Towards Better Railway Service: Passengers Counting in Railway Compartment

Yuanzhi Liang,Xueming Qian*,and Li Zhu

***Abstract*—Counting passengers in railway compartments is an essential problem for improving service quality, user experience, public security, and disaster relief in the railway system. Considering many limitations in the compartment, the infrared sensor, 3D camera, etc. are not practical in this scene. Due to the flexibility and lower cost, solutions with standard cameras attract much attention in real applications. However, since the problem with scale variation in the narrow space is different from universal detection or counting problems, the specific benchmark of dataset and methods should be provided and proposed for this task. In this paper, we provide a passenger counting dataset. Relying on this dataset, we propose a passenger counting method. The solution contains a motion supervised multi-scale representation method which provides proposals against scale variation, a spatially-temporally enhanced counting which provides precise counting numbers, and a partial proposal method which conducts methods to be utilized in reality. With the proposed solution, the passengers counting task is solved in higher accuracy and practicable in the compartment environment. In experiments, the results show that all the modules in our solution are useful and efficient, and our method outperforms in comparison with others in the compartment scene.**

***Index Terms*—passenger counting, image processing**

I. INTRODUCTION

For higher efficiency in checking tickets, more flexible allocation to attendants, and more thoughtful and personalized services in compartments, passengers counting becomes a general and essential requirement in the railway industry. With the valuable counting system in the compartment, the railway corporations can also make a better decision on both sides of passengers' experience and profits. Besides, the passengers counting also useful in public security and disaster relief in the railway system.

To count people in a specific scene, we can solve this task via three kinds of devices: infrared sensor, 3D camera (e.g., TOF camera), and the standard camera. Generally, the infrared sensor can fix the most simple condition in counting. However, in the railway compartment, the sensor is not practical due to three defeats: 1. Considering the movement of the compartment itself and multiple people simultaneously, the complicate responses of sensor interweave at the same time, which leads to the sensor useless in real sense for counting. 2. The luggage from passengers always occluded and disturb the sensor and cause inaccurate results. 3. The installation of the sensor should be designed for each type of compartments. It is impossible to develop a solution by sensors to fix every condition of different types of compartments. Besides, the 3D cameras can give the most precise results in some counting problems, but still hard to be utilized in compartments. 1. The 3D camera is expensive. The 3D camera is about 20 times higher in price than the standard cameras. If installing the 3D camera in each compartment, the overall system is too expensive to be practised. 2. The 3D camera is sensitive to the height, which makes the 3D camera can not fit a different type of compartments. Due to the difference in the internal spaces (e.g., height and width of compartments, types of seats), deviations and errors occur when using the original settings of the 3D camera.

Meanwhile, the standard camera has many advantages in this scene. The standard camera is cheap, easy to be installed and flexible to various conditions. Thus, in the real senses of compartments, the solutions with a typical video camera are the most practical.

However, in both applications and researches, gathering data in the railway compartment and using the images captured in the unique scene are not widely concerned. From real applications, counting passengers is an essential task in many railway corporations. From the academic side, the data in this problem is unique, which can not be fixed entirely by face detection like in wider face [6] or COCO [7], or by crowd counting like in Shanghaitech [17] and UCF_CC_50 [16] as shown in Fig.1. The condition of occlusion, the density of the crowd, and the range and distribution of target scales in images are different between the passengers counting and the other problems. Thus, the passenger counting is hard to solve by detection, since the smallest head in images are on a tiny scale and occluded by seats. Since the distribution of targets is not as dense as the open scenes, it is also hard for crowd counting methods. It is valuable to focus on the compartment scene and provide a benchmark in the dataset and method. In this work, we offer a dataset for counting passengers in the compartment and methods for counting.

Specifically, in method design, considering some unique properties in the surveillance videos from the compartments, the ordinary methods from detection or counting are not always available. We summarize the challenges in this task as follow:

This work was supported in part by the NSFC under Grant 61772407, and 61732008).

Yuanzhi Liang is with the Faculty of Software Engineering, Xian Jiaotong University, Xi'an, Shaanxi 710049, China. Xi'an China. (e-mail: liangyzh13@stu.xjtu.edu.cn).

