [9], [3] and image segmentation [10], [11], [12], [13], [14], [15], but has been rarely investigated in crowd counting. Our model projects a collection of pixels from a spatial-aware density feature map 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 [8], which directly treated cross-granularity feature maps as graph vertices and utilized a cascaded Graph Neural Network (GNN) to reason the cross-scale relationships.

In this work we present a novel loss function for density map regression. The commonly adopted Least Absolute Error (L1) or Least Square Error (L2) loss [16], [4], [17] assumes pixel-wise independence. However, two major flaws exist: (1) The estimated density map is over-smoothed [5], underestimating high-density regions and overestimating low-density parts. The model may focus on reducing count mistakes rather than regressing high-quality density maps, therefore it cannot reflect the true density levels. (2) Without a large receptive field, pixel-wise loss functions may ignore regional density level information during training [18]. Unbalanced low- and high-level density distributions might cause bias in training, reducing network resiliency. To overcome these concerns, we present a new loss function for density map regression called Dilated Contrastive Density Loss (L_{DCD}), where the density difference between dilated adjacent pixels provides extra regional supervision. Ablation studies conducted show that our proposed regional loss function outperforms pixel-wise losses in all datasets used in this work.

We conducted extensive experiments on seven well-known challenging counting benchmarks. Quantitative and qualitative results demonstrate that our model achieves state-of-the-art performance. To the best of our knowledge, we achieved the **best** counting performance among other auxiliary task-based counting methods on the NWPU-Crowd [19] benchmark [2], which is currently the largest crowd counting benchmark. Our model is robust and generalizable, indicating incorrectly labeled or mislabeled object ground truths in the test datasets. Please refer to Fig 2 for more details.

In summary, this work makes the following contributions:

- We address the feature learning issues of the backbone network for auxiliary task-based methods in crowd counting challenges, by enabling task-shareable and task-specified feature learning simultaneously with a primary focus on the main task.
- We propose 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-granularity feature relations between density map regression and other auxiliary tasks.

[2]https://www.crowdbenchmark.com/nwpucrowd.html

- We propose a novel loss function tailored for density map regression, strengthening the network's generalizability and improving the counting accuracy.

II. Related Work

In recent years, density map regression-based counting methods [20], [21], [22], [23], [24], [25], [26], [27], [17], [28], [29], [30], [31], [32], [33], [34] using *CNNs* have achieved good performance. As mentioned previously, they employ different learning strategies to address difficult issues such as variations in scale, alternate density levels, and complicated background scenes. Specifically, attention-based methods [35], [18], [36], [37], [38], [39], [33], [40], [41], [42], auxiliary task-based methods [43], [44], [45], [46], [47], [48], [49], [50], [6], [51], [52], and different supervision-based methods [53], [54], [55], [56], [57], [58], [59], [60], [61], [62], [63], [64] align closely with our proposed method presented in this work. We have elaborated the related works of the aforementioned learning strategies in the following contents.

A. Attention-Based Counting

Visual attention mechanisms were applied among several works [35], [18], [36], [37], [38], [39], [33], [40], [41], [42] in crowd counting applications, which helps the network focus on valuable information and addresses several challenges. For example, *Miao et al.* [35] utilized a shallow feature-based attention module to highlight the regions of crowd interest and filter out the noise from background clutter. To tackle various density levels issues, *Jiang et al.* [18] employed an attention mask to refine the density map for adapting to different density levels. Furthermore, *Zhang et al.* [36] proposed the *Attention Neural Field* that incorporates non-local attention modules with conditional random fields to maintain multi-scale features and long-range dependencies, enabling control over the large-scale variation challenge of input crowd images. *Wan et al.* [65], [33] exploited the self-attention mechanism to adaptively generate density maps with different Gaussian kernel sizes, which is then used as the ground truth to supervise the model. The aforementioned methods adopt the attention mechanism as a feature enhancement module to implicitly address the crowd counting task challenges emphasised throughout this paper, including notable scale changes, large-scale density level variability, and complex scene backgrounds. Our model explicitly addresses these challenges through auxiliary tasks. On the other hand, our model adopts the attention mechanism to construct an adaptively shared backbone network, enabling task-shared and task-specific feature learning simultaneously.

B. Auxiliary Task-Based Counting

Recently, auxiliary task learning-based counting methods [43], [44], [45], [46], [47], [48], [49], [50], [6], [51], [52], [66], [67], [68] have attracted research attention because of their ability to capture extra granularity information and contextual dependencies for density map regression. Most methods utilize 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 [4], [69], [70], [6], [71], [72], [73] can enhance patch-wise density-level information, which helps to address the underestimation and overestimation problems of density map regression. 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-wise and patch-wise feature learning. In contrast, our model proposes a density level segmentation auxiliary task, which can be regarded as the 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 appearances, the crowd segmentation task, adopted by previous methods [74], [4], [8], [7], [75], can provide spatial location information for the crowd, which highlights the foreground over the background and guides the network focus onto the region of interest. Our model also adopts the crowd segmentation task because of its superiority in spatial location information enhancement. In particular, *Luo et al.* [8] adopted crowd segmentation as the auxiliary task, then proposed a cascaded graph-based model to tackle the fusion of features between the crowd segmentation and density map regression tasks. This is similar to our learning paradigm, however, there are two significant difference: (1) They did not consider the density level information and only treated the features of the density map and crowd segmentation as the vertices in their proposed model. Alternatively, we incorporate the spatial information of crowd location, the semantic information of density level, and the main task of density map features, into the proposed vertices in our model. (2) They treated the vertices equally. Specifically, they regarded the crowd segmentation and density map features as independent vertices, fusing and aggregating the information among them. However, the main task to estimate the counting number should be density map regression, hence they may introduce inevitable noise into the training process if the auxiliary task takes over. Differently, we project a collection of pixels from a spatial-aware density feature map with similar density levels to each graph vertex, thereby enhancing the main task vertices' spatial location awareness. Also, we project the long-range density level dependency among every pixel into the adjacency matrix, boosting the main task vertices' semantic density level awareness. Please see Section III-E and Fig. 5 for details.

C. Learn to Count with Different Supervisions

Instead of tackling the counting task through different learning frameworks or strategies, recent methods [53], [54], [55], [56], [57], [58], [59], [60], [61], [62], [63], [64], [76], [77], [78] have paid attention to the way of supervisions. For example, *Sravya et al.* proposed a bin loss [55] to enable the data distribution-aware optimization, which helped to address the domain variation challenges from different crowd data

sources. *Song et al.* [56] studied the counting problem in a different way, where a combination of *Euclidean* loss and *Cross Entropy* loss was used for point location learning, instead of density map regression. Along the same line, *Bayesian* loss was proposed by [58] to provide more reliable supervisions at each annotated point. Alternatively, *Wan et al.* [57] studied the combination of pixel-wise loss and point-wise loss, which investigated the density map representation through an unbalanced optimal transport problem. [59] 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, [60] proposed a distribution matching loss to tackle the weakened generalizability of Gaussian smoothed density maps. Moreover, *Wang et al.* [61] treated the counting with density maps 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 a model, such as point locations, bounding boxes, matching, ranking, classification, *etc.*. However, the mainstream counting methods still rely on pixel-wise supervision with the density map ground truth [1], such as the *L1* or *L2* loss functions. In this work, we propose a Dilated Contrastive Density Loss (L_{DCD}) to improve the pixel-wise loss' receptive field and to increase the regional supervision.

III. Methodology

A. Ground Truth Generation

Following [79], 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 the pixel in the mask $\{B_i\}_{i=1}^N$ is set to 1 if its pixel value in the density map is larger than zero, otherwise it is set to 0 .

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:

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

where L represents the overall levels of density. Following previous patch-based density level classification methods [4], [6], we set L equal to 4 in our work. D_i is the pixel value in the i_{th} density map ground truth. Specifically, given a density map and Eq. 1, we can generate the density level map with *L* levels of object density. In other words, we set all the pixels of the density map into *L* categories or classes according to their own pixel value. In this way, each pixel is assigned to a semantic label to represent the high-level sparseness or denseness.

