Frequency Feature Pyramid Network With Global-Local Consistency Loss for Crowd-and-Vehicle Counting in Congested Scenes

Xiaoyuan Yu, Yanyan Liang, *Member, IEEE*, Xuxin Lin, Jun Wan, *Senior Member, IEEE*, Tian Wang, and Hong-Ning Dai, *Senior Member, IEEE*

***Abstract*— Context prediction plays a crucial role in implementing autonomous driving applications. As one of important context-prediction tasks, crowd-and-vehicle counting is critical for achieving real-time traffic and crowd analysis, consequently facilitating decision-making processes for autonomous vehicles. However, the completion of crowd-and-vehicle counting also faces challenges, such as large-scale variations, imbalanced data distribution, and insufficient local patterns. To tackle these challenges, we put forth a novel frequency feature pyramid network (FFPNet) in this paper. Our proposed FFPNet extracts the multi-scale information by frequency feature pyramid module, which can tackle the issue of large-scale variations. Meanwhile, the frequency feature pyramid module uses different frequency branches to obtain different scale information. We also adopt the attention mechanism to strength the extraction of different scale information. Moreover, we devise a novel loss function, namely global-local consistency loss, to address the existing problems of imbalanced data distribution and insufficient local patterns. Furthermore, we conduct extensive experiments on six datasets to evaluate our proposed FFPNet. It is worth mentioning that we also construct a novel crowd-and-vehicle dataset (CROVEH), which is the only dataset that contains both crowd-and-vehicle annotations. The experimental results show that FFPNet achieves the best performance on different backbones, e.g., 52.69 mean absolute error (MAE) on P2PNet with FFP module. The codes are available at: https://github.com/MUST-AI-Lab/FFPNet.**

***Index Terms*— Context prediction, frequency feature pyramid, discrete cosine transformation, global-local consistency loss.**

Manuscript received 26 February 2021; revised 28 September 2021, 31 January 2022, and 15 May 2022; accepted 21 May 2022. Date of publication 10 June 2022; date of current version 8 July 2022. This work was supported in part by the National Key Research and Development Plan under Grant 2021YFE0205700; in part by the External Cooperation Key Project of Chinese Academy Sciences under Grant 173211KYSB20200002; in part by the Chinese National Natural Science Foundation Project under Grant 61876179 and Grant 61961160704; in part by the Science and Technology Development Fund of Macau under Grant 0008/2019/A1, Grant 0010/2019/AFJ, Grant 0025/2019/AKP, Grant 0004/2020/A1, and Grant 0070/2021/AMJ; in part by the Guangdong Provincial Key Research and Development Programme under Grant 2019B010148001; in part by the National Natural Science Foundation of China (NSFC) under Grant 62172046; and in part by the Special Project of Guangdong Provincial Department of Education in Key Fields of Colleges and Universities under Grant 2021ZDZX1063. The Associate Editor for this article was S. Wan. *(Corresponding author: Yanyan Liang.)*

Xiaoyuan Yu, Yanyan Liang, and Xuxin Lin are with the Faculty of Innovation Engineering, School of Computer Science and Engineering, Macau University of Science and Technology, Macau, China (e-mail: acreateryxy@gmail.com; yyliang@must.edu.mo; linxuxin6@gmail.com).

Jun Wan is with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China (e-mail: jun.wan@nlpr.ia.ac.cn).

Tian Wang is with the Guangdong Key Laboratory of AI and Multi-Modal Data Processing, BNU-HKBU United International College, and the BNU-UIC Institute of Artificial Intelligence and Future Networks, Beijing Normal University (BNU Zhuhai), Zhuhai, Guangdong 519088, China (e-mail: cs_tianwang@163.com).

Hong-Ning Dai is with the Department of Computing and Decision Sciences, Lingnan University, Hong Kong, China (e-mail: hndai@ieee.org).

Digital Object Identifier 10.1109/TITS.2022.3178848

I. Introduction

As a crucial component in the future smart transportation system, self-driving vehicles or autonomous vehicles have received tremendous attention recently [1]. However, both the safety and reliability of autonomous driving are major concerns before the wide proliferation of this technology [2]. The lack of context awareness and context prediction in transportation systems is one of the obstacles for autonomous driving. Contexts in transportation systems include various events, traffic flows, pedestrian crowd, and surroundings [3]–[8], among which the number of vehicles and persons is an important indicator to determine a transportation context. Therefore, recent studies [9], [10] concentrate on crowd (or vehicle) counting. Crowd (or vehicle) counting aims to predict the number of persons or the number of vehicles by estimating the crowd or vehicle density distribution in a scene. This task can be achieved by a density map estimation of the crowd or vehicle distribution.

A. Motivation

Despite recent advances in crowd-and-vehicle counting, we still face three major challenges.

Challenge 1 (Large-Scale Variations of Persons and Vehicles): Fig. 1(a) shows a real scenario at a crosswalk, in which there are pedestrians and vehicles. Some of them are close to the camera while some of them are far from the camera. Therefore, there are large-scale variations caused by the different sizes of heads and vehicles. To address this issue, the extraction of multi-scale information is very important. For extracting multi-scale information, two common methods include using multiple convolution kernels of different sizes [12]–[17] and leveraging multiple down-sampling operations with different scaling factors [18], [19]. Multi-column Convolutional Neural Network (MCNN) [12] uses three parallel

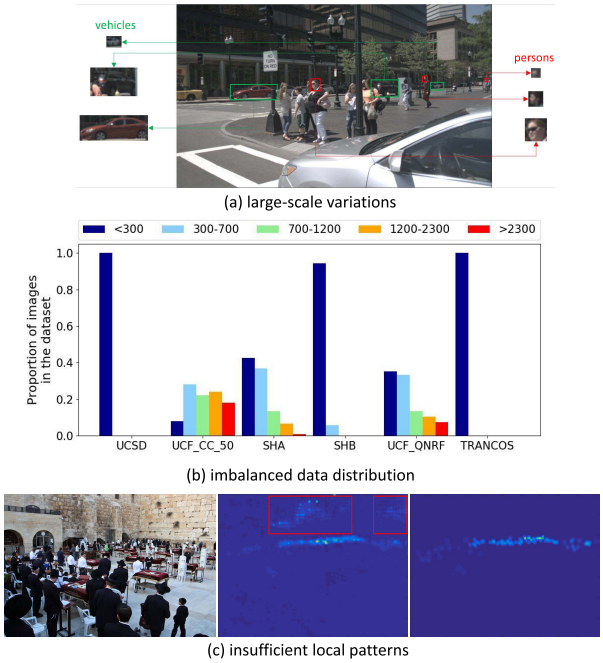

Fig. 1. Three challenges in crowd-and-vehicle counting: (a) large-scale variations; (b) imbalanced data distribution; (c) insufficient local patterns. In Fig. 1(c), the input image is on the left, the predicted density map generated by CSRNet [11] is in the middle, and the ground truth is on the right.

sub-networks (subnets), which have the convolutional kernel with the varied size to capture heads or vehicles with different sizes while it suffers from insufficient capacity. Although Context-Aware Network (CAN) [19] extracts the multi-scale information using the pooling layers with varied-size kernels, the multi-scale information obtained from arbitrary size input is represented as the fixed-size feature maps, consequently losing much detailed information. Despite the advent of Path Aggregation Network (PANet) [20], Neural Architecture Search Feature Pyramid Network (NAS-FPN) [21] and bi-directional feature pyramid network (BiFPN) [22] based on the Feature Pyramid Network (FPN) [23], they all suffer from the interference of shallow features containing too much redundant information on the regression of highly abstracted density maps.

Challenge 2 (Imbalanced Data Distribution): There are several major datasets for crowd-and-vehicle counting: UCSD dataset [24], UCF_CC_50 dataset [14], ShanghaiTech dataset [12], UCF-QNRF dataset [25], and TRANCOS dataset [26]. However, those five commonly-used datasets have *imbalanced data distribution*. Fig. 1(b) plots the distribution of these five datasets. Kindly note that ShanghaiTech dataset is divided into two datasets: SHA (for dense dataset) and SHB (for sparse dataset). We observe that SHA dataset suffers from strong imbalanced data distribution, e.g., most of data samples of them are sparsely distributed (i.e., $<$ 300 samples). The imbalanced-data-distribution problem also exists in the other three datasets, UCF_CC_50 dataset, SHA dataset, and UCF-QNRF dataset. The imbalanced data distribution is a general problem in both classification and regress tasks though the difference between classification and regression tasks is caused by different evaluation metrics of them. The evaluated metric, e.g., mean Average Precision (mAP) of classification tasks suggests the model be partial to the categories of fewer samples via resampling or reweighting methods such as the Focal Loss. However, the evaluated metrics, such as Mean Absolute Error (MAE) and mean squared error (MSE) of regression tasks induce the model to discard outliers, i.e., fewer samples of extremely dense and sparse. Although SmoothL_1 [27] can alleviates the negative influence of outliers, it does not take into account inliers and is not everywhere differentiable. Similarly, BalancedL_1 [28] can be applied in this regression task while it is designed to balance classification loss and localization loss in detection tasks.