Xueming Qian is with the Ministry of Education Key Laboratory for Intelligent Networks and Network Security, School of Information and Communication Engineering, and SMILES LAB, Xi'an Jiaotong University, Xi'an 710049, China. and also with Zhibian Technology Company, Ltd., Taizhou 317000, China (* Correspond author, e-mail: qianxm@mail.xjtu.edu.cn).

Li Zhu is with the Department of Software Engineering, Xi'an Jiaotong Univeristy, Xi'an, Shaanxi 710049, China. (e-mail: zhuli@mail.xjtu.edu.cn).

Fig. 1. Examples for common detection datasets(Wider face [6], COCO [7]), crowd counting datasets (Shanghaitech datasets [17], UCF_CC_50 [16] datasets) and our passengers counting datasets. The range of head scales in passengers counting dataset is different from others. There are 5 situations in passengers counting datasets. In every situations, the scale changings are apparently larger than other datasets as well.

1. Scale changing: The scale problem in the compartments have two aspects: the extreme range and dynamic variation. 1) The extreme range of scale: from a single frame, images of passengers in the first raw are merely 20 times larger than the smallest ones in the last raw, which is far from the typical variation in crowd counting tasks as shown in the second column of Fig.1. 2) complicated occlusion with scale changing: considering the passengers walk around in this cramped space, some passengers have conspicuous scale changing in adjacent frames, which makes the scale changing dynamically. Meanwhile, the narrow space also leads to severe occlusion. The large scale targets near to the camera often occlude the others, and multiple scale interlace in the same area makes the condition more complicated.

2. Limited hardware requirement: Considering the unique electric system, narrow space for equipment, and the cost of widespread deployment, machines with powerful GPUs and radiator fans are not available. Meanwhile, since some passengers in the last several raws may show a part of heads, the input images should keep the detailed information for passengers, which induce the high-resolution ratio is necessary. (Double cameras for one compartment is also infeasible due to the cost and resource limitation.) Directly downsampling is not useful in this task as well. Thus, the imbalance between high-quality inputs and limited computational resources should be attention. Specifically, in our work, we need to design the methods applied in embedded chips like Hi3519a.

Considering the practicability of the standard camera in railway compartment and many defeats in current methods, we'd like to propose a novel approach to handle this problem, handle the above defeats and obtain better performance in passengers counting. In this paper, we focus on the scene in railway compartments and propose a novel framework with an apposite method for counting passengers. Our framework contains two parts: the motion supervised multi-scale representation method and the spatially-temporally enhanced counting method. The representation method provides proposals for passengers' heads against scale variation and complicated occlusion. Then, the counting method works out the number by the proposals. Moreover, to overcome equipment limitations, we also propose a partial proposal method to adapt the tiny memory in an embedded device. The experiments show the efficiency and high performance of our methods in compartment videos.

The main contributions of this paper are summarized as follows:

1. We propose a motion supervised multi-scale representation method. The method solves the extreme-scale changing by a multi-scale network and introduces inter-frame motion knowledge in videos against disturbance from complicated occlusion. The network with motion knowledge enables to offer a solid proposal for counting method and have better performance in passengers' perception in the compartment.

2. We proposed a novel spatially-temporally enhanced counting method to achieve precise counting numbers from proposals. In this method, we fully consider the dynamic changing of passengers' scales and design a module for learning the spatial and temporal information in videos, which further boost performances in counting.

3. The partial proposal method is designed to provide a solution in limited hardware conditions in railway compartments. With this method, our counting methods can be operated with smaller memory cost, which can adapt to the actual application scenes.

4. Considering the scarcity of images in the real scene in the railway compartment, we propose a passenger counting dataset. Considering the particularities and values of this task in both academic researches and practical applications, our dataset is an essential step for improving counting accuracy and boost the applications in railway compartments.

Our counting framework provides an efficient solution for the passengers counting problem in surveillance videos. In experiments, the proposed methods show better performances than other counting methods and. We would provide our surveillance video dataset and annotations to support further researches in passengers counting like video summarization [56]. The dataset will be available soon.

The remainder of this paper is organized as follows. In section II, we review the related work on counting. Section III demonstrates the details of our method. Experiments of

different methods and analysis of the proposed method are set up in section IV. Finally, the conclusion is in section V.