B. Task Adaptive Backbone Network

Intuitively, our motivation is that the backbone network should be able to produce both universal (or generic) and specialised features that are applicable to all tasks and can also be tailored to specific tasks. To this end, instead of using a shared backbone network to extract generalizable features for different tasks, we propose an auxiliary-task based adaptive 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. 3 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 [69], [8], [80], [6], *etc.*, the truncated VGG-16 [81] 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. V. The shared backbone adopts the first 13 layers of VGG-16 to extract multi-level features. To exploit the global contextual dependencies, we propose a Feature Fuse Block (*FFB*), which aggregates and fuses 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 high-level features to low-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. 3. Each attention module consists of a global average pooling (GAP) layer to capture global context through different feature map channels, generating an attention tensor to lead the emphasis of feature learning. Then, two blocks with a convolutional layer followed by a Batch Normalization *(BN)* [82] layer with *ReLu* and sigmoid as the activation functions are added. For the convolutional layer filter, the kernel size is 1×1. Element-wise multiplication is then performed between the outputs of a 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 introduce a cross-granularity feature fusing operator in a particular order to focus on optimizing the density map regression task. Specifically, the crowd segmentation 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.

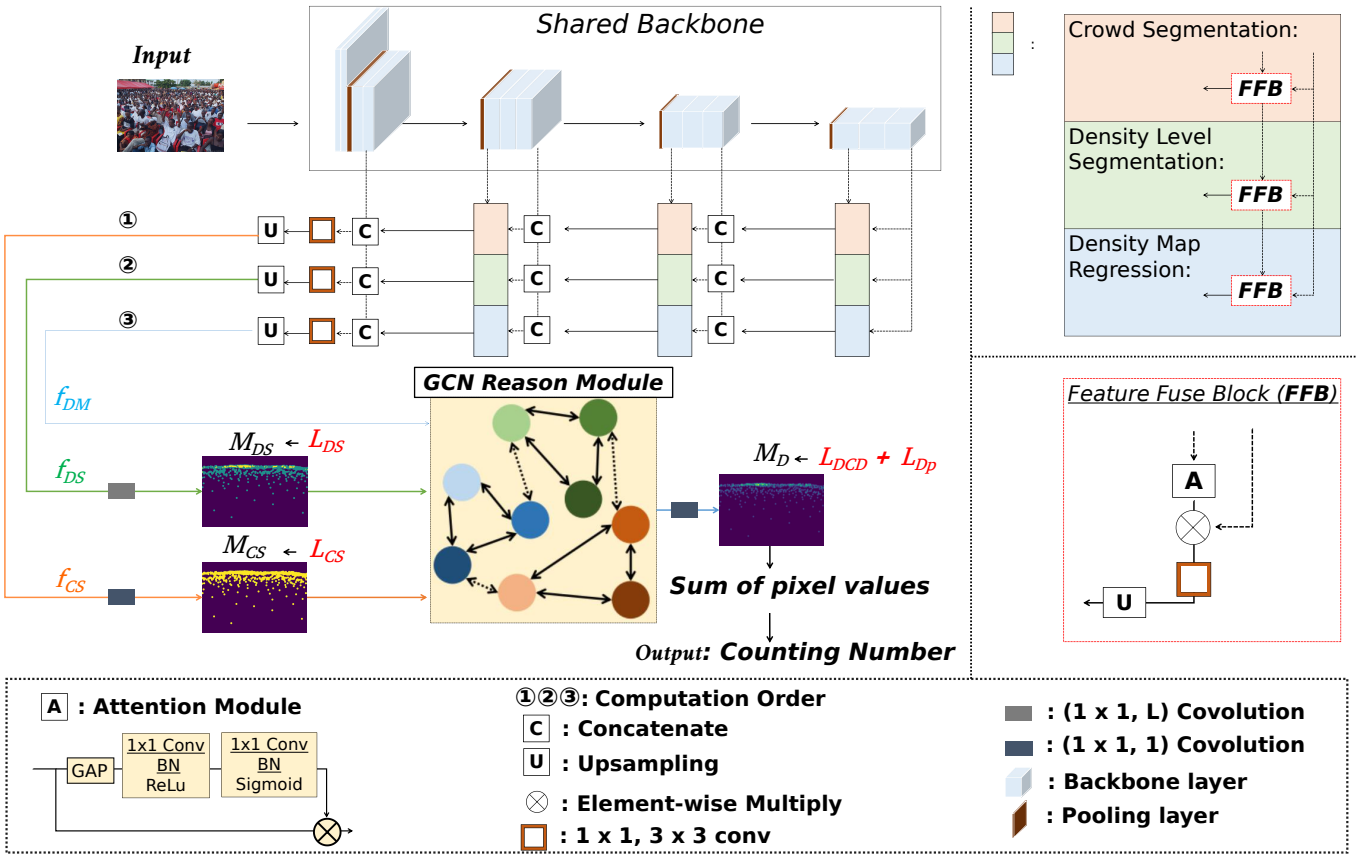

Fig. 3. Illustration of our proposed network. The adaptively shared backbone network has three outputs of f_{CS}, f_{DS}, f_{DM}, representing crowd segmentation, density level segmentation, and density map regression branches' output feature map, respectively. The order of their involvements indicates that the density map regression branch can benefit from the extra density level and crowd spatial supervision from the other two branches gradually.

C. Auxiliary Tasks

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

Crowd Segmentation. We introduce crowd segmentation as one of the auxiliary tasks for two reasons. Firstly, the pixel value of the density map should be zero in areas devoid of people. However, the predicted density map can be inaccurate and noisy when the background is cluttered and complex. The task of crowd segmentation provides a spatial focus to the density map regression procedure by setting the pixel values of non-crowd regions to zero. Secondly, given the standard setup of single density map regression, pixels within a specific range of the point annotations should contribute more to the final counting results; however, most irrelevant pixels dominate the loss [1]. In order to circumvent this constraint, crowd segmentation can provide additional information enhancement in terms of the spatial indicator via 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 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: $M_{CS} = Sigmoid(\theta_{CS}, f_{CS}) \in \mathbb{R}^{1\times H\times W}$. Fig. 4 demonstrates an example of the location map, which is the M_{CS} after using 0.5 as the thresholding, resulting in a binary map. The colors represent different classes, where there is a foreground class and background class. Crowd segmentation focuses on the spatial information, and indicates the geometry-aware supplementary as the auxiliary task.

Density Level Segmentation. Density map regression is a pixel-wise task that focuses on the learning of low-level features but may disregard high-level semantic information, such as the density level information [28]. However, such semantic information is critical in the counting system because the density map's pixel values should rely not solely on their own pixel-wise characteristics but also on regions with varying densities [17]. To address the issues, we perform density level segmentation as another auxiliary task. Compared with previous patch-based density level classification methods [4], [69], [70], [6], our proposed pixel-based density level segmentation can provide pixel level density information and high-level semantic features at the same time. Fig. 4 demonstrates an example of the density level map, where colors represent different classes. From class 3 down to class 0, the density level decreases. Density level segmentation focuses on the semantic information, and indicates the density level-aware

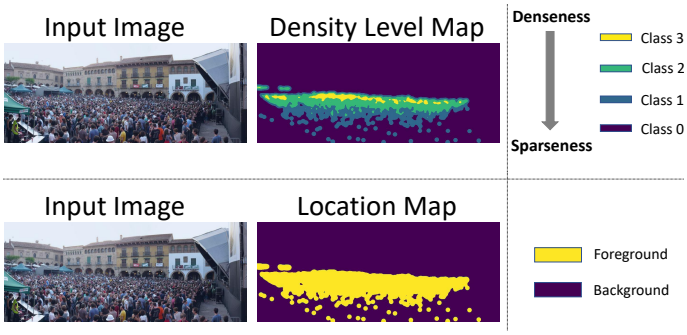

Fig. 4. Example of the density level map (top) and location map (bottom). For the density level, the colors represent different classes, which corresponds to different density levels. From class 3 down to class 0, the density level decreases from denseness to sparseness. The class 0 represents the background, where there is no objects. As for the location map, the colors represent the different classes, where there is a foreground class and a background class.