Challenge 3 (Insufficient Local Patterns): Many deep models also suffer from insufficient local patterns. Fig. 1(c) presents a complex scenario, in which lots of people stand in front of a wall. As shown in Fig. 1(c), we observe that a part of the wall is mistakenly regarded as a dense crowd (i.e., the red box) by Congested Scene Recognition Network (CSRNet) [11]. This problem often occurs when the scene contains umbrella surfaces, asphalt roads, leaves, leaves shadows, bushes, and weeds, which may contain features similar to the dense crowd. Most of existing methods consider the entire image while failing to restrict the area of the crowd, thereby leading to inaccurate predictions. To address this problem, Scale Aggregation Network (SANet) [13] uses a combined loss that considers the local correlation and Adversarial Cross-Scale Consistency Pursuit (ACSCP) [29] adopts a scale-consistency regularizer that enforces the sum-up of people counts from local patches to be the same as the overall counts of the whole region.

B. Contributions

In this paper, we propose a frequency feature pyramid network (FFPNet) to address the above challenges. Different from previous approaches of extracting multi-scale information from the spatial domain, we replace the pooling and up-sampling operations within the feature pyramid using discrete cosine transformation (DCT) and inverse discrete cosine transformation (IDCT) to extract the multi-scale information from the frequency domain, as shown in Fig. 2. Meanwhile, as shown in recent studies [30] and [31], the attention mechanism can enhance the valuable features in the region of people gathered. Inspired by these studies, we also incorporate the attention mechanism to improve the scale expression ability for feature maps. Moreover, we design a global-local consistency loss, which contains weighted log-cosh loss and the gradient magnitude similarity. The weighted log-cosh loss mitigates the drastic gradient changes caused by the imbalanced data distribution which is not of interest to researchers. Meanwhile, it also represents the global pattern of the density map. The gradient magnitude similarity is used to refine the insufficient local patterns. Details about our FFPNet will be given in Section III and Section IV.

Furthermore, most of existing datasets only contain a single type of objects (either vehicles or persons) in a scene. Therefore, they are not suitable for more complex scenarios of crowd-and-vehicle counting. To this end, we introduce a hybrid crowd-and-vehicle dataset (CROVEH) that includes 7,587 persons and 6,361 vehicle annotations. It allows us to estimate the number of multiple targets in the scene at the

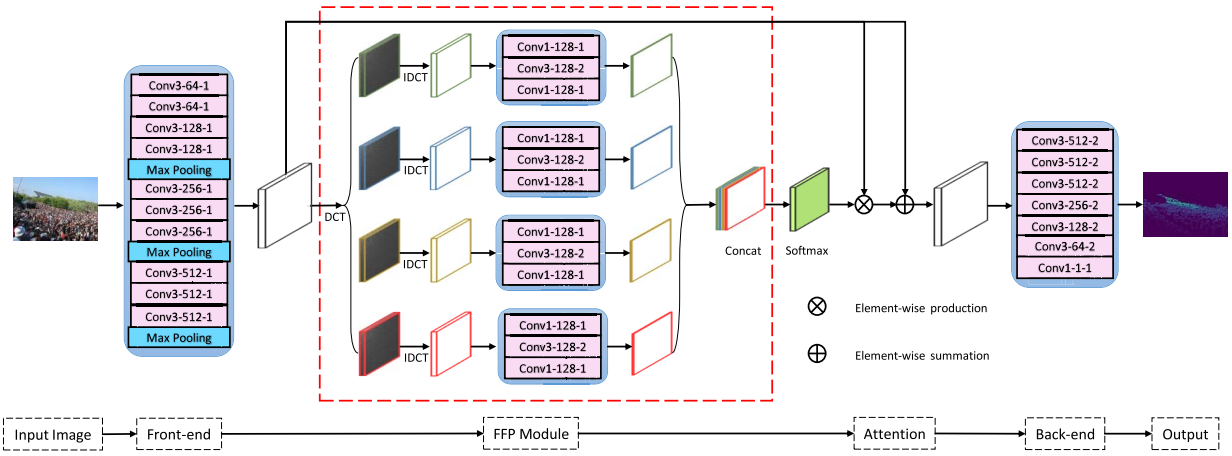

Fig. 2. Architecture of our FFPNet, which consists of front-end, FFP module, attention module, and back-end.

same time, especially for complex scenarios in autonomous driving and intelligent transportation.

Our contributions can be summarized as follows:

- We propose a module called frequency feature pyramid to address large-scale variations. It uses 3D DCT and IDCT to extract the multi-scale contextual information from the frequency domain. We also utilize the attention mechanism to enhance the response of the crowd in complex scenarios.
- We design a global-local consistency loss, which weakens the drastic gradient changes resulted from imbalanced data distribution and refines the local patterns while keeping the global prior.
- We construct CROVEH dataset, which allows us to estimate the density map of multiple types of targets in complex scenes at the same time.
- We conduct extensive experiments on six challenging datasets: CROVEH, TRANCOS [26], UCF-QNRF [25], UCSD [24], UCF_CC_50 [14] and ShanghaiTech [12]. Experimental results show that our FFPNet achieves superior performance on these datasets (refer to Section IV).

II. Related Work

In the following, we review crowd (or vehicle) counting methods from deep learning approaches, which mainly include methods to extract multi-scale information, methods to add additional supervision information, and novel strategies.

The aim of effective crowd-counting methods [12]–[19], [32], [33] is to enhance the density map response by extracting multi-scale information. Cao *et al.* [13] uses a scale aggregation module composed of multiple convolution kernels of different sizes to extract multi-scale information. Hossain *et al.* [17] Introduce multi-scale feature extractors with different convolution filter sizes to capture multi-scale information, and use the attention mechanism on different scales. Jiang *et al.* [15] construct a Multi-scale Encoder with different sizes of convolution kernels in trellis encoder-decoder network (TEDnet) to extract multi-scale information and uses a multi-path decoder to fuses multi-scale information. Scale pyramid network (SPN) [16] employs multiple sizes dilated convolutions in parallel instead of traditional convolutions with different sizes to extracting multi-scale information. In summary, all the above methods use different sizes of convolutions to extract multi-scale information from the spatial domain. By contrast, our method uses DCT and IDCT to extract multi-scale information from the frequency domain. Jiang *et al.* [32] utilized multi-scale information by generating and using different scaling factors for areas with different densities. Zhao *et al.* [33] propose a new depth embedding module to exploit the depth cues to obtain multi-scale information. Both of these two works proposed a novel method of using multi-scale information. Kang and Chan [18] propose a method of dividing the input image into multiple scales before training to obtain multi-scale information. In contrast to our multi-scale information acquisition in the network, this work is to obtain multi-scale information by constructing image pyramids in data preprocessing. Tian *et al.* [34] proposed a transformer with CNN, which consists of a pyramid vision transformer backbone, a pyramid feature aggregation module and an efficient regression head with multi-scale dilated convolution.

Moreover, adding additional supervision information is also a common way to improve the accuracy of estimation. Cao *et al.* design SANet [13], which consists of the encoder that captures multi-scale information by scale aggregation modules and the decoder that uses transposed convolutions to generates higher quality density map. Shen *et al.* [29] design a novel scale-consistency loss, which causes the total crowd counts of the local patch to be consistent with the total crowd counts of the regional union. In [15], a new combination loss is used to measure the map similarity with respect to local coherence and spatial correlation. Although these methods considered the local correlation of the density map, they ignored the imbalanced data distribution on the datasets.

Some researchers have proposed effective training strategies to generate higher quality density maps. Sam *et al.* [35] propose a hierarchical clustering method, which constructs multiple different image clusters and creates a group of expert CNNs for each cluster. Shi *et al.* [36] propose to use deep negative correlation learning to generalize features. Sam *et al.* [37] develop a switch CNN that can automatically

transfer patches to the corresponding regressor. Shi *et al.* [38] propose a novel perspective-aware CNN, which integrates the perspective information to solve the large-scale variations of persons.