II. RELATED WORK

Crowd counting is a challenging task due to the complicated background in different scenes and various crowd distribution [18]–[20]. The end-to-end counting gives counting numbers directly by processing single images like in [21]. Density map based counting provides a representation of the crowd distribution first, and then count number through the density map like in [15], [18], [22]–[26]. Compared with the end-to-end method, the density map gives more details and more convenient to apply in multi-task, like crowd velocity estimation and crowd motion analysis. The scale problem also appears in other computer vision tasks like retrieval. There are also some works [55], [57] explore the scale problem and propose some novel methods for representation of multi-scale targets in images.

Moreover, methods depend on object detection models, like Faster RCNN [27], YOLO [28], SSD [29] and MPNET [58], have excellent performance in target detection, which provides new ideas and a great improvement to various counting tasks.

A. Counting based on density maps

Zhang et al. give a cross-scene crowd counting solution via a deep convolutional neural network for the first time [18]. The proposed structure produces a density map for crowd distribution and also contains a fully connected layer for counting number regression. In the crowd counting task, scale problem is one of the most important factors that affect the performance of the algorithm directly. The method in [22] uses a combination with several shallow networks that correspond to multiple scales in the crowd. The thought of applying various networks to capture different scales of crowd develops in many studies. A CNN structure called multiple-column network (MCNN) [15] is proposed for handling crowd images on different scales. Each column of the network applies to a particular scale of the crowd, and all of them are merged by 1*1 convolution to get a density map. Switchable network [23] provides a flexible way to change sub-networks for different scales. It has a switch layer to make the decision for which sub-network can be used for the current image area, and then work out a density map for the crowd. Sindagi et al. [26] showed a Cascaded Multi-task CNN model for a similar but better way. This model can be divided into two stages, which are high-level prior stage and density estimation stage. The prior stage classifies crowd density level and sends prior density information to the estimation stage for boosting performance.

In recent, more interesting crowd counting methods are proposed, Rather than regressing ordinary heatmaps, Sheng et al. [51] propose an attribute map based counting method and provide more information for a network in learning the scene. To better handle the scale variation, the SaCNN [52] provides a novel structure for counting and a geometry-adaptive Gaussian kernel used for better representation of crowd densities. Moreover, infusing more domain knowledge, Huang et al. [53] propose a method to utilize the body part in images, which is more robust for occlusion. Issam et al. [31] give a solution to solve the problem with point annotation. Viresh et al. [32] propose an iterative thought to count with two-branch CNN for both low and high-resolution density maps. To better fuse multi-scale information, the adaptive fusing module is proposed in [13] and achieve better predictions. Rather than using multi-column networks with heavy computational costs, Ze el al. [33] proposed SCNet to handle the task in limited network width. Liu et al. [48] focus on the self-supervised problem in crowd counting and propose a ranking strategy combined with the siamese network to solve the problem. PCC Net [49] is proposed to deal with the problems of high appearance similarity, perspective changes, and severe congestion and obtain better performance. Zheng et al. [50] further explore the problem of diverse densities in the same scene and propose a method to solve cross-line pedestrian counting. Shen et al. [54] focus on the problems of the averaging effects and the cross-scale inconsistency, and proposed the Adversarial Cross-Scale Consistency Pursuit (ACSCP) framework to solve the problems.

B. Detection Based Counting

In low-density crowd estimation, as the targets can be recognized respectively, the counting task can be converted to a detection task. Many detection methods are also practicable for counting.

In tradition solutions of detection [34], potential regions proposals give probable regions for targets first. Then, the regression or classification process is adopted for proposal areas and produces detection output. Moreover, some binary classifiers like Bayes method [35], random forest [36] also broaden the thought of detection and counting tasks. With the increasing research of deep learning, many approaches [27]–[29], [37] seeking the multiple target detection solutions are proposed. The region-based convolutional neural networks (R-CNNs) [38] greatly improve the performance in detection. It uses features extracted from CNN models rather than traditional hand-craft features like Haar [39], etc. Then, in Faster-RCNN [27], region proposal network (RPN) is designed and to reduce calculation in the proposal and also provide faster detection in speed.