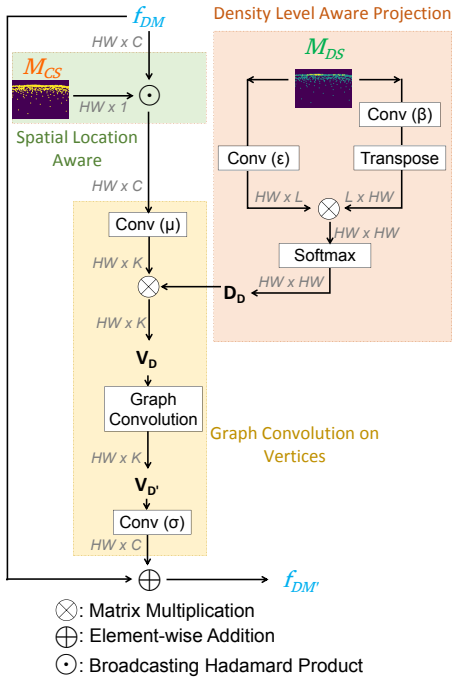

Fig. 5. Architecture of the proposed *GCN* reasoning module. $f_{DM} \in \mathbb{R}^{C\times H\times W}$ is the feature map of the density map regression branch, $C = 32$ is the channel size; $M_{CS} \in \mathbb{R}^{1\times H\times W}$ is the prediction of the crowd segmentation branch; $M_{DS} \in \mathbb{R}^{L\times H\times W}$ 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.

supplementary as the auxiliary task. 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 activation function are applied. The prediction of the density level segmentation branch M_{DS} is defined as: $M_{DS} = softmax(\theta_{DS}, f_{DS}) \in \mathbb{R}^{L\times H\times W}$, 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 analysed for fusion into a combined reasoned feature to feed to 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 feature map derived from the density map branches $f_{DM} \in \mathbb{R}^{C\times H\times W}$ into the *GCN* reasoning module to understand 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 a 1×1 convolution layer with a *ReLU* activation function.

E. GCN Reasoning Module

Deep feature extraction and fusion have been explored in previous studies, such as discriminant correlation analysis [83], [84], and multi-canonical correlation analysis [85], [86], [87], where they adaptively selected and fused CNN features from different layers, such that resulting representations have a high linear correlation. Following the same line, we propose a GCN model to fuse the correlated and supplementary features from auxiliary tasks that contribute to the counting task.

Different granularity representations are utilised for the crowd segmentation and density level segmentation feature domains. Direct fusion (element-wise multiplication or channel-wise concatenation) of the outputs of three task branches may lead to domain conflicts [8]. Our *GCN* reason model projects a collection of pixels from a spatial-aware density feature map with similar density levels to each graph vertex and exploits a *GCN* to reason about the relations among graph vertices. In other words, our graph is formed with fused-information of spatial locations and density levels from auxiliary tasks via initialising the adjacency matrix and vertices (D_D and V_D shown in Fig. 5). The proposed *GCN* reasoning module structure is shown in Fig. 5. In detail, there are three primary modules: *Spatial Location Aware* module, *Density Level Aware Projection* module, *Graph Convolution on Vertices* module.

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 for this: (1) M_{CS} is a one-channel crowd segmentation map, with encoded probabilities of the non-crowd regions' pixel values approaching zero and crowd regions' pixel values approaching one; the value of one serves as a filter to zero out the non-crowd region's pixel value of the density map. (2) the broadcasting Hadamard Product can achieve crowd spatial awareness for every channel of f_{DM} through zeroing out the non-crowd region's pixel value. This addresses the challenge of complex scene backgrounds in crowd images.

Density Level Aware Projection Module. As mentioned above, the pixel-wise density level information can help to address the challenges of 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 [8]. We exploited the nature of *GCN* and projected the density level information into the graph vertices for further reasoning, which benefited the long-range relationship reasoning ability of *GCN* and the multi-granularity information enhancement from density level. Inspired by the non-local module [88], 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:

$$D_D = softmax\Big(\epsilon(M_{DS}) \otimes \beta^{\mathsf{T}}(M_{DS})\Big), \tag{2}$$

where Conv β and Conv ϵ are two convolution layers with 1×1 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 larger weights. The dependence matrix might itself reflect the pixel-by-pixel density level dependency. In addition, with Eq. 2, we projected the density level map as a precondition to the graph domain via matrix multiplication, which simultaneously improves high-level semantic dependence.

Graph Convolution on Vertices. In this module, we learnt how to reason the region-based relationship in the density map through *GCN* in graph domain. Formally, the constructed vertices V_D is defined as:

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

where $\otimes$ is matrix multiplication; $\odot$ is the broadcasting Hadamard product. Specifically in Eq. 3, we projected the spatial aware feature map of f_{DM} into graph domain with K vertices, and each vertex is represented by an embedding of shape $H \times W$. This is achieved by Conv (μ), which is a 1×1 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 have similar density levels to graph vertices, where each vertex represents a region in the crowd image. With the constructed vertices (V_D), the long-range region-wise relationship is further reasoned in the graph domain through *GCN*. Formally, the output vertices of our proposed *GCN* ($V_{D'}$) are calculated as:

$$V_{D'} = ReLU\Big((I - A) \otimes V_D \otimes W_D\Big), \tag{4}$$

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 along the training process. The identity matrix I serves as a residual connection that alleviates the optimization difficulties. Specifically, in Eq. 4, 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 [89], resulting the output vertex features $V_{D'} \in \mathbb{R}^{K \times HW}$. Based on the learned graph, the information propagated 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 a 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 Conv (σ), which is a 1×1 convolution layer. Finally, we summed up the re-projected and the original density feature maps to form the final feature map. The final pixel-wise density feature map $f_{DM'}$ is thus computed as: $f_{DM'} = f_{DM} + \sigma(V_{D'})$. This can be regarded as the residual connection.

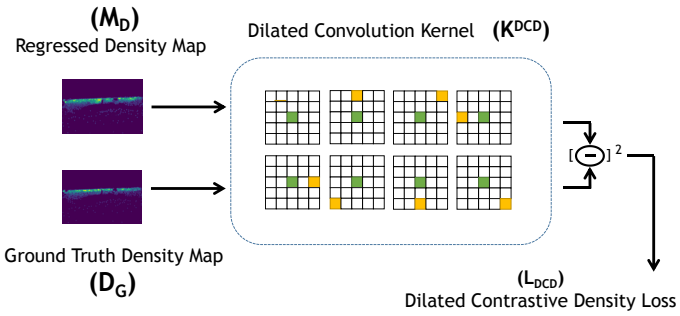

Fig. 6. Dilated Contrastive Density Loss (L_{DCD}). There are eight dilated contrastive kernels with green, white, yellow blocks representing 1, 0, -1, respectively. The least-square error of two outputs from the regression and ground truth is treated as the final L_{DCD}.

F. Loss Function

The whole network is end-to-end trainable, which includes four loss functions. The total loss function is defined in Eq. 5 as follows:

$$L_{total} = L_{CS} + L_{DS} + \gamma \cdot (L_{Dp} + L_{DCD}), \tag{5}$$

where γ is empirically set as 2, which is a hyper-parameter to trade-off between the auxiliary losses and main loss. Please note that extensive experiments have been conducted to determine the weights of the losses for the two auxiliary tasks. 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 the crowd segmentation auxiliary task; categorical cross-entropy (L_{DS}) is used for the density level segmentation auxiliary task; *L2* loss is used for pixel-wise density map regression supervision (L_{Dp}). However, the pixel-wise *L2* loss assumes pixel-wise independence, which results in an over-smooth density map prediction [5] and the underlying bias from unbalanced low- and high-level density distributions of crowd images. To address this issue, we propose a Dilated Contrastive Density Loss (L_{DCD}), where we take into account more adjacent pixels for regional density difference. In detail, we applied a single layer convolution on the regressed density map M_D and the ground truth density map D_G. The single layer convolution has eight filters; each filter contains a 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} in Eq. 6 as below:

$$L_{DCD} = \sum_{i} ||K_i^{DCD} \otimes M_D - K_i^{DCD} \otimes D_G||_2^2, \quad (6)$$

where K_i^{DCD} is the i^{th} dilated contrastive convolution kernel, $i \in [1, 8]$. Details of the kernel are shown in Fig. 6, where a 3×3 convolution layer with the dilated rate of 2 is applied; this gives a larger receptive field as 5×5. The kernel value is empirically set as 0 ,-1, and 1 because we do not find any significant difference regrading different kernel values. On the other hand, the kernel value is designed to achieve a contrastive learning purpose to include regional relationships among pixels instead of single pixel-wise L2 or L1 loss. We performed extensive experiments to evaluate the effectiveness of the proposed L_{DCD} loss; quantitative results in the *Ablation Study* (Section V-D) demonstrates that the proposed L_{DCD} loss can improve the counting accuracy not only for our model but also for previous single *L2* loss-based methods.