III. Methodology

A. Architecture

Fig. 2 depicts the proposed FFPNet, in which we also adopt VGG-16 [39] due to its powerful and effective network structure. We first train front 14 layers (both convolution and pooling layers) with the pre-trained model to extract lower semantic features. Meanwhile, we remove the first pooling layer to extract more valid contextual information. Moreover, we also use the FFP module to gather and emphasize the frequency-based multi-scale information. Each sub-network in the FFP module is denoted by "Conv(κ) − (γ) − (δ)", where κ, γ, and δ are the kernel size, the number of channels, the dilation rate, respectively. Finally, we deploy the dilated convolution to aggregate long-ranged information while keeping the output resolution for generating a high-quality density map.

Inspired by [30] and [31], we fuse and enhance the multi-scale information using the attention mechanism as shown in Fig. 2. The input feature map is denoted by x. The output at spatial position i and channel c is denoted by $F_{i,c}(x)$. We then have the following expression for $F_{i,c}(x)$

$$F_{i,c}(x) = \left(1 + H_{i,c}(x)\right) \times G_{i,c}(x), \tag{1}$$

where $H(x)$ is the learnable mask with a range from 0 to 1 and $G(x)$ is the output of the front-end sub-network. Through the concatenation operation, we fuse four scales of feature maps. After using softmax, element-wise production, and element-wise summation, we emphasize significant features.

B. Frequency Feature Pyramid Module

One of the effective methods for extracting multi-scale information is to construct the spatial feature pyramid through down-sampling operations, as shown in the upper part of Fig. 3. As an orthogonal transformation, DCT captures the important multi-scale information of the given images more directly and efficiently while avoiding learning extra parameters. Therefore, we propose the FFP module to extract the multi-scale information by applying different coefficients of DCT and IDCT. Our FFP module retains more detailed information on the corresponding scale than the spatial feature pyramid.

As shown in Fig. 3, the spatial-based multi-scale information is generated by down-sampling and up-sampling. The down-sampling is used to build the feature pyramid of different scales while the up-sampling is deployed to keep the feature maps that have the same size. The frequency-based multi-scale information is obtained by coding of DCT with low-pass filters and reconstructing of IDCT. Columns a, b, c, and d depict the difference between the spatial-based multi-scale information and the frequency-based multi-scale information with the same coefficient. Regarding Column a, the spatial feature map obtained by down-sampling the image to 1/4 and then up-sampling to the original size, the frequency feature

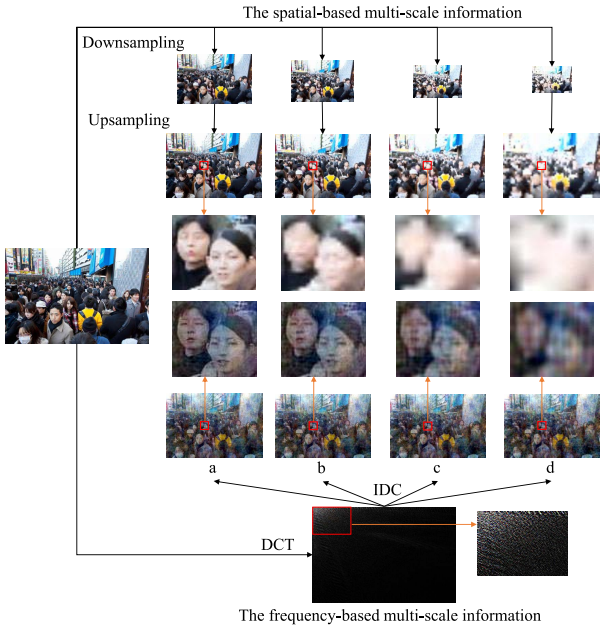

Fig. 3. Comparison between the spatial-based multi-scale information and the frequency-based multi-scale information. The corresponding coefficients for columns a, b, c, and d are 1/4, 1/16, 1/64, and 1/256, respectively. For the spatial-based multi-scale information, the coefficients indicate the scaling ratio. For the frequency-based multi-scale information, the coefficients represent the passing low-frequencies to reconstruct images. For facilitate viewing, the first row is not strictly displayed in proportion.

map is depicted by using the 1/4 low frequencies of DCT and the reconstruction of IDCT. Remaining Columns b, c, and d are similar operations to Column a. For the same coefficient feature map, we found that the feature map captured by frequency domain has more contextual information than the feature map generated by spatial domain, as shown in the second row and the third row of Fig. 3.

To capture the correlation between the feature map channels, the FFP module extracts the multi-scale information by using 3D DCT and IDCT. We denote the number of rows, columns, and channels of the feature map by N, M, and C, respectively. We denote the spatial domain signal at location (x, y, z) by $f(x, y, z)$ and the cosine transform coefficient at frequency (u, v, w) by $F(u, v, w)$. We then have the 3D DCT of $f(x, y, z)$ as follows,

$$\begin{aligned} F(u,v,w) = c(u)c(v)c(w) \sum_{x=0}^{N-1} \sum_{y=0}^{M-1} \sum_{z=0}^{C-1} f(x,y,z) \\ \cdot \cos\frac{(2x+1)u\pi}{2N} \cos\frac{(2y+1)v\pi}{2M} \\ \times \cos\frac{(2z+1)w\pi}{2C}, \end{aligned} \tag{2}$$

where u, v, and w are the frequency variables corresponding to different dimensions and $c(u)$, $c(v)$, and $c(w)$ are the constant terms. We have $c(v) = c(w) = c(u)$, where $c(u)$ is given by

$$c(u) = \begin{cases} \sqrt{\dfrac{1}{N}}, & \text{if } u = 0 \\ \sqrt{\dfrac{2}{N}}, & \text{otherwise.} \end{cases} \tag{3}$$

The definition of IDCT is as follows,

$$f(x,y,z) = c(u)c(v)c(w)\sum_{u=0}^{N-1}\sum_{v=0}^{M-1}\sum_{w=0}^{C-1}F(u,v,w) \cdot \cos\frac{(2x+1)u\pi}{2N}\cos\frac{(2y+1)v\pi}{2M} \times \cos\frac{(2z+1)w\pi}{2C}. \quad (4)$$

As shown in Fig. 2, the FFP module has four branches with different frequency coefficients, i.e., 1/4, 1/16, 1/64, 1/256. Each branch has 1×1 convolution and 3×3 convolution where the 1×1 convolution is a powerful operation to reduce the number of parameters. Therefore, we use 1×1 convolution to compress the channels in each branch. The 3×3 convolution layer is used to extract high-level semantic features. The integration of high-level semantic features uses the concatenation operation to restore the number of channels.

C. Global-Local Consistency Loss

In order to address both the imbalanced-data distribution problem and the insufficient local-pattern problem, we propose a novel global-local consistency loss. Firstly, we use the weighted log-cosh loss to overcome the imbalanced data distribution. Then, the gradient magnitude similarity is used to constrain the response of crowd or vehicle features on the local pattern. We then design two forms of the global-local consistency loss to illustrate how to solve the imbalanced data distribution and the insufficient local patterns.

1) Weighted log-cosh*:* For imbalanced data distribution, the outliers produce the excessively large gradients that are harmful to training. SmoothL_1 [27] that only focuses on the negative influence of outliers is not differentiable everywhere. Therefore, our proposed weighted $\log-\cosh$ loss can not only reduce the gradient of outliers and increase the gradient of interior points, but also ensure that it is differentiable everywhere. The weighted log-cosh loss is defined as follows,

$$L_{\text{wlc}} = \frac{1}{N}\sum_{i=1}^{N}\alpha\log\left(\cosh\left(Y_i - Y_i^P\right)\right), \quad (5)$$

where α is 1.313. When L_{wlc} satisfies $\mathsf{error} = 1$ and $\mathsf{gradient} = 1$, we can get $\alpha = 1.313$. This ensures that L_{wlc} has a larger gradient of inliers than SmoothL_1 though the cost of the outlier gradient only increases a little, as shown in Fig. 4. Compare to SmoothL_1, L_{wlc} not only increases the gradient of inliers but also controls the gradient of outliers to be a small range. This implies the increased overall gradients, thereby speeding up the convergence of the model.

2) Gradient Magnitude Similarity: The gradient of the image is sensitive to image distortion. Xue *et al.* [40] propose the gradient magnitude similarity (GMS) to measure the image similarity by capturing local gradient changes. GMS has better efficiency, in both speed and performance. Therefore, we use GMS to measure the local correlation between the estimated density map and ground truth density map.