Above RCNN related methods can be summarized to the same thought of two stages processing: proposal and classification. This kind of methods have high precision in recognition, but hard to be accelerated. YOLO [28] provides a new solution for detection and efficiently improves the speed. It converts the thought of classification to regression in sub-areas. The input images are divided into several areas; YOLO computes bounding boxes and probabilities in each area. Then, compared with given thresholds, the method decides whether the output contains targets. SSD [29] further improves YOLO and combines the idea of an anchor box. It adds convolutional layers after the baseline network, which makes SSD be able to calculate in multiple scales and has better performance in the small object than YOLO.

Though detection methods have good performance in multiple detections, these methods have a problem in dealing

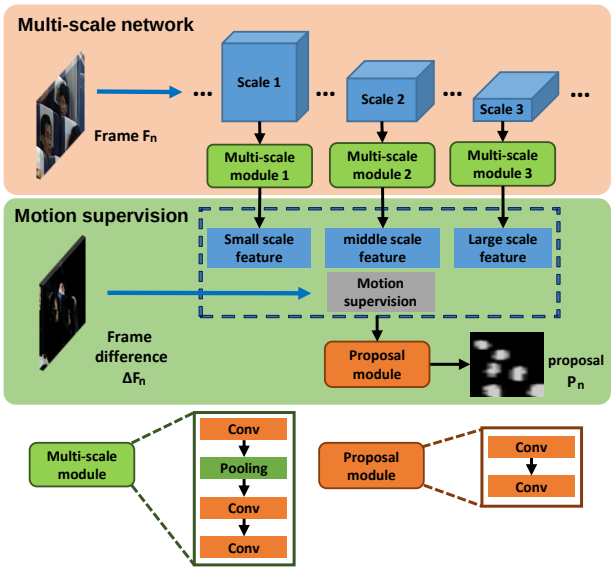


Fig. 2. Overview of our motion supervised the multi-scale representation method. With the design of a multi-scale network and utilization of motion knowledge, the method offers high-quality proposals for counting.

with tiny targets and severe scale changing. Since the primary purpose of these methods is detection, they would reject some imperfect targets, which may make detection based methods inaccuracy for counting.

III. METHODOLOGY

In this work, our solution contains a counting framework and a unique training method. In the counting framework, we proposed a motion supervised multi-scale representation method to offer proposals for passengers' heads. Then, to get accurate counting results, a spatially-temporally enhanced counting method is proposed to deal with proposals from representation methods and gives counting numbers. Holistically, we train the counting framework with our novel training method and solve the passengers counting in actual scenes. More details would be discussed in this section.

A. Motion supervised multi-scale representation

To overcome the severe changing of scale, we proposed the motion supervised multi-scale representation method to get proposals of passengers' heads. In this method, we design a multi-scale network to capture and learning targets in images. We also introduce motion information to lead the network attention on some person related area and conduct better head proposals. More details are given as follows.

1) multi-scale network: We propose a multi-scale network to adapt the scale changing in head sizes. inspired by the works in [15] and [23], three scales in representation are designed in networks, which correspond to the large, middle and small size of head images. As shown in Fig. 2, we first apply Resnet50 [40] as feature extractor and provide basic feature for the n-th frame images F_n. Then, three CNN networks are designed for three scales of features extracted from the backbone network.

For features selection, generally, the shallow layers are available to express detailed texture, which corresponds to local features. The deep layers suit semantic information, which corresponds to global features. The coalition of shallow and deep layers features capture information in diverse scales in images [40]. Thus, we fuse three feature maps on different scales from Resnet50 as shown in Fig.2. We also provide further discussion in experiments.

For the network design, rather than small kernel size used in [40], we applied larger kernel size to capture information from features and given the scale robustness to the networks. In detail, our multi-scale CNN extracts three scales features from Resnet correspondingly, which are 150*150 with 64 channels, 75*75 with 256 channels, and 38*38 with 512 channels. We denote these features as $Scale1$, $Scale2$, and $Scale3$, respectively. Then, the different scale features are feed into three multi-scale modules, as shown in Fig. 2. Concretely, each multi-scale module contains three convolution layers and one pooling layer. The kernel sizes of convolution layers in module 1 are 7*7, 7*7, and 5*5 with 32, 8, and 1 channel. The kernel sizes of convolution layers in module 2 and module 3 are 3*3, 5*5, and 5*5 with 64, 8, and 1 channel. The outputs of three multi-scale modules generate small, middle and large scale features for the further processing in motion supervision part.