IV. Experiments

A. Datasets

ShanghaiTech [16] consists of 1,198 images, containing a total amount of 330,165 people with head centre point annotations. This dataset has been 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 subsets as specified in [16]. ***UCF-QNRF*** [90] is a large crowd dataset, consisting of 1,535 images with around 1.25 million annotations in total. The number of people in these images varies largely with a wide range spanning from 49 to 12,865. As indicated by [90], for training, 1,201 images are used, the remaining 334 images form the test set. ***JHU-Crowd++*** [91] is a recent challenging large-scale dataset that contains 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*** [19] is currently 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 vehicle counting dataset: ***Trancos*** [92] with 403 images for training, 420 images for validation and 421 images for testing. These experiments further demonstrate 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 128×128 at a random location, then randomly flipped the image patches horizontally with a probability of 0.3. We trained our model with 400 epochs for all experiments, with a starting learning rate of $1e-4$ and a cosine decay schedule [93]. The batch size is set to 96. Five-fold cross-validation is used for fair comparison and hyper-parameter tuning is applied in all settings. We implemented the proposed method with *PyTorch 1.7, CUDA 10.2* using *Python 3.6*. All the training processes are performed on a server with four TESLA V100, and all the test experiments are conducted on a local workstation with *Intel(R) Xeon(R) W-2104 CPU* and *Geforce RTX 2080Ti GPU*. Our model takes average 19.5 hours to train on *JHU-Crowd++* [91] and *NWPU-Crowd* [19] datasets and average 8.5 hours to train on *ShanghaiTech* [16], *UCF-QNRF* [90] and *Trancos* [92]. Our implementation code is publicly available at: https://github.com/smallmax00/Counting_With_Adaptive_Auxiliary.

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 in Eq. 7 as below:

$$GAME(L) = \frac{1}{N} \sum_{n=1}^{N} (\sum_{l=1}^{4^L} |y_n^l - \hat{y}_n^l|), \quad (7)$$

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, $GAME$ is equivalent to MAE.

V. Results

A. Counting Results

In this section, we present our experimental results on the crowd 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-E), we show that our model could indicate some mislabeled or incorrectly 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 counting datasets. Fig. 7 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

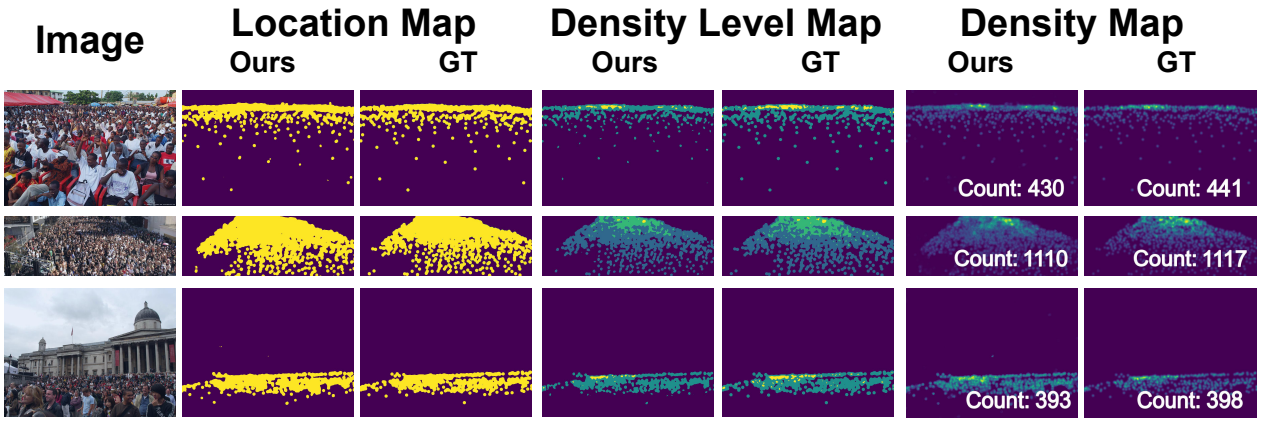

Fig. 7. Qualitative results of the density, crowd location and density level map in *SHA* test dataset. Our model can produce accurate density maps compared with the ground truth (*GT*), along with accurate auxiliary crowd segmentation and density level segmentation results.

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	*SHA*		*SHB*		*QNRF*		*JHU-Crowd++*		*NWPU-Crowd*	
	MAE	*RMSE*	*MAE*	*RMSE*	*MAE*	*RMSE*	*MAE*	*RMSE*	*MAE*	*RMSE*
CP-CNN [69]	73.6	106.4	20.1	30.1	-	-	-	-	-	-
DecideNet [5]	-	-	21.53	31.98	-	-	-	-	-	-
CFF [4]	65.2	109.4	7.2	12.2	93.8	146.5	83.6	400.7	80.8	364.1
AT-CSRNet [74]	-	-	8.11	13.53	-	-	-	-	-	-
SHRGBD [67]	70.3	111.0	8.8	15.3	113.3	177.6	107.9	446.7	103.0	478.1
HA-CCN [70]	62.9	94.9	8.1	12.7	118.1	180.4	-	-	-	-
RAZ-Net [66]	65.1	106.7	8.4	14.1	116	195	-	-	151.5	634.6
HYGNN [8]	60.2	94.5	7.5	12.7	100.8	185.3	-	-	-	-
LSC-CNN [80]	66.4	117.0	8.1	12.7	120.5	218.2	112.7	454.4	90.4	388.8
ASCC [18]	57.8	**90.1**	7.5	13.1	91.6	159.7	84.6	355.1	95.7	398.0
UMRNet [7]	62.6	103.3	7.2	11.5	86.3	153.1	-	-	-	-
DAMNet [6]	63.1	106.3	9.1	16.3	101.5	186.9	-	-	-	-
MATT [49]	59.5	97.3	6.9	**10.3**	-	-	-	-	-	-
SGANet [63]	57.6	101.1	**6.3**	10.6	87.6	152.5	-	-	-	-
Ours	**57.0**	98.6	7.1	12.3	**85.3**	**129.4**	**66.6**	**254.9**	**76.4**	**327.1**

methods in terms of *MAE* on all five datasets. In particular, our model outperforms the patch-based density level classification based method *HA-CCN* [70] by 14.7% via average *MAE*. Notably, the *JHU-Crowd++* dataset [91] and *NWPU-Crowd* dataset [19] are recent publicly available datasets, which are more challenging due to large variations in scale, occlusion, and complex weather scenes. Specifically, *NWPU-Crowd* is the current largest crowd counting benchmark [3]. To the best of our knowledge, we achieved the greatest 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 was able to enhance CACC (100.1 *MAE*) [17] by 18.3% and the dilated kernel-based method CSR-Net (85.9 *MAE*) [30] by 4.8% via *MAE*.