Given an input image, GMS is defined as the root mean square of directional gradients of the image along both vertical and horizontal directions. The gradient is computed using the convolutional operation with the 3×3 Prewitt filters. We denote the gradient magnitude of the estimated density map Y^P and ground truth density map Y by m_{Y^P} and m_Y, respectively. They can be computed by convolving h_x and h_y on Y^P and Y, where h_x and h_y indicate the horizontal directions and the vertical directions of Prewitt filters, respectively. We then have the definition of m_{Y^P} as follows,

$$m_{Y^P} = \sqrt{\left(Y^P\otimes h_x\right)^2 + \left(Y^P\otimes h_y\right)^2}. \quad (6)$$

Similarly, we have the definition of m_Y as follows,

$$m_Y = \sqrt{\left(Y\otimes h_x\right)^2 + \left(Y\otimes h_y\right)^2}, \quad (7)$$

where symbol $\otimes$ denotes the convolution operation.

Then GMS (with range from 0 to 1) is computed as follows,

$$\mathsf{GMS} = \frac{2m_{Y^P}\times m_Y + c}{m_{Y^P}^2 + m_Y^2 + c}, \quad (8)$$

where c is the positive constant that avoids division by zero. A higher GMS means a higher similarity to a ground truth density map. The gradient magnitude similarity loss denoted by L_{GMS} is defined as follows,

$$L_{\mathsf{GMS}} = \frac{1}{N}\sum_{i=1}^{N}\left(1-\mathsf{GMS}_i\right), \quad (9)$$

where the range of L_{GMS} is the same as GMS.

3) Global-Local Consistency Loss: To constrain the response of crowd or vehicle feature on the global and local pattern simultaneously, we combine the weighted log-cosh loss and the gradient magnitude similarity loss to construct the global-local consistency (GLC) loss. It has two forms of combination. The first form of the global-local consistency loss denoted by $L_{\mathsf{GLC}}^{(1)}$ is defined as follows:

$$L_{\mathsf{GLC}}^{(1)} = \frac{1}{N}\sum_{i=1}^{N}\left(\alpha\log\left(\cosh(Y_i - Y_i^P)\right) + \beta_1\left(1-\mathsf{GMS}_i\right)\right), \quad (10)$$

where β_1 is the weight to balance L_{wlc} and L_{GMS}. The term L_{wlc} enlarges the gradients of inliers and restricts the maximum gradient of outliers. It reflects the error of the estimated

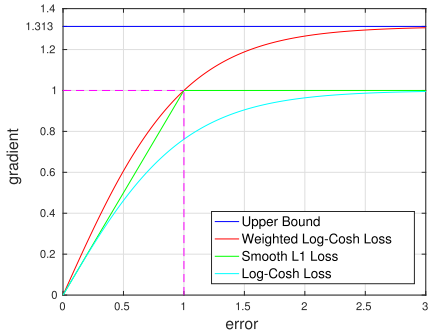

Fig. 4. The curves show that the inlier gradient of the weighted log-cosh loss is larger than SmoothL_1 loss while the outlier gradient of the weighted log-cosh loss has a small upper bound, i.e., 1.313.

density map on the global pattern. The term L_{GMS} constrains the similarity between the ground truth and the estimated density map on the local pattern. In experiments, we let α be 1.313 and β_1 be 1.6.

Because the generalization error of the crowd-and-vehicle counting task is measured by MAE and MSE. In order to make the empirical error closer to the generalization error, we design the second form of the global-local consistency loss denoted by $L_{\mathsf{GLC}}^{(2)}$ given as follows,

$$L_{\mathsf{GLC}}^{(2)} = \frac{1}{N}\sum_{i=1}^{N}\left(\alpha \log\left(\cosh(Y_i - Y_i^P) + \beta_2\left(1 - \mathsf{GMS}_i\right)\right)\right), \tag{11}$$

where α is 1.313 and β_2 is 1.0 in our experiments.

According to the definitions of two forms of global-local consistency loss as defined in Eq. (10) and Eq. (11), we find the following relation between them.

Lemma 1: The second form of the global-local consistency loss is no greater than the first form of the global-local consistency loss,

$$L_{\mathsf{GLC}}^{(2)} \leq L_{\mathsf{GLC}}^{(1)}, \quad \text{if} \quad \alpha\beta_2 \leq \beta_1. \tag{12}$$

Proof: Let $x = \cosh(Y_i - Y_I^P)$ and $y = 1 - \mathsf{GMS}(i)$. Suppose that the inequality holds for $x \geq 1$ and $0 \leq y \leq 1$. The above inequality can be represented as follows,

$$\alpha \log(x + \beta_2 y) \leq \alpha \log(x) + \beta_1\ y. \tag{13}$$

Rearranging it, we have,

$$\alpha \log(x + \beta_2 y) - \alpha \log(x) \leq \beta_1\ y. \tag{14}$$

We then have,

$$\alpha \log(1 + \beta_2 \frac{y}{x}) \leq \beta_1\ y. \tag{15}$$

We next obtain

$$\alpha \log(1 + \beta_2 \frac{y}{x}) \leq \alpha \frac{\beta_2 y}{x} \leq \beta_1\ y. \tag{16}$$

Inequality (15) is true. Therefore, inequality (12) is proved. ■

We observe that $L_{\mathsf{GLC}}^{(2)}$ has a lower upper bound of generalization deviation than $L_{\mathsf{GLC}}^{(1)}$. This means $L_{\mathsf{GLC}}^{(2)}$ will make the model obtain better generalization than $L_{\mathsf{GLC}}^{(1)}$ on testing dataset. We will show that $L_{\mathsf{GLC}}^{(2)}$ is better than $L_{\mathsf{GLC}}^{(1)}$ through experiments in Section IV.

IV. Experiments

A. Implementation Details

At present, the method of the density map estimation is the most effective for the crowd-and-vehicle counting task. Our density map generation method follows MCNN [12]. Parameter setting follows CSRNet [11]. We crop each image into nine patches. Each patch is 1/4 of the original image. The four quarters and the centre are the first five patches. The remaining four patches are located in the left-centre, right-centre, top-centre and bottom-centre of the original image. After flipping all of the nine patches, we get 18 patches. Because some patches have fewer persons (or even zero), we remove those patches containing persons smaller than the minimum number of persons of the dataset, e.g., the patches with a number of persons smaller than 33 on ShanghaiTech part_A dataset are removed. Since TRANCOS [26], CROVEH, and UCSD datasets [24] are too sparse, we just flip samples without cropping.

We train FFPNet using first 14 layers of VGG-16 as our pre-trained model. Based on experience and experimental results, we use a Gaussian distribution with mean of 0 and standard deviation of 0.01 to initialize the rest layers. Meanwhile, Adam optimizer is deployed to training. The learning rate is fixed at $5e-6$, momentum is 0.95 and decay is $5e-4$.

B. Evaluation Metrics

We use MAE and MSE to assess the performance of the model, where MAE indicates the accuracy of the estimated crowd (or vehicle) count and MSE is a measure of the robustness. For testing images, MAE and MSE are defined as $\mathsf{MAE} = \frac{1}{N}\sum_{i=1}^{N}\left|C_i - C_i^{\mathrm{GT}}\right|$ and $\mathsf{MSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(C_i - C_i^{\mathrm{GT}}\right)^2}$, where N is the number of images in the testing set, C_i is the predicted count of persons (or vehicles) in the i-th testing image, and C_I^{GT} is the corresponding ground truth count. Particularly, C_i is defined as $C^i = \sum_{h=1}^{H}\sum_{w=1}^{W} p_{h,w}$, where H and W indicate the height and width of the density map, C^i is the pixel value at location (h, w) of the density map.

C. Datasets

We choose six crowd-and-vehicle counting datasets:

- *CROVEH dataset* has 426 samples that contain both 7,587 persons and 6,361 vehicle annotations. Each sample has at least 11 persons and 11 vehicles, and at most 98 persons and 38 vehicles. The average numbers of persons and vehicles are 17.8 and 14.9. These samples are selected from nuImages [44] and Cityscapes [45] datasets. We eliminate ego vehicles, bicycles, and motorcycles from vehicle annotations. We also change the annotations of riders to persons. Due to the large size of the original images, we scale each image to a half of its original size. We divide the source dataset into the training set and testing set. CROVEH dataset can reflect the real congestion on the street. This may help the context awareness and context prediction of intelligent transportation systems.
- *TRANCOS dataset* [26] is a public traffic and congestions dataset, which includes 1,244 images with 46,797 vehicle annotations for different congested scenes. The perspectives of images are not fixed and the images are collected from very different scenarios. The region of interest (ROI) is also provided. Grid Average Mean Absolute Error (GAME) [26] is used to assess the counting accuracy.
- *ShanghaiTech Dataset* [12] contains 330,165 annotations in 1,198 images. This dataset includes two parts: SHA has 482 images of congested scenes captured from the Internet; SHB consists of 712 sparse scenes got from urban streets. We follow the splits of the training and testing sets provided by the authors for our experiments.