2) motion supervised method: In surveillance videos, the frames are not isolated. Considering the movement of passengers, the relationships between frames contain valuable information about passengers' locations and states.

In our method, we fully utilized the information in videos and introduce the motion knowledge into multi-scale network training. As shown in Fig. 2, we take the frame difference $\Delta F_n = F_n - F_{n-1}$ into account. An area with a higher value in ΔF_n stands for a more significant probability of containing passengers. Thus, ΔF_n can be viewed as prior knowledge for passenger location.

In the proposed method, we concatenate frame difference ΔF_n with the features of the multi-scale network. With motion information in ΔF_n, the network enables attention to some area with passengers' movements easily, instead of attention on the background. Meanwhile, we loosen the supervision to prevent overfitting and forbid the network over-relying on some conspicuous motion. The motion supervision $S_n(t)$ can be defined as follows:

$$S_n(t) = \mu(t) * \Delta F_n \tag{1}$$

$$\mu(t) = \begin{cases} 1 - \frac{t}{T} & t < T \\ 0 & t \geq T \end{cases} \tag{2}$$

where t is the training epoch, $\mu(t)$ is a variable changing with training procedure, and T means the decreasing threshold. In this paper, we set the $T = 30$. $\mu(t)$ declines from one to zero linearly during the training procedure and sustains zero in the last several epochs.

In detail, as shown in Fig.2, we concatenate the motion supervision $S_n(t)$ with features from the multi-scale network. Then, we use a proposal module to generate proposal P_n, as

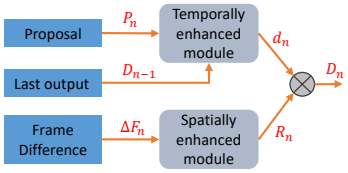


Fig. 3. Pipeline of our spatially-temporally enhanced counting method. The enhanced proposals D_n use the spatial information in ΔF_n and temporal knowledge in D_{n-1} jointly.

shown in Fig.2. The proposal module contains two convolution layers with kernel size in 3*3 and 5*5 and channels in 4 and 1.

B. Spatially-temporally enhanced counting method

To get a precise counting number, we proposed a novel method to process proposals from the representation methods. Our counting method not only utilizes the information in the proposals but also exploits the spatial knowledge in videos and temporal knowledge in previous proposals. Specifically, the whole counting methods can be divided into three parts, the spatially enhanced module, the temporally enhanced module, and the proposal counting module. More details are given in the following parts.

1) Spatially enhanced module: Since the motion information in ΔF_n always provides the instantaneous spatial information and is valuable in counting, we utilize the ΔF_n to enhance the proposals in the spatial side further.

We divide ΔF_n into three scaled sub-areas: 32*32, 16*16, 4*4, which corresponding to the condition of $k = 1, 2, 3$, and then we compute their mean value m_k.

$$m_k(i,j) = \frac{\sum_{\lfloor \frac{i*sub_k}{w} \rfloor}^{\lfloor \frac{i*sub_k}{w} \rfloor + sub_k} \sum_{\lfloor \frac{j*sub_k}{h} \rfloor}^{\lfloor \frac{j*sub_k}{h} \rfloor + sub_k} \Delta F_n(i,j)}{w * h/(sub_k)^2} \tag{3}$$

where i and j are the position value of the horizontal and vertical coordinate. The sub_k indicates sub-area scale in the k-th partitioning pattern (e.g., $sub_1 = 32$). $\lfloor * \rfloor$, w, and h indicate rounding down operation, width and height of ΔF_n. On each scale, we get the corresponding spatial proposal candidates of pattern k as follows:

$$U_n(i,j,k) = \begin{cases} 1 & if \Delta F_n(i.j) \geq m_k(i,j) \\ 0 & else \end{cases} \tag{4}$$

Then we fuse the three scale spatial proposal candidates to get the jointly spatial proposal C_n of frame n as follows:

$$C_n(i,j) = \frac{\sum_{k=1}^{3} U_n(i,j,k)}{3} \tag{5}$$

Next, we get a smoothed response R_n by filtering C_n with a 5*5 Gaussian kernel as follows:

$$R_n(i,j) = Conv(C_n(i,j), G_5) \tag{6}$$

where $Conv$ is convolutional operation.