Vehicle Counting Results. We conducted experiments on vehicle (*Trancos* [92]) counting datasets to show our model's

[3]https://www.crowdbenchmark.com/nwpucrowd.html

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

Methods	Trancos	
	MAE	*RMSE*
PPPD [94]	9.7	-
CSRNet [30]	3.5	5.1
BL-Crowd [58]	2.9	6.7
MD-Crowd [60]	3.1	6.6
Auto-Scale [40]	2.9	6.1
SUANet-Fully [3]	4.9	6.9
SASNet [46]	2.9	4.7
Gau-SANet [76]	2.5	**2.8**
STNet [64]	3.8	5.0
ASCC [18]	3.8	4.9
DM-Count [60]	3.9	5.2
P2PNet [56]	3.8	4.9
WSNet [41]	4.3	5.8
Ours	**2.3**	4.8

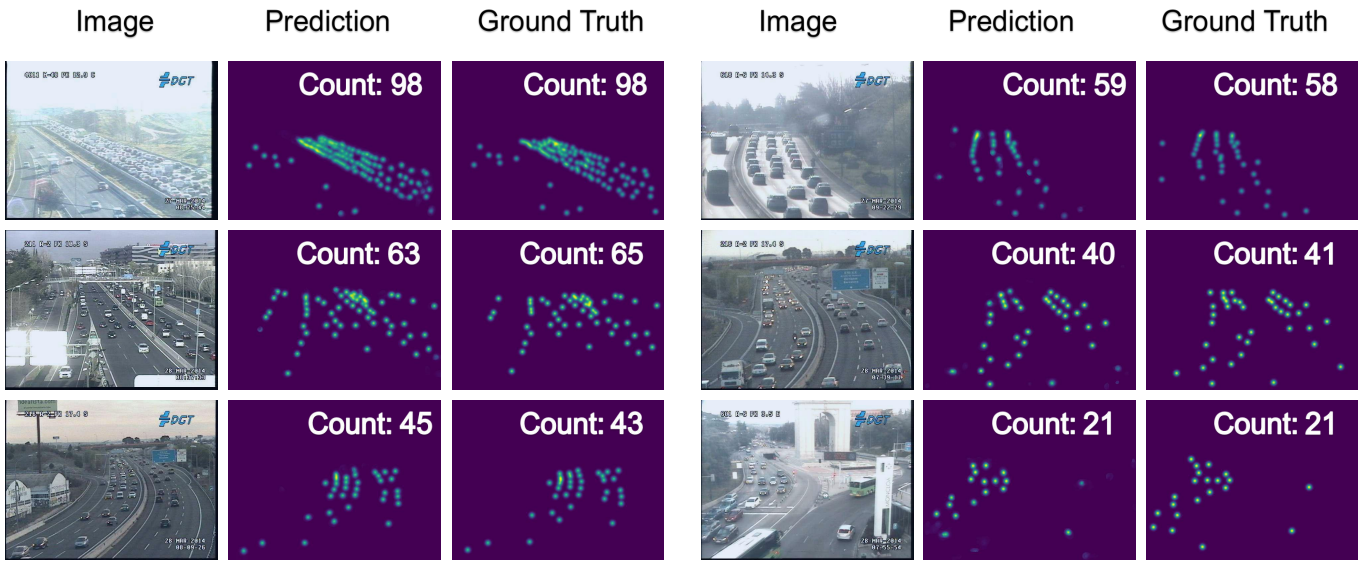

Fig. 8. Qualitative results on the Trancos dataset. The density map ground truth and our predictions are shown, with counting number presented in the figure. Our model adapts well with scale variations, where the scale of the vehicles varies from the distance between the camera and vehicle locations. Specifically, the vehicles that are far from the camera only contain a few pixels in the image, while the near-camera vehicles have more pixels. The scale of such pixel occupation changes can be well handled by our methods and the predicted density maps can clearly show the location correspondence.

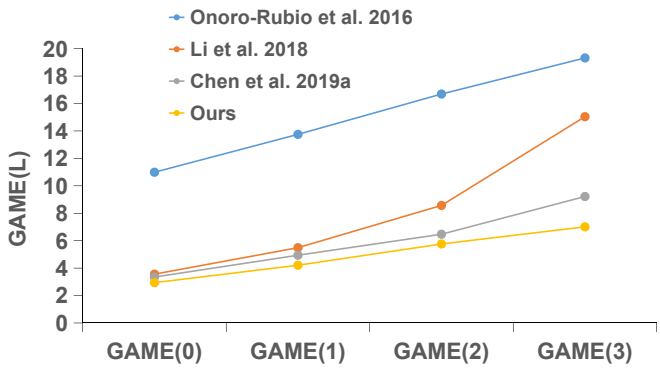

Fig. 9. Comparison of $GAME$ performance on the *Trancos* dataset among the proposed approach and the state-of-the-arts, such as *Onoro-Rubio et al.* [2], *Li et al.* [30], *Chen et al.* [29]. Note that, a small range of increase among different $GAME$ values indicates that our method counts and localizes overlapping vehicles more accurately.

broad applicability and robustness. Fig. 8 shows the qualitative results, and TABLE. II shows the quantitative results compared with the previous state-of-the-art methods. Due to the different scenes in the vehicle 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 designed our model especially for crowd counting tasks; still, our model achieves superior performance when compared with previous methods. Specifically, our model outperformed the distribution matching supervised methods *BL-Crowd* [58], *MD-Crowd* [60], *P2PNet* [56] and *DM-Count* [60] by 20.7, 25.8, 39.5 and 41.0 % of MAE; outperformed the auxiliary task assisted methods *Auto-Scale* [40], *SASNet* [46], *STNet* [64] and *ASCC* [18] by 20.7, 20.7, 39.5, and 39.5 % of MAE. Notably, *WSNet* [41] is specially designed for traffic density estimation and vehicle counting, where an attention-based Transformer [95] is used to extract the local-global consistent features. This is because the traffic scenario can be easily affected by weather and scale changes, which results in weakened semantic and spatial content of the captured images. Our proposed graph-based multi-granularity information fusion paradigm had a similar intuition, to enhance the relevant semantic and spatial information. Our model outperformed *WSNet* [41] by 46.5 % by MAE in Trancos test dataset. Furthermore, we present local comparison performance through the $GAME$ metric to indicate the model's ability to recognize the objects' locations. Fig. 9 shows the comparison results in terms of the $GAME$ on the Trancos dataset. As illustrated, our method localizes and counts overlapping vehicles more accurately.

Results on Weather Changes Among the seven datasets used in this work, *JHU-Crowd++* [91] provided the weather condition-based labels. For example, the test dataset (a total of 1600 images) contained 168 images weather labels; for example, 49 images are labeled as 'rain'; 78 images are labeled as 'snow'; 64 images are labeled as 'fog'. In this section, we provide the quantitative and qualitative counting results on different weather conditions. Following *JHU-Crowd++* [91] benchmark's setting, we report the counting performance on the test images with weather labels. Specifically in TABLE. III, our method achieved 110.2 MAE and 598.2 RMSE, which outperformed previous state-of-the-art methods *LSC-CNN* [80], and *MBTTBF* [42] by 38.1 and 20.5 % MAE. Benefiting from the proposed auxiliary task and the graph-based multi-granularity feature fusion mechanism, our model can extract the spatial and semantic features from the input image, especially when weather degradation causes a weakened image quality. Fig. 10 shows the qualitative results of our model under different weather conditions. Our model can

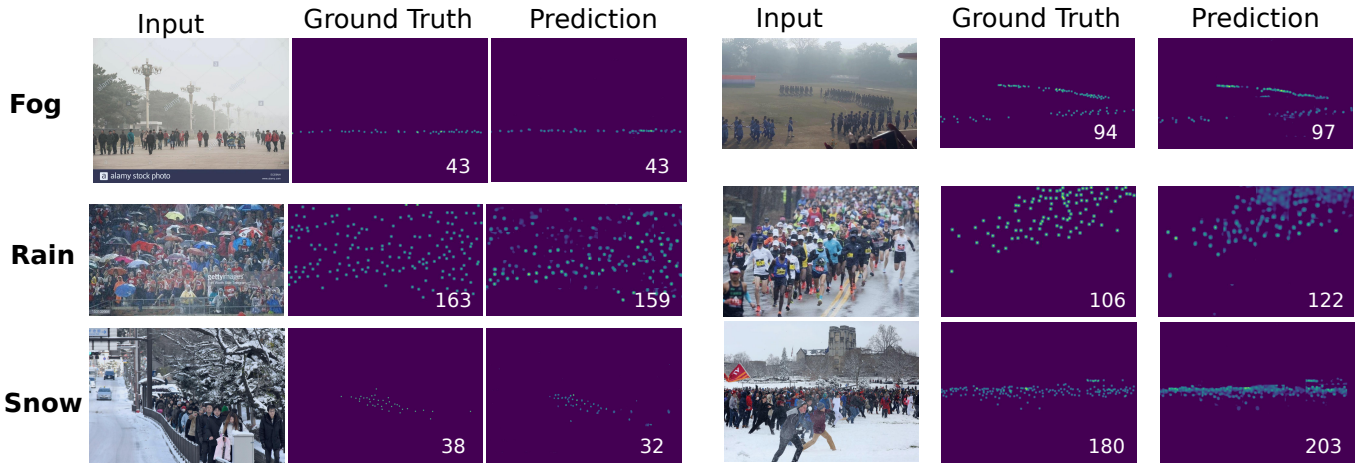

Fig. 10. Qualitative results on different weather conditions of the *JHU-Crowd++* dataset. The density map ground truth and our predictions are shown, with the counting number presented in the figure. In total, three conditions, fog, rain, and snow, are demonstrated in the respective rows of the figure. Our model can handle severe weather degradation well and indicates precise crowd locations.

handle the severe weather degradation well, which is critical in the intelligent transportation system because weather can easily affect traffic scenarios.