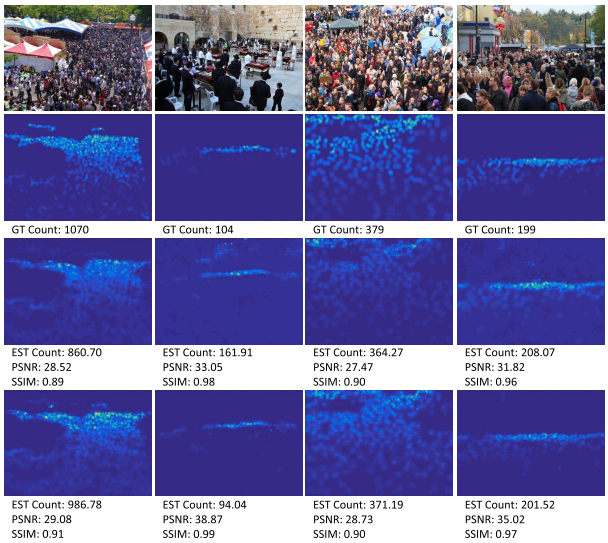

Fig. 5. The first row shows samples from testing sets of UCF-QNRF and SHA. The second row shows the ground truth for each sample. The third and fourth rows show the density maps predicted by CSRNet [11] and FFPNet, respectively.

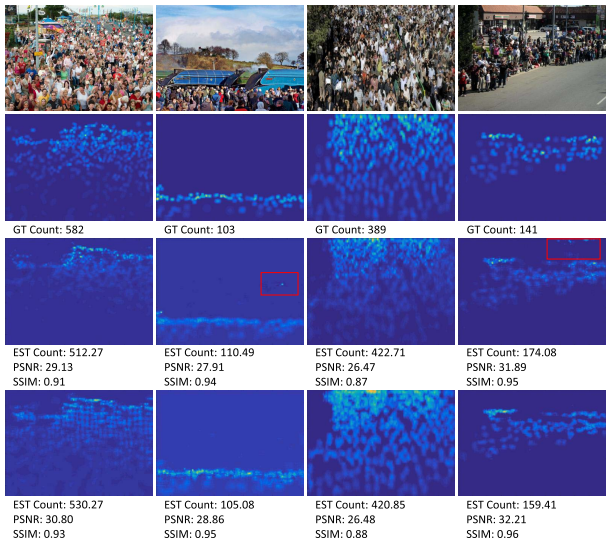

Fig. 6. The first row shows samples from the testing set of UCF-QNRF and SHA. The second row shows the ground truth for each sample. The third and fourth rows show the density maps predicted by FFPNet [11] and FFPNet with the GLC loss, respectively.

- *UCF-QRNF Dataset* [25] collects 1,535 dense-crowd images with 1,251,642 annotations from the Internet, which is divided into training and test splits of 1,201 and 334 images, respectively.
- *UCF_CC_50 Dataset* [14] contains 50 extremely-dense crowd images with 64,000 annotations. The number of persons per image is between 94 to 4,543 and the average count is 1,280. We also conduct 5-fold cross-validation suggested by [14].
- *UCSD dataset* [24] has 2,000 images collected from a pedestrian walkway using surveillance cameras. These sparse crowds range from 11 to 43 have an average of 25. We follow the setting of [24] to divide the dataset with 2,000 frames: taking the frames from 601 to 1,400 as the training split and the rest as the testing split. ROI is provided as well for the whole dataset.

D. Results and Analysis

Figs. 5 and 6 present representative estimated density maps for the test scenes from ShanghaiTech [12] and UCF-QNRF [25] datasets. In Fig. 5, FFPNet utilizes the FFP module to strengthen the crowd or vehicle response and eliminate the incorrect feature, implying the effectiveness of the FFP module in extracting multi-scale information. As shown in the second and fourth columns of Fig. 6, the GLC loss eliminates the inaccurate response in red boxes. In the first and third columns, density maps estimated using GLC loss have more obvious responses locally, even on large-scale targets, indicating that the GLC loss can handle insufficient local patterns.

We evaluate the overall performance of our approach in comparison to existing approaches on these six datasets as shown in Table I. Compared with existing methods, our FFPNet achieves superior performance in crowd-and-vehicle counting. For example, compared with CSRNet [11], FFPNet

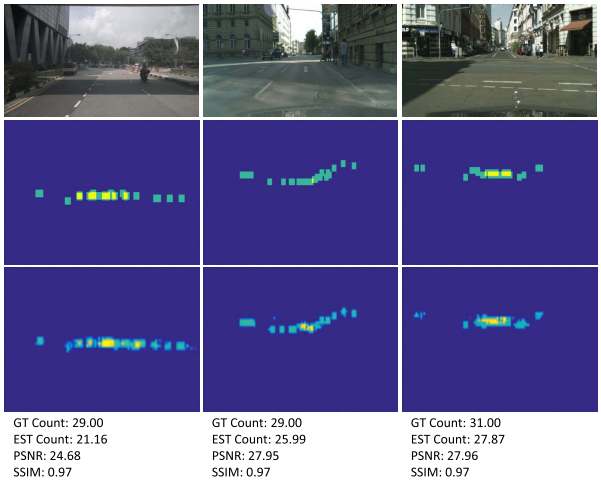

Fig. 7. The first row shows samples from the testing set of CROVEH. The second row shows the ground truth for each sample. And the third row shows the predicted density map.

reduces MAE and MSE by 8.2% and 6.5%, respectively on CROVEH_C (crowd) dataset; it reduces MAE and MSE by 25.1% and 25.6%, respectively on CROVEH_V (vehicle) dataset. On CROVEH dataset, FFPNet can estimate the distribution of crowd and vehicles separately at the same time. In order to facilitate the display of congestion, we merged the crowd density map and vehicle density map into one, as shown in Fig. 7. Particularly, FFPNet outperforms other methods in estimating the density map of crowd and vehicles. On TRANCOS [26], we conduct another experiment to count the number of vehicles. The comparison results are shown in Table II. Fig. 8 illustrates some examples of the predicted density maps. Compared with existing methods, our approach achieves the best performance in GAME(0), GAME(1), and

TABLE I

COMPARISON WITH STATE-OF-THE-ART METHODS ON SHANGHAITECH, UCF-QNRF, UCF_CC_50, UCSD, CROVEH DATASETS

	SHA		SHB		UCF-QNRF		UCF_CC_50		UCSD		CROVEH_C		CROVEH_V	
Methods	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE
MCNN [12]	110.2	173.2	26.4	41.3	227.0	426.0	377.6	509.1	1.07	1.35	37.1	47.1	26.1	32.1
Switch-CNN [37]	90.4	135.0	21.6	33.4	228.0	334.0	318.1	439.2	1.62	2.10	-	-	-	-
CL-CNN [25]	-	-	-	-	132.0	191.0	-	-	-	-	-	-	-	-
ACSCP [29]	75.7	102.7	17.2	27.4	-	-	291.0	404.6	1.04	1.35	-	-	-	-
Deep-NCL [36]	73.5	112.3	18.7	26.0	-	-	288.4	404.7	-	-	-	-	-	-
IG-CNN [35]	72.5	118.2	13.6	21.1	-	-	291.4	349.4	-	-	-	-	-	-
CSRNet [11]	68.2	115.0	10.6	16.0	-	-	266.1	397.5	1.16	1.47	20.7	29.1	24.7	31.3
SANet [13]	67.0	104.5	8.4	13.6	-	-	258.4	334.1	-	-	-	-	-	-
SPN [16]	61.7	99.5	9.4	14.4	-	-	259.2	335.9	1.03	1.32	-	-	-	-
TEDnet [15]	64.2	109.1	8.2	12.8	113	188	249.4	354.5	-	-	-	-	-	-
ADCrowdNet [41]	63.2	98.9	8.2	15.7	-	-	266.4	258.0	1.39	1.68	-	-	-	-
CAN [19]	62.3	100.0	7.8	12.2	107	183	212.2	243.7	-	-	-	-	-	-
SMANet [42]	59.7	102.1	7.3	12.9	92.5	176.7	178.4	256.3	-	-	-	-	-	-
ASNet [32]	57.78	90.13	-	-	**91.59**	**159.71**	174.84	251.63	-	-	-	-	-	-
PDANet [43]	58.5	93.4	**7.1**	10.9	-	-	**119.8**	**159**	**0.93**	1.21	24.8	33.5	23.1	29.8
FFPNet	**57.5**	**88.0**	7.2	**10.7**	97.8	168.7	210.8	240.0	0.97	**1.20**	**19.0**	**27.2**	**18.5**	**23.3**

CROVEH_C denotes the CROVEH (crowd) and **CROVEH_V** denotes the CROVEH (vehicle).