2) Temporally enhanced module: With the video data, continuous proposal outputs are available on the output side. The proposals are not isolated and also contains useful temporal information to improve proposals. In this part, we fully explore the temporal connection between proposals and enhanced this information to get better proposals.

Based on Hebb learning rule [42], we fuse current raw proposal P_n from motion supervised multi-scale network with the final proposal of the previous frame D_{n-1} to get the temporal enhancement. The Hebb learning rule can be explained to self-adapt network weights: increasing weight when neurons have high response simultaneously and vice versa [42]. As shown in Fig. 3, d_n is obtained by a weighted sum of D_{n-1} and P_n as follows:

$$d_n(i,j) = sigmoid(W_n^1(i,j) * P_n(i,j) + W_n^2(i,j) D_{n-2}(i,j)) \tag{7}$$

where W_n^1 and W_n^2 are the weight matrix for each pixel, which has the same sizes as the proposals.

$$\begin{aligned} W_n^1(i,j) &= (1-\gamma_1) * W_{n-1}^1(i,j) + \alpha_1 * D_{n-1}(i,j) * P_{n-1}(i,j), \\ W_n^2(i,j) &= (1-\gamma_2) * W_{n-1}^2(i,j) + \alpha_2 * D_{n-1}(i,j)^2 \end{aligned} \tag{8}$$

where γ_1, γ_2, α_1 and α_2 are predetermined parameters [42], for example, $\gamma_1 = \gamma_2 = 0.5$, $\alpha_1 = \alpha_2 = 0.5$. In initial setting, W_0^1 is all-one matrix and W_0^2 is all-zero matrix.

At last, we get the spatially-temporally enhanced proposal D_n by the certainty based weighting as follow:

$$D_n(i,j) = R_n(i,j) * d_n(i,j) \tag{9}$$

D_n take both enhancements into account, which is more robust and accurate than original proposals for counting.

3) Proposal counting module: In proposal counting, we first define a series of counting kernels. Each kernel is a two-dimensional matrix with the two factors: $\{width, height\}$. Concretely, the quadrate kernel is $Q_{\{S_q,S_q\}}$ and the rectangular kernel is $R_{\{2S_r,S_r\}}$, where S_q means the width of quadrate kernel and S_r means the height of rectangular kernels. The kernels can be defined as follows:

$$\begin{aligned} Q_{\{S_q,S_q\}}(i,j) &= \frac{1}{\sqrt{10\pi S_q}} e^{-\frac{(i-\frac{S_q}{2})^2 + (j-\frac{S_q}{2})^2}{10 S_q}}, \\ R_{\{2S_r,S_r\}}(i,j) &= \frac{1}{\sqrt{20\pi S_r}} e^{-\frac{(i-S_r)^2 + j^2}{20 S_r}} \end{aligned} \tag{10}$$

where S_q , $S_r \in \{30 + n * k | k = 10, n = 0, ..., 7\}$. We operate convolution to the proposal with generated kernels respectively and get the result by counting the peak of convoluted maps. The pseudo-code description of the method is given as Algorithm 1.

C. Partial proposal training method

We proposed a scene-aware partial input method for the limited computation resources. We keep the inputs in a small size cropping from the raw images corresponding to the specific scene. Details are given as follows.

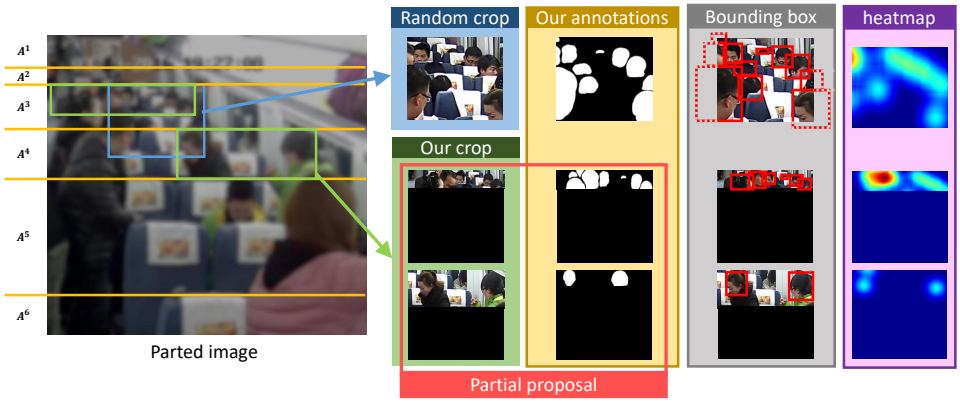


Fig. 4. In our partial proposal method, according to the prior knowledge of the scene, the images are parted into 6 sub-areas. Rather than random crops in the overall images, the partial inputs are cropped in the sub-areas. Then, to better represent the targets, pixel-level annotations are given for network.