TABLE III
RESULTS ON *JHU-Crowd++* [91] COUNTING DATASET UNDER WEATHER SETTING. WE FOLLOW THE *JHU-Crowd++* [91] BENCHMARK'S SETTING AND REPORT THE COUNTING PERFORMANCE. OUR MODEL ACHIEVES SUPERIOR PERFORMANCE TO THE PREVIOUS STATE-OF-THE-ART METHODS.

Methods	JHU-Weather	
	MAE	*RMSE*
CSRNet [30]	141.4	640.1
SA-Net [77]	154.2	685.7
CACC [17]	155.4	617.0
DSSI-Net [25]	229.1	760.3
MBTTBF [42]	138.7	631.6
LSC-CNN [80]	178.0	744.3
JHU-Crowd++ [91]	138.6	654.0
SFCN [78]	122.8	606.3
BL-Crowd [58]	140.1	675.7
Ours	**110.2**	**598.2**

B. Auxiliary Task Results

In this section, we report the performance of the 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. Fig. 7 shows examples of those tasks' predictions from our model.

TABLE IV
COMPUTATIONAL EFFICIENCY. THE NUMBER OF PARAMETERS IN MILLIONS (*M*), FLOATING-POINT OPERATIONS (*FLOPs*) AND INFERENCE TIME IN MILLISECOND (*ms*) OF DIFFERENT COUNTING METHODS ON A FIXED SIZE OF 128×128 INPUT IMAGE.

Methods	*Params* (*M*)	*FLOPs* (*G*)	*Inference Time* (*ms*)
DM-Count [60]	21.5	6.7	1.9
SUANet-Fully [3]	15.9	6.5	5.3
LSC-CNN [80]	35.1	25.4	4.6
BL-Crowd [58]	21.5	6.7	1.9
ASCC [18]	30.4	10.2	3.2
SASNet [46]	38.9	14.6	7.8
Ours	18.8	8.5	8.8

C. Computational Efficiency

Table.IV presents the number of parameters in millions (*M*), floating-point operations (*FLOPs*) and inference time in millisecond (*ms*) of the compared models. Our model adopts *VGG-16* [81] as the backbone, which leads to a relatively smaller model size of 18.8 *M* parameters, compared to other models, such as *LSC-CNN [80]* (35.1 *M*), *ASCC* [18] (30.4 *M*), and *SASNet* [46] (38.9 *M*). On the other hand, our model is computationally effective, only requiring 8.5 *FLOPs*. This is comparable to other light-weight models such as *DM-Count* [60], *SUANet-Fully* [3], and *BL-Crowd* [58]. Due to the auxiliary task-based nature, our model required a relatively longer inference time, such as 8.8 *ms* per image. However, our method can still be used for a real-time counting application (*inference speed* $>$ *24 frame per second*).

D. 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. V with several different backbone networks. In general, *VGG*-based backbone networks achieved comparable counting performance, compared with *ResNet*-based backbone networks in relatively large-scale datasets, such as *QNRF*, *JHU-Crowd++* and *NWPU-Crowd*. While, *ResNet*-based backbones work better on small-scale counting datasets, such as *SHA*

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

Methods	*SHA*		*SHB*		*QNRF*		*JHU-Crowd++*		*NWPU-Crowd*	
	MAE	*RMSE*	*MAE*	*RMSE*	*MAE*	*RMSE*	*MAE*	*RMSE*	*MAE*	*RMSE*
VGG-16 [81]	**57.0**	98.6	7.1	12.3	**85.3**	129.4	**66.6**	254.9	**76.4**	**327.4**
VGG-19 [81]	59.7	99.8	8.4	13.2	87.8	144.0	73.7	320.1	79.9	360.0
ResNet-50 [96]	57.8	**96.6**	**7.0**	**11.7**	85.5	**128.7**	77.9	318.1	79.3	344.4
ResNet-101 [96]	61.1	100.8	9.1	14.5	93.3	147.9	69.7	**253.3**	81.4	361.5

and *SHB*. We report 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 evaluate the effectiveness of the auxiliary tasks, adaptively shared backbone network, and *GCN*-enabled reasoning module. Please note that, in order to eliminate the performance improvement from a bigger model, we add feed-forward *CNN* blocks containing (3×3 convolution with Batch Normalization) into other ablation study models in TABLE. VI 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 (*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 to be 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. VI. 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. 11, 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. VII, 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. VII, 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*.

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

Methds	*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	**57.0**	**98.6**	**66.6**	**254.9**

TABLE VII
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.

Methds	*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	**57.0**	**98.6**	**66.6**	**254.9**

Ablation on Graph Reasoning Module. In this section, we evaluate the effectiveness of the proposed graph reasoning module. We specially designed our graph reasoning module

TABLE VIII
ABLATION STUDY RESULTS ON GRAPH REASONING MODULES. ONLY OUR PROPOSED GRAPH REASONING MODULE CAN EFFICIENTLY UTILIZE THE AUXILIARY INFORMATION FROM OTHER TASKS TO COMPLEMENT THE DENSITY MAP REGRESSION TASK.

Methds	*SHA*		*JHU-Crowd++*	
	MAE	*RMSE*	*MAE*	*RMSE*
classic GCN	67.1	109.0	79.2	308.7
SGR [97]	60.3	101.0	73.1	301.0
DualGCN [98]	63.8	105.7	80.8	307.3
GloRe [99]	61.0	105.4	71.3	317.7
Ours	**57.0**	**98.6**	**66.6**	**254.9**

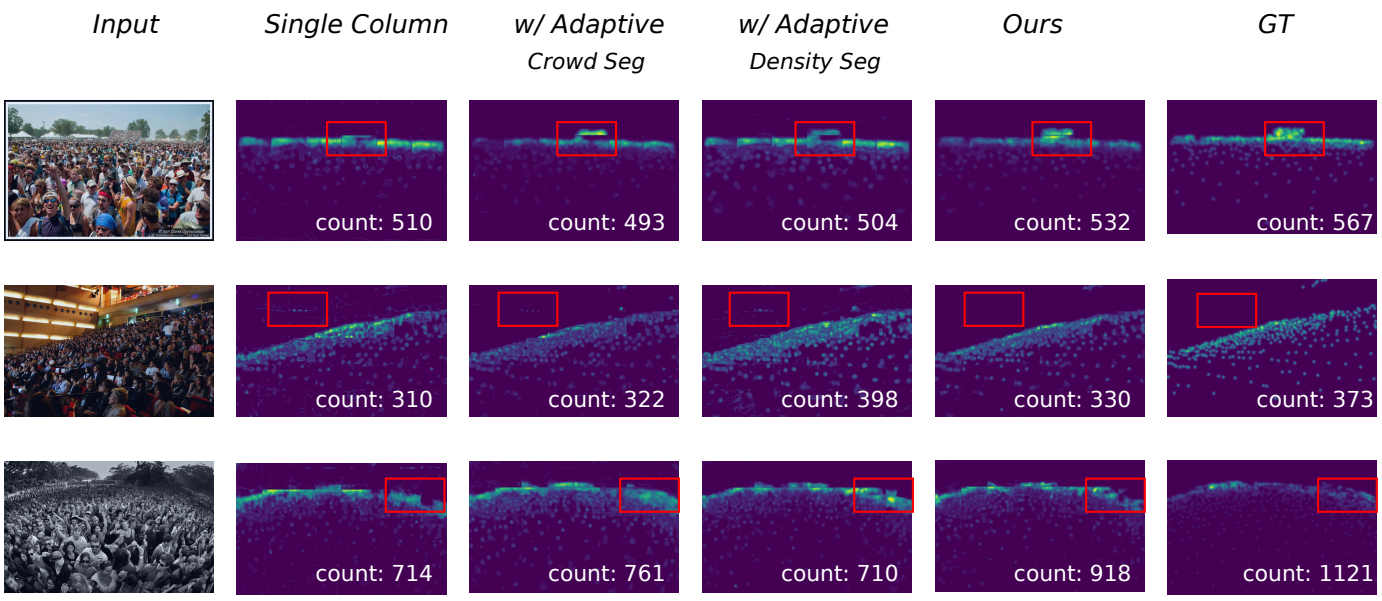

Fig. 11. The qualitative results of the predicted density maps of ablation studies about auxiliary tasks. The red bounding boxes are used for better visualization and comparison. *Ours* and *w/ Adaptive Crowd Seg* can know the crowd's spatial regions (first and third rows), and filter out the background 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).