TABLE II

COMPARISON ON TRANCOS DATASET

Methods	GAME0	GAME1	GAME2	GAME3
Fiaschi et al. [46]	17.77	20.14	23.65	25.99
Lempitsky et al. [47]	13.76	16.72	20.72	24.36
Hydra-3s [48]	10.99	13.75	16.69	19.32
FCN-HA [49]	4.21	-	-	-
CSRNet [11]	3.56	5.49	8.57	15.04
ADCrowdNet [41]	2.69	4.61	7.13	14.14
SPN [16]	3.35	4.94	6.47	**9.22**
FFPNet	**2.17**	**4.41**	**5.50**	11.13

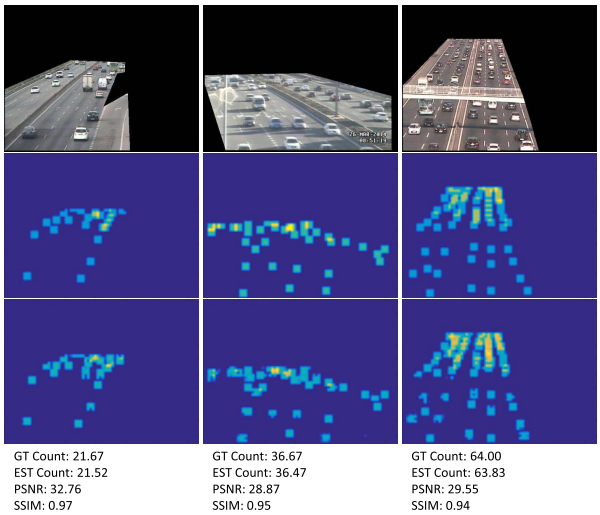

Fig. 8. The first row shows samples from the testing set of TRANCOS with ROI. The second row shows the ground truth for each sample. The third row shows the predicted density map.

GAME(2) (except for GAME(3)), e.g., 19.3% lower than the second best result in GAME(0); this demonstrates that our approach is more robust and generalized than other methods.

SHA [12], UCF-QNRF [25], and UCF_CC_50 datasets [14] contain a lot of extremely-dense samples, leading to the imbalanced data distribution. They also suffer from large-scale variations caused by camera heights and angles. On SHA [12], we get 7.7% MAE lower than CAN [19] and 12% lower than ADCrowdNet [41]. Our approach reduces MAE and MSE by 8.6% and 7.8% on UCF-QNRF, respectively. Since UCF_CC_50 dataset [14] only contains 50 samples, it is difficult for training supervised DL models though our FFPNet still achieves competitive results. It illustrates that our approach can better handle extremely-dense scenes, imbalanced data distribution, and large-scale variations. For the relatively-sparse datasets like SHB [12], the samples are disturbed by the complex background in streets. Our approach delivers 5.2% MAE lower than ADCrowdNet [41] and 12.3% MSE lower than CAN [19]. This validates our approach can reduce the negative influence of complex background. For UCSD dataset [24] with extremely-sparse scenes, we get 1.0% MAE lower than ADCrowdNet [41] and much lower MSE than MMCNN [50]. These experiments demonstrate that our approach achieves superior performance not only on extremely-dense scenes but also on sparse scenes. Therefore, our approach is better in dealing with crowd estimation in dense scenes. Compared with the methods such as MCNN [12], CSRNet [11], and CAN [19], our approach can better extract and utilize multi-scale information, consequently solving imbalanced data distribution and insufficient local patterns.

E. Ablation Experiments

1) Ablation Studies on Architecture: To analyse the effectiveness of our model, we conduct a series of ablation experiments with several configurations including CSRNet [11], P2PNet [51], data augmentation, attention mechanism, the FFP module, global-local consistency loss on SHA dataset. Table III presents the results. Firstly, CSRNet* reduces MAE and MSE by 2.8% and 7.3%, respectively in comparison to the baseline. With our data augmentation, CSRNet further reduces MAE and MSE by 1.0% and 4.6%, respectively. The experimental results show that it is effective for model training by removing the patches with fewer persons to eliminate the negative effects. We then use this model as our new baseline and fine-tune it in the following experiments. In particular, FFPNet without (w/o) attention mechanism increases MAE

TABLE III

ABLATION STUDIES OF ARCHITECTURE ON SHA DATASET. FFPNET INDICATES FFPNET3D WHICH USE 3D DCT AND IDCT. THE SYMBOL * MEANS OUR RE-IMPLEMENTS

Settings	Backbone	MAE	MSE
Baseline	CSRNet	68.2	115.0
CSRNet *	CSRNet	66.28	106.66
CSRNet+dataAug	CSRNet	65.60	101.79
FFPNet1D+dataAug	CSRNet	63.45	100.37
FFPNet2D+dataAug	CSRNet	63.14	101.50
FFPNet+dataAug	CSRNet	62.11	98.99
FFPNet (SMS)+dataAug	CSRNet	65.08	101.75
FFPNet (w/o attention)+dataAug	CSRNet	63.69	101.12
FFPNet+dataAug+$L_{\text{GLC}}^{(2)}$ loss	CSRNet	**57.46**	**88.04**
Baseline	P2PNet	52.70	85.10
P2PNet *	P2PNet	53.69	87.90
P2PNet w/ FFP	P2PNet	**52.69**	**84.89**

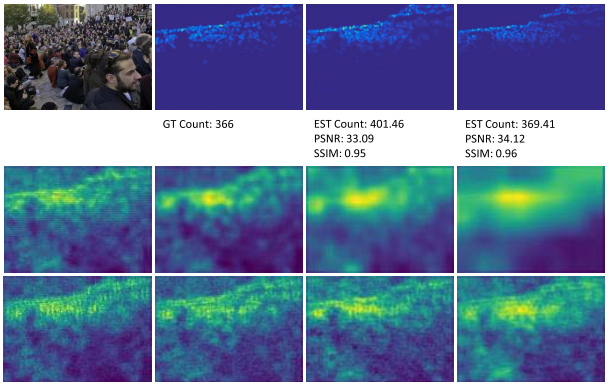

Fig. 9. The first row shows the testing image from SHA, the ground truth, the estimated density map from FFPNet with SMS and FFPNet. The second and the third row show the density maps of four branches from FFPNet with SMS and FFPNet, respectively.

and MSE by 2.9% and 0.7%, respectively, implying that the attention mechanism can significantly enhance the valid features of the crowd data. Moreover, FFPNet with data augmentation also reduces MAE and MSE by 2.5% and 2.1%, respectively, implying that our FFPNet can effectively extract more multi-scale information by FFP module. At last, FFPNet with the proposed global-local consistency loss has achieve the best performance in terms of 7.5% MAE and 11.1% MSE decrement in comparison to the variant without the global-local consistency loss. This result demonstrates that our global-local consistency loss can alleviate the influence of imbalanced data distribution and insufficient local patterns. To further demonstrate the effectiveness of FFPNet, we conduct experiments using P2PNet, one of the best-performing methods, as the baseline. P2PNet estimates the number of people by predicting a set of heads in the image directly. In our experiments, P2PNet with the FFP module achieves better results, in terms of 0.02% MAE and 0.25% MSE decrement than P2PNet. This shows that FFPNet can effectively extract and utilize multi-scale information on non-heatmap regression counting. Compared with the state-of-the-art CCTrans [34], P2PNet with FFP module still reduces MSE by 0.02%.

2) Ablation Studies on Feature Pyramid: Moreover, we also study different multi-scale configurations, as shown in Table III. In order to study the effectiveness of DCT and IDCT by replacing the pooling and up-sampling operations. We construct a spatial multi-scale module (SMS) using pooling and up-sampling operations to extract the multi-scale information. We observe that FFPNet with SMS achieves 2.8% higher MSE and 4.8% higher MAE than FFPNet (DCT), respectively, implying that FFPNet with DCT captures the multi-scale information more effectively than FFPNet with SMS. Fig. 9 presents visualized results to further demonstrate this effect. In Table III, FFPNet1D and FFPNet2D represent using 1D DCT with IDCT and 2D DCT with IDCT to construct the feature pyramid of the frequency-based multi-scale information, respectively, where 1D transformation indicates the operations on the columns of the feature map and 2D transformation indicates the operations consisting of separated row transformation and column transformation on the entire feature map. Similarly, FFPNet denotes FFPNet3D, where 3D transformation uses 1D transformation on the channel dimension based on 2D transformation. The results of FFPNet are better than FFPNet2D and FFPNet1D, suggesting that FFPNet with the channel dimension can effectively extract more high-level semantic features. The final values of MAE and MSE reach 62.11 and 98.99, respectively.