Algorithm 1: Proposal counting algorithm

Input: proposal P_n, quadrate kernels Q, rectangular kernels R, proposal width w_p, proposal height h_p
Output: counting numbers Num_n

ξ = 1;
for kernel T_k in $\{Q, R\}$ **do**
  $E_q = Conv(P_n, T_k)$;
  for i = 1 to w_p, j = 1 to h_p **do**
    $\hat{i}, \hat{j} = \underset{i^* \in [i-10, i+10], j^* \in [j-10, j+10]}{\arg\max} E_q(i^*, j^*)$
    $\Theta_\xi = (\hat{i}, \hat{j})$;
    $\xi + +$;
  end
end
$Num_n = 0$;
for μ = 1 to ξ, ν = 1 to ξ **do**
  $\eta = \|\theta_\mu - \theta_\nu\|_2$;
  if $\eta \geq 30$ **then**
    $Num_n + +$;
  end
end

1) Scene-aware partition method: As the specialty of compartment scene, regular crop method always produces broken head images and exacerbates the difficulty in counting. Because of the diversity of passengers' heads images that contain faces, hairs, necks, headwears, etc., the incomplete image data may further increase the difficulty of network learning. Besides, the random crop in the images brings more uncertainty of the scale distribution.

To avoid getting too many incomplete samples, we propose the scene-aware partial input method. As shown in Fig. 4, our method offers more complete head images as useful samples. The details of the method are given as follows.

To adapt the changing of compartments and some deviations in the installation of the camera, we partition the images via the corresponding scene. We calculate the partition according to the seats in the compartment based on images without passengers, e.g., Fig.1. According to [41], the connection of the real world distance and the pixels of the image is shown as follows:

$$y = f_y \frac{Y}{Z} + C_y \tag{11}$$

where y and Y denote the object height in images and the real world, Z is the distance between object and camera, f_y and c_y are the coefficients for mapping the coordinate system from the real world to images. Considering that every row of seats has the same distance to each other, and every seat has the same height Y in the real world coordinate, the position of a ϵ-th row of seat $y(\epsilon)$ can be rewritten as follows:

$$y(k) = \alpha/\epsilon + \beta \tag{12}$$

where α and β are coefficients for mapping. $\epsilon = 1, 2, ..., 10$

These coefficients, as the basis of partition images, would be estimated by following three steps:

(a) We used Hough transform [44] to extract horizontal lines via the edge image obtained by Sobel edge detector [45] in an image without passengers. We choose the lines by constraining the absolute value of slope less than 0.1. K-medoids [46] is adopted to find 10 clustering centers in lines intercepts from Hough transfer. The 10 clustering centers would be used to fitting a $\hat{y}$ and estimate the coefficients $\hat{\alpha}$ and $\hat{\beta}$, which is inversely proportional to ϵ. (b) We take $\hat{y}(\epsilon)$, $\epsilon = 1, 2, ..., 10$ as the position of 10 rows. Then, we classify every two rows into a sub-area and generate five sub-areas. The farthest part of the camera (without seats) would be divided into an independent sub-area. Thus, we partition each image into 6 horizontal sub-areas: $A^1, A^2, ..., A^6$, as shown in Fig. 4. (c) The input data are cropped according to the partitions' edges. As shown in Fig.4, by cropping in sub-areas, we avoid generating too many incomplete head images for training.

2) Proposal annotation: As shown in Fig. 4 and 5, Rather than common counting problem, the intense variation of scale in passenger counting make the heat map hard to represent targets. The bounding box annotation in detection is also impracticable due to the universal phenomenon of incomplete head images and occlusion in images. As shown in Fig.4, the bounding boxes are dense and have too many overlaps in annotations.