TABLE IX
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)	**57.0**	**98.6**	**66.6**	**254.9**

TABLE X
ABLATION STUDY RESULTS (*MAE*) ON OUR COMBINED LOSS (CONTRASTIVE AND *L2* LOSS), COMPARED WITH SINGLE *L2* LOSS (*base*). MOREOVER, WE APPLIED THE COMBINED LOSS FUNCTION TO OPTIMIZE PREVIOUS SINGLE *L2* 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 [16]	110.2	**108.1**	188.9	**168.3**
CSRNet [30]	68.2	**65.9**	85.9	**84.1**
CACC [17]	62.3	**60.8**	100.1	**97.9**
Ours	59.5	**57.0**	70.8	**66.6**

to incorporate the auxiliary tasks and for fusing information into the adjacency matrix to form the information-fused graph. So for the ablation study, we had to only apply other *GCN* on the density map. Firstly, we employed the classic graph convolution [89] to reason the correlations between regions in density feature maps (f_{DM}). Additionally, we adopted potent graph convolution operations to show the superiority of our proposed *Graph Reasoning Module*. In detail, we applied the *SGR* [97], *DualGCN* [98], and *GloRe* module [99] respectively, where the *SGR* module exploited a knowledge graph mechanism; *DualGCN* explored the coordinate space and feature space graph convolution; and *GloRe* utilized a projection and re-projection mechanism to reason the semantics between different regions. These methods achieved state-of-the-art performance on different computer vision tasks, however, they can only process single task rather than using auxiliary information. Tab. VIII shows that our model achieves more accurate and reliable results than [89] and outperforms the *SGR*, *DualGCN*, and *GloRe* by 7.2 %, 20.0 % and 6.6 % in terms of mean *MAE* on the two test datasets.

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 3×3 convolution layer into 1, 2, 3, 4, which resulted in the receptive field of the L_{DCD} being like 3, 5, 7, 9. TABLE. IX shows the comparison results; when the dilated rate is 2, our model achieves the best performance on *SHA* and *JHU-Crowd++* datasets.

Furthermore in TABLE. X, 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*). Furthermore, we applied the proposed combined loss function (*w/ contrastive*) into previous single *L2*-based methods [16], [30], [17]. We re-implemented their network with their open-source code and used the same experimental setting as our method. TABLE. X 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 *L2* loss function via average *MAE* on two datasets. Our proposed L_{DCD} also helps to reduce the original *MCNN* [16]

by 6.4%, the *CSRNet* [30] by 2.7%, and the *CACC* [17] by 2.3% over average *MAE* on two datasets. Please note that we did not compare with other loss functions that were proposed in the recent crowd counting models [56], [58], [57], [60], [61], [59]. Those methods are not pure density map regression-based methods, thus it is unfair to compare.

E. Discussion: Comparison with Ground Truth

Underlying labeling errors (noisy ground truth) exist in most datasets due to human annotator error. However, a robust model can omit noisy ground truths during training and produce a more accurate prediction. This section showed that our model could indicate some mislabeled or incorrectly labeled 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. 2 shows a wrongly labeled point annotation (top left) case of the crowd counting test dataset, and the other cases are mislabeled point annotation of vehicle counting test dataset. We highlighted the incorrectly labeled or mislabeled area with red bounding boxes for better visualization and comparison.

F. Limitation and Future Work

In this work, we presented an object counting framework assisted by auxiliary multi-granularity information, achieving cutting-edge counting performance in seven large-scale counting datasets. This significantly contributes to transportation systems, including many applications such as security alerts, public space design, *etc.*. However, one limitation of our method is that the complexity of inference is increased due to the enlarged number of optimized tasks. This is a typical issue of auxiliary-task based counting methods [43], [44], [45], [46], [47], [48], [49], [50], [6], [51], [52], which has been discussed before. However, our method only required 8.8 milliseconds per image, which is comparable to other single-task-based methods (please refer to TABLE. IV). In other words, our method can also be used for a real-time counting application (*inference speed* $>$ *24 frame per second*). The trade-off between accuracy and complexity can be determined when applied to a real-world task.

A future extension of our work could be multiple objects tracking (*MOT*), such as vehicles or crowd tracking. Most of the *MOT* approaches [100], [101], [102] follow the classic paradigm of tracking-by-detection, where object trajectories are obtained by associating per-frame outputs of object detectors. Recently, a new prediction scheme [103], [68] is gaining attention that uses a tracking-by-counting mechanism. Specifically, using the crowd density maps, the detection, counting, and tracking of multiple targets as a network flow program is achieved. In the future, our model could be integrated into such learning pipelines to tackle *MOT* with dense crowds or vehicles.

VI. Conclusion

We proposed an auxiliary-task-based object counting methodology via a graph-based multi-granularity information fusion paradigm. The proposed task-adaptive backbone enabled the task-shared and task-specific features to be learned simultaneously. We have demonstrated its potential in maintaining state-of-the-art performance upon seven challenging benchmarks. Our approach is anticipated to be widely applicable in the real world.

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. Meng, H. Zhang, Y. Zhao, X. Yang, X. Qian, X. Huang, and Y. Zheng, "Spatial uncertainty-aware semi-spervised crowd counting," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 15 549–15 559.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] Y. Li and A. Gupta, "Beyond grids: learning graph representations for visual recognition," in *Advances in Neural Information Processing Systems*, 2018, pp. 9225–9235.

[10] Y. Meng, W. Meng, D. Gao, Y. Zhao, X. Yang, X. Huang, and Y. Zheng, "Regression of instance boundary by aggregated cnn and gcn," in *Proceedings of the European Conference on Computer Vision*, 2020.

[11] 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.

[12] Y. Meng, H. Zhang, Y. Zhao, X. Yang, Y. Qiao, J. C. M. Ian, X. Huang, and Y. Zheng, "Graph-based region and boundary aggregation for biomedical image segmentation," in *IEEE Transactions on Medical Imaging*, 2021.

[13] Y. Meng, H. Zhang, D. Gao, Y. Zhao, X. Yang, X. Qian, X. Huang, and Y. Zheng, "BI-GCN: boundary-aware input-dependent graph convolution network for biomedical image segmentation," in *32nd British Machine Vision Conference 2021, BMVC 2021, Online, November 22-25, 2021*. BMVA Press, 2021, p. 223. [Online]. Available: https://www.bmvc2021-virtualconference.com/assets/papers/0097.pdf

[14] Y. Meng, H. Zhang, Y. Zhao, D. Gao, B. Hamill, G. Patri, T. Peto, S. Madhusudhan, and Y. Zheng, "Dual consistency enabled weakly and semi-supervised optic disc and cup segmentation with dual adaptive graph convolutional networks," *IEEE Transactions on Medical Imaging*, p. in press, 2022.