TABLE IV

COMPARISONS FOR DIFFERENT LOSS FUNCTIONS ON SHA DATASET

Methods	MAE	MSE
L_{MSE}	62.11	98.99
SmoothL_1 [27]	63.13	97.70
BalancedL_1 [28]	62.68	98.47
L_{wlc}	62.36	97.39
L_{SANet} [13]	63.11	99.93
$L_{\text{GLC}}^{(1)}$	58.04	90.40
$L_{\text{GLC}}^{(2)}$	**57.46**	**88.04**
$L_{\text{GLC}}^{(3)}$	62.82	96.47

TABLE V

ABLATION STUDIES OF GLOBAL-LOCAL CONSISTENCY LOSS FUNCTIONS ON SHA DATASET

	$L_{\text{GLC}}^{(1)}$		$L_{\text{GLC}}^{(2)}$	
Settings	MAE	MSE	MAE	MSE
$\beta = 0.6$	58.04	90.34	59.05	92.10
$\beta = 0.8$	58.12	91.86	59.41	90.90
$\beta = 1.0$	58.08	96.98	**57.46**	**88.04**
$\beta = 1.2$	59.91	95.27	60.12	97.16
$\beta = 1.4$	59.98	97.09	60.49	96.76
$\beta = 1.6$	**57.54**	**89.39**	60.68	94.71
$\beta = 1.8$	58.62	92.91	61.57	93.12
$\beta = 2.0$	58.36	93.06	59.60	92.77

3) Ablation Studies on Global-Local Consistency Loss: Finally, we compare nine types of losses on SHA, as shown in Table IV. Table V also presents the ablation studies of Global-Local Consistency loss. In particular, SmoothL_1 and L_{wlc} slightly increase MAE than L_{MSE}, implying that there are a few outliers in SHA dataset. Libra R-CNN [28] proposes the Balanced BalancedL_1 loss to solve the sampling imbalance problem. SANet [13] proposes a joint loss to combine MSE and weighted structural similarity index measure (SSIM) to enforce the local structural similarity. The experiment results show that FFPNet with L_{SANet} has no improvement than L_{MSE} (i.e., decreasing MAE and MSE). When we use L_{GMS} to

constrain the local correlation of density map, both $L_{\mathsf{GLC}}^{(1)}$ and $L_{\mathsf{GLC}}^{(2)}$ contribute to the performance improvement. Compared with $L_{\mathsf{GLC}}^{(1)}$, $L_{\mathsf{GLC}}^{(2)}$ get the best performance. It validates $L_{\mathsf{GLC}}^{(2)}$ has the lower upper bound of generalization deviation than $L_{\mathsf{GLC}}^{(1)}$. We further define the global-local consistency loss with SSIM as follows,

$$L_{\mathsf{GLC}}^{(3)} = \frac{1}{N}\sum_{i=1}^{N}\left(\alpha \log\left(\cosh(Y_i - Y_i^P) + \beta_3\left(1 - \mathsf{SSIM}\right)\right)\right), \tag{17}$$

where $\alpha = 1.3$. Because $L_{\mathsf{GLC}}^{(3)}$ and $L_{\mathsf{GLC}}^{(2)}$ are the same in form, we set $\beta_3 = 1.0$. This result demonstrates that SSIM does not apply to our approach compared to GMS. From Table V, we observe that β_1=1.6 is the best for $L_{\mathsf{GLC}}^{(1)}$ and β_2=1.0 for $L_{\mathsf{GLC}}^{(2)}$. The best performance is achieved by $L_{\mathsf{GLC}}^{(2)}$ at β_2=1.0, i.e., MAE is 57.46 and MSE is 88.04.

V. Conclusion

In this paper, we propose an effective crowd-and-vehicle counting approach that consists of the FFPNet and global-local consistency loss. The FFPNet uses FFP module to effectively extract more multi-scale contextual information. The global-local consistency loss can alleviate the negative influence caused by the large gradient of the outliers and improve the local pattern correlation of estimated density map. Extensive experiments show that our FFPNet is robust and generalized in crowd-and-vehicle counting. Meanwhile, it can effectively produce high-quality density maps, which depict the distribution of crowd and vehicles in the natural scenes. We also construct the crowd-and-vehicle dataset for multi-target counting on a complex scene. It can help us estimate the number of persons and vehicles at the same time. Finally, we conduct several groups of experiments to evaluate the proposed FFPNet in comparison with recent state-of-the-art (SOTA) methods. The results demonstrate that FFPNet achieves the best results on different backbones, e.g., SOTA performance in terms of 52.69 MAE on P2PNet with FFP module.

References

[1] L. Claussmann, M. Revilloud, D. Gruyer, and S. Glaser, "A review of motion planning for highway autonomous driving," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 5, pp. 1826–1848, May 2020.

[2] J. Guo, U. Kurup, and M. Shah, "Is it safe to drive? An overview of factors, metrics, and datasets for driveability assessment in autonomous driving," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 8, pp. 3135–3151, Aug. 2019.

[3] C. Chen, Z. Liu, S. Wan, J. Luan, and Q. Pei, "Traffic flow prediction based on deep learning in Internet of Vehicles," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 6, pp. 3776–3789, Jun. 2021.

[4] S. Wan, X. Xu, T. Wang, and Z. Gu, "An intelligent video analysis method for abnormal event detection in intelligent transportation systems," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 7, pp. 4487–4495, Jul. 2021.

[5] M. Usman, M. A. Jan, and A. Jolfaei, "SPEED: A deep learning assisted privacy-preserved framework for intelligent transportation systems," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 7, pp. 4376–4384, Jul. 2021.

[6] C. Chen, L. Liu, S. Wan, X. Hui, and Q. Pei, "Data dissemination for industry 4.0 applications in Internet of Vehicles based on short-term traffic prediction," *ACM Trans. Internet Technol.*, vol. 22, no. 1, pp. 1–18, Feb. 2022.

[7] Y. Wu, H. Guo, C. Chakraborty, M. Khosravi, S. Berretti, and S. Wan, "Edge computing driven low-light image dynamic enhancement for object detection," *IEEE Trans. Netw. Sci. Eng.*, early access, Feb. 14, 2022, doi: 10.1109/TNSE.2022.3151502.

[8] W. Wei, R. Yang, H. Gu, W. Zhao, C. Chen, and S. Wan, "Multi-objective optimization for resource allocation in vehicular cloud computing networks," *IEEE Trans. Intell. Transp. Syst.*, early access, Aug. 3, 2021, doi: 10.1109/TITS.2021.3091321.

[9] Q. Zhou, J. Zhang, L. Che, H. Shan, and J. Z. Wang, "Crowd counting with limited labeling through submodular frame selection," *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 5, pp. 1728–1738, May 2019.

[10] X. Ding, F. He, Z. Lin, Y. Wang, H. Guo, and Y. Huang, "Crowd density estimation using fusion of multi-layer features," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 8, pp. 4776–4787, Aug. 2021.

[11] Y. Li, X. Zhang, and D. Chen, "CSRNet: Dilated convolutional neural networks for understanding the highly congested scenes," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1091–1100.

[12] Y. Zhang, D. Zhou, S. Chen, S. Gao, and Y. Ma, "Single-image crowd counting via multi-column convolutional neural network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 589–597.

[13] X. Cao, Z. Wang, Y. Zhao, and F. Su, "Scale aggregation network for accurate and efficient crowd counting," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 734–750.

[14] H. Idrees, I. Saleemi, C. Seibert, and M. Shah, "Multi-source multi-scale counting in extremely dense crowd images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 2547–2554.

[15] X. Jiang *et al.*, "Crowd counting and density estimation by trellis encoder-decoder networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 6133–6142.

[16] X. Chen, Y. Bin, N. Sang, and C. Gao, "Scale pyramid network for crowd counting," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2019, pp. 1941–1950.

[17] M. Hossain, M. Hosseinzadeh, O. Chanda, and Y. Wang, "Crowd counting using scale-aware attention networks," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2019, pp. 1280–1288.