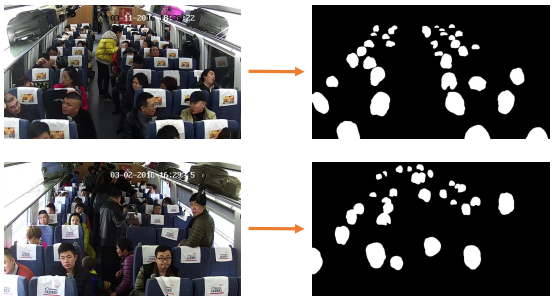


Fig. 5. Examples for image annotations. Examples in the left column are raw images and in the right are corresponding annotations.

In this case, we apply the mask annotation to give 0 or 1 labels for every pixel. Through this measure, even a small part of an incomplete target can be noticed and provided for the network to learn. Furthermore, attributing to the mask annotation, we convert the common regression-based counting thought (regression for the heatmaps, or bounding box position values) to a classification problem (whether the pixel belongs to a head or not). That reduces the difficulty of network learning. The annotations would also be released in our dataset.

IV. EXPERIMENT

In this section, we evaluate our method for passengers counting on the compartment surveillance video. We compare the proposed passenger counting approach with the detection based approaches, such as Faster RCNN [27], YOLO [28], and SSD [29]. We also test the common crowd counting method like CSRNet [10], SFCN [9], and SANet [8]. To train the detectors with passenger data, we use ground truth proposals to generate a bounding box. Utilizing our counting method, we get the peak points as the center of bounding boxes and take the kernel size as box size. Then, we train the data together with VOC2007-2012 [47] to avoid overfitting. For counting methods, the heatmaps are annotated according to methods in [15], [18]. In the setting of our experiment, with the partial proposal method, the inputs to our network are 300*300. The network train with SGD optimizer in the learning rate of 0.01 and the decay step is 0.1 for every ten epochs. We finally train 75 epochs in our experiments.

A. Dataset

TABLE I
RATIO OF DIFFERENT SITUATION IN OUR COMPARTMENT COUNTING DATASET

Situations	DJ	NP	DB	BE	DE
Ratio	0.616	0.21	0.062	0.074	0.038

The data we have contained 500 video sequences, each video is 8-10 minutes, and the frame size is 1280*720*3. The video data can be repartitioned into five situations, which are: during the journey (DJ), during boarding (DB), during exiting (DE), boarding and exiting simultaneously (BE), and no passengers (NP). The examples of situations are shown as Fig. 1. The ratio of every situation in compartment video data shown as TABLE I. Our partition settings and crop method are also available soon in the open-source version of the dataset.

In the training stage, we choose 120 videos as the training set. We randomly capture two adjacent frames from each video to avoid repetition (one as the training input, the other for calculating ΔF_n). We annotate the 120 frames with the pixel-wise annotation, as mentioned in the last section, and shown in Fig. 4 and 5. In the test stage, we take 1000 frames from the other 380 videos. All training and test images are annotated with counting numbers.

B. Metric

We used the mean squared error (MSE) and the absolute error (MAE) as in [15], [18], [23], which are defined as follows:

$$\begin{aligned} MSE &= \sqrt{\frac{1}{N}\sum_{i}^{N}(x_i - \hat{x_i})^2}, \\ MAE &= \frac{1}{N}\sum_{i}^{N}\|x_i - \hat{x_i}\| \end{aligned} \tag{13}$$

where N is the number of test frames, x_i and $\hat{x_i}$ are the ground truth and values of predictions. Generally, MAE and MSE manifest the accuracy and stability of estimation correspondingly.

C. Objective Performances

Comparisons of different methods under different situations are shown in TABLE II and TABLE III, respectively. All detection based methods [27]–[29] have low MSE and MAE values in NP since the numbers in this situation are less than 5, and occlusion is not severe. The detectors can distinguish people in the scene. But the actions of people may cause deformation and occlusion, which may disturb the detector. On the contrary, the counting based methods [8]–[10] can handle some crowded scenes. But when the passengers have complicated interaction in BE, performances of these methods become worse. In NP situation, the counting results are also affected by the background of seats or adversarial. In the experiments, our method is more robust to these disturbances and obtains better performances.

Meanwhile, we find that the DJ situation is the hardest but is the common case of the dataset. Due to the severe occlusion for passengers behind the seats, detection-based methods are