[15] Y. Meng, X. Chen, H. Zhang, Y. Zhao, D. Gao, B. Hamill, G. Patri, T. Peto, S. Madhusudhan, and Y. Zheng, "Shape-aware weakly/semi-supervised optic disc and cup segmentation with regional/marginal consistency," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2022.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] H. Wu, Q. Li, C. Wen, X. Li, X. Fan, and C. Wang, "Tracklet proposal network for multi-object tracking on point clouds," in *Proceedings of the 30th International Joint Conference on Artificial Intelligence*, 2021, pp. 1165–1171.

[21] 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.

[22] 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.

[23] J. Yang, Y. Zhou, and S.-Y. Kung, "Multi-scale generative adversarial networks for crowd counting," in *2018 24th International Conference on Pattern Recognition*. IEEE, 2018, pp. 3244–3249.

[24] 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*, 2021.

[25] 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.

[26] L. Liu, H. Wang, G. Li, W. Ouyang, and L. Lin, "Crowd counting using deep recurrent spatial-aware network," in *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, 2018, pp. 849–855.

[27] X. Ding, F. He, Z. Lin, Y. Wang, H. Guo, and Y. Huang, "Crowd density estimation using fusion of multi-layer features," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 8, pp. 4776–4787, 2020.

[28] 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.

[29] 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*. IEEE, 2019, pp. 1941–1950.

[30] 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.

[31] 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.

[32] Z. Yan, R. Zhang, H. Zhang, Q. Zhang, and W. Zuo, "Crowd counting via perspective-guided fractional-dilation convolution," *IEEE Transactions on Multimedia*, 2021.

[33] 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.

[34] 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.

[35] 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.

[36] 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.

[37] 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*, 2021.

[38] H. Duan, S. Wang, and Y. Guan, "Sofa-net: second-order and first-order attention network for crowd counting," in *31st British Machine Vision Conference*, 2020.

[39] Q. Wang, W. Lin, J. Gao, and X. Li, "Density-aware curriculum learning for crowd counting," *IEEE Transactions on Cybernetics*, 2020.

[40] C. Xu, D. Liang, Y. Xu, S. Bai, W. Zhan, X. Bai, and M. Tomizuka, "Autoscale: learning to scale for crowd counting," *International Journal of Computer Vision*, vol. 130, no. 2, pp. 405–434, 2022.

[41] Y.-X. Hu, R.-S. Jia, Y.-B. Liu, Y.-C. Li, and H.-M. Sun, "Wsnet: a local–global consistent traffic density estimation method based on weakly supervised learning," *Knowledge-Based Systems*, p. 109727, 2022.

[42] V. A. Sindagi and V. M. Patel, "Multi-level bottom-top and top-bottom feature fusion for crowd counting," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 1002–1012.

[43] 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.

[44] 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.

[45] S. Abousamra, M. Hoai, D. Samaras, and C. Chen, "Localization in the crowd with topological constraints," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, 2021, pp. 872–881.

[46] 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.

[47] 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.

[48] 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.

[49] 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.

[50] J. Wan, N. S. Kumar, and A. B. Chan, "Fine-grained crowd counting," *IEEE Transactions on Image Processing*, vol. 30, pp. 2114–2126, 2021.

[51] W. Liu, M. Salzmann, and P. Fua, "Counting people by estimating people flows," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[52] ——, "Estimating people flows to better count them in crowded scenes," in *European Conference on Computer Vision*. Springer, 2020, pp. 723–740.

[53] 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.

[54] ——, "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.

[55] 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.

[56] 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*, 2021.

[57] 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.

[58] 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.

[59] J. Wan and A. Chan, "Modeling noisy annotations for crowd counting," *Advances in Neural Information Processing Systems*, vol. 33, 2020.

[60] B. Wang, H. Liu, D. Samaras, and M. H. Nguyen, "Distribution matching for crowd counting," *Advances in Neural Information Processing Systems*, vol. 33, 2020.

[61] 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*, 2021.

[62] 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.

[63] Q. Wang and T. P. Breckon, "Crowd counting via segmentation guided attention networks and curriculum loss," *IEEE Transactions on Intelligent Transportation Systems*, 2022.

[64] M. Wang, H. Cai, X. Han, J. Zhou, and M. Gong, "Stnet: scale tree network with multi-level auxiliator for crowd counting," *IEEE Transactions on Multimedia*, 2022.

[65] 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.

[66] 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*. IEEE, 2019, pp. 1217–1226.

[67] 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.

[68] 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.

[69] 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.

[70] V. Sindagi and V. Patel, "Ha-ccn: hierarchical attention-based crowd counting network," *IEEE Transactions on Image Processing*, vol. 29, pp. 323–335, 2019.

[71] 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.

[72] 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.

[73] 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.

[74] 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. 12 736–12 745.

[75] 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.

[76] Z.-Q. Cheng, Q. Dai, H. Li, J. Song, X. Wu, and A. G. Hauptmann, "Rethinking spatial invariance of convolutional networks for object counting," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 19 638–19 648.

[77] X. Cao, Z. Wang, Y. Zhao, and F. Su, "Scale aggregation network for accurate and efficient crowd counting," in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 734–750.

[78] Q. Wang, J. Gao, W. Lin, and Y. Yuan, "Learning from synthetic data for crowd counting in the wild," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 8198–8207.

[79] V. Lempitsky and A. Zisserman, "Learning to count objects in images," in *Advances in Neural Information Processing Systems*, 2010, pp. 1324–1332.

[80] 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.

[81] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *International Conference on Learning Representations*, 2015.

[82] 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.

[83] M. Amrani, K. Yang, D. Zhao, X. Fan, and F. Jiang, "An efficient feature selection for sar target classification," in *Pacific Rim Conference on Multimedia*. Springer, 2017, pp. 68–78.

[84] M. Amrani, F. Jiang, Y. Xu, S. Liu, and S. Zhang, "Sar-oriented visual saliency model and directed acyclic graph support vector metric based target classification," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 11, pp. 3794–3810, 2018.

[85] M. Amrani, M. Hammad, F. Jiang, K. Wang, and A. Amrani, "Very deep feature extraction and fusion for arrhythmias detection," *Neural Computing and Applications*, vol. 30, pp. 2047–2057, 2018.

[86] M. Amrani and F. Jiang, "Deep feature extraction and combination for synthetic aperture radar target classification," *Journal of Applied Remote Sensing*, vol. 11, p. 042616, 2017.

[87] M. Amrani, A. Bey, and A. Amamra, "New sar target recognition based on yolo and very deep multi-canonical correlation analysis," *International Journal of Remote Sensing*, pp. 1–20, 2021.

[88] 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.

[89] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," *International Conference on Learning Representations*, 2017.

[90] 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*, 2018, pp. 532–546.

[91] 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.

[92] 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.

[93] I. Loshchilov and F. Hutter, "SGDR: stochastic gradient descent with warm restarts," *International Conference on Learning Representations*, 2017.

[94] 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.

[95] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[96] 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.

[97] X. Liang, Z. Hu, H. Zhang, L. Lin, and E. P. Xing, "Symbolic graph reasoning meets convolutions," in *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, 2018, pp. 1858–1868.

[98] L. Zhang, X. Li, A. Arnab, K. Yang, Y. Tong, and P. H. Torr, "Dual graph convolutional network for semantic segmentation," in *BMVC2019*, 2019.

[99] Y. Chen, M. Rohrbach, Z. Yan, Y. Shuicheng, J. Feng, and Y. Kalantidis, "Graph-based global reasoning networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 433–442.

[100] A. Kim, A. Ošep, and L. Leal-Taixé, "Eagermot: 3D multi-object tracking via sensor fusion," in *2021 IEEE International Conference on Robotics and Automation*. IEEE, 2021, pp. 11 315–11 321.

[101] M. P. Muresan, S. Nedevschi, and R. Danescu, "Robust data association using fusion of data-driven and engineered features for real-time pedestrian tracking in thermal images," *Sensors*, vol. 21, no. 23, p. 8005, 2021.

[102] H. Wu, W. Han, C. Wen, X. Li, and C. Wang, "3d multi-object tracking in point clouds based on prediction confidence-guided data association," *IEEE Transactions on Intelligent Transportation Systems*, 2021.

[103] L. Wen, D. Du, P. Zhu, Q. Hu, Q. Wang, L. Bo, and S. Lyu, "Detection, tracking, and counting meets drones in crowds: a benchmark," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 7812–7821.