[18] D. Kang and A. Chan, "Crowd counting by adaptively fusing predictions from an image pyramid," 2018, *arXiv:1805.06115*.

[19] W. Liu, M. Salzmann, and P. Fua, "Context-aware crowd counting," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5099–5108.

[20] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path aggregation network for instance segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8759–8768.

[21] G. Ghiasi, T.-Y. Lin, and Q. V. Le, "NAS-FPN: Learning scalable feature pyramid architecture for object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7036–7045.

[22] M. Tan, R. Pang, and Q. V. Le, "EfficientDet: Scalable and efficient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10781–10790.

[23] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2117–2125.

[24] A. B. Chan, Z.-S. J. Liang, and N. Vasconcelos, "Privacy preserving crowd monitoring: Counting people without people models or tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–7.

[25] H. Idrees *et al.*, "Composition loss for counting, density map estimation and localization in dense crowds," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 532–546.

[26] R. Guerrero-Gómez-Olmedo, B. Torre-Jiménez, R. López-Sastre, S. Maldonado-Bascón, and D. Onoro-Rubio, "Extremely overlapping vehicle counting," in *Proc. Iberian Conf. Pattern Recognit. Image Anal.* Cham, Switzerland: Springer, 2015, pp. 423–431.

[27] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1440–1448.

[28] J. Pang, K. Chen, J. Shi, H. Feng, W. Ouyang, and D. Lin, "Libra R-CNN: Towards balanced learning for object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 821–830.

[29] Z. Shen, Y. Xu, B. Ni, M. Wang, J. Hu, and X. Yang, "Crowd counting via adversarial cross-scale consistency pursuit," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5245–5254.

[30] F. Wang *et al.*, "Residual attention network for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3156–3164.

[31] Y. Du, C. Yuan, B. Li, L. Zhao, Y. Li, and W. Hu, "Interaction-aware spatio-temporal pyramid attention networks for action classification," in *Proc. ECCV*, 2018, pp. 373–389.
[32] X. Jiang *et al.*, "Attention scaling for crowd counting," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 4706–4715.
[33] M. Zhao, C. Zhang, J. Zhang, F. Porikli, B. Ni, and W. Zhang, "Scale-aware crowd counting via depth-embedded convolutional neural networks," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 10, pp. 3651–3662, Oct. 2020.
[34] Y. Tian, X. Chu, and H. Wang, "CCTrans: Simplifying and improving crowd counting with transformer," 2021, *arXiv:2109.14483*.
[35] D. B. Sam, N. N. Sajjan, R. V. Babu, and M. Srinivasan, "Divide and grow: Capturing huge diversity in crowd images with incrementally growing CNN," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3618–3626.
[36] Z. Shi *et al.*, "Crowd counting with deep negative correlation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5382–5390.
[37] D. B. Sam, S. Surya, and R. V. Babu, "Switching convolutional neural network for crowd counting," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4031–4039.
[38] M. Shi, Z. Yang, C. Xu, and Q. Chen, "Revisiting perspective information for efficient crowd counting," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7279–7288.
[39] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.
[40] W. Xue, L. Zhang, X. Mou, and A. C. Bovik, "Gradient magnitude similarity deviation: A highly efficient perceptual image quality index," *IEEE Trans. Image Process.*, vol. 23, no. 2, pp. 684–695, Feb. 2014.
[41] N. Liu, Y. Long, C. Zou, Q. Niu, L. Pan, and H. Wu, "ADCrowdNet: An attention-injective deformable convolutional network for crowd understanding," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3225–3234.
[42] M. Wang, H. Cai, J. Zhou, and M. Gong, "Stochastic multi-scale aggregation network for crowd counting," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 2008–2012.
[43] S. Amirgholipour, X. He, W. Jia, D. Wang, and L. Liu, "PDANet: Pyramid density-aware attention net for accurate crowd counting," 2020, *arXiv:2001.05643*.
[44] H. Caesar *et al.*, "NuScenes: A multimodal dataset for autonomous driving," 2019, *arXiv:1903.11027*.
[45] M. Cordts *et al.*, "The cityscapes dataset for semantic urban scene understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3213–3223.
[46] L. Fiaschi, U. Köthe, R. Nair, and F. A. Hamprecht, "Learning to count with regression forest and structured labels," in *Proc. ICPR*, 2012, pp. 2685–2688.
[47] V. Lempitsky and A. Zisserman, "Learning to count objects in images," in *Proc. Adv. Neural Inf. Process. Syst.*, 2010, pp. 1324–1332.
[48] D. Onoro-Rubio and R. J. López-Sastre, "Towards perspective-free object counting with deep learning," in *Proc. ECCV*, 2016, pp. 615–629.
[49] S. Zhang, G. Wu, J. P. Costeira, and J. M. Moura, "FCN-RLSTM: Deep spatio-temporal neural networks for vehicle counting in city cameras," in *Proc. ICCV*, 2017, pp. 3667–3676.
[50] B. Yang, J. Cao, N. Wang, Y. Zhang, and L. Zou, "Counting challenging crowds robustly using a multi-column multi-task convolutional neural network," *Signal Process., Image Commun.*, vol. 64, pp. 118–129, Mar. 2018.
[51] Q. Song *et al.*, "Rethinking counting and localization in crowds: A purely point-based framework," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 3365–3374.

Xiaoyuan Yu received the B.S. degree in software engineering from the Beijing Institute of Technology, Zhuhai, China, in 2014, and the M.S. degree from the Macau University of Science and Technology (MUST) in 2017, where he is currently pursuing the Ph.D. degree. His research interests include computer vision and pattern recognition.

Yanyan Liang (Member, IEEE) received the B.S. degree from the Chongqing University of Posts and Telecommunications, Chongqing, China, in 2004, and the M.S. and Ph.D. degrees from the Macau University of Science and Technology (MUST), Taipa, Macau, in 2006 and 2009, respectively. He is currently an Associate Professor with MUST. He has published more than 40 papers related to pattern recognition, image processing, and computer vision in IEEE TRANSACTIONS and international conferences, including IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON CYBERNETICS, IEEE TRANSACTIONS ON MULTIMEDIA, IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY, IJCAI, CVPR, ICPR, and FG. He is also working on smart city applications with computer vision and big data. His research interests include computer vision, image processing, and machine learning.

Xuxin Lin received the B.S. degree in software engineering from the Beijing Institute of Technology, Zhuhai, China, in 2014, and the M.S. and Ph.D. degrees from the Macau University of Science and Technology (MUST) in 2016 and 2019, respectively. He is currently working as a Post-Doctoral Researcher with MUST. His research interests include computer vision and pattern recognition.

Jun Wan (Senior Member, IEEE) received the B.S. degree from the China University of Geosciences, Beijing, China, in 2008, and the Ph.D. degree from the Institute of Information Science, Beijing Jiaotong University, Beijing, in 2015. Since January 2015, he has been a Faculty Member with the National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Science (CASIA), China, where he currently serves as an Associate Professor. His main research interests include computer vision and machine learning. He is an Area Chair of ICME 2021 and 2022. He served as the Co-Editor for a special issues on IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE and IEEE TRANSACTIONS ON BIOMETRICS, BEHAVIOR, AND IDENTITY SCIENCE. He is an Associate Editor of the *IET Biometrics* from 2020 to 2022.

Tian Wang received the B.Sc. and M.Sc. degrees in computer science from Central South University in 2004 and 2007, respectively, and the Ph.D. degree in computer science from the City University of Hong Kong in 2011. Currently, he is a Professor with the Institute of Artificial Intelligence and Future Networks, Beijing Normal University and UIC. He has 40 patents and has published more than 200 papers in high-level journals and conferences. He has more than 9500 citations, according to Google Scholar. His H-index is 56. He has managed six national natural science projects (including two sub-projects) and four provincial-level projects. His research interests include the Internet of Things, edge computing, and mobile computing.

Hong-Ning Dai (Senior Member, IEEE) received the Ph.D. degree in computer science and engineering from the Department of Computer Science and Engineering, Chinese University of Hong Kong. He is currently an Associate Professor with the Department of Computing and Decision Sciences, Lingnan University, Hong Kong. His current research interests include big data analytics, blockchain technology, and the Internet of Things. He is a Senior Member of ACM. He has served as an Associate Editor for IEEE TRANSACTIONS ON INDUSTRIAL INFORMATICS, IEEE SYSTEMS JOURNAL, IEEE ACCESS, and *Ad Hoc Networks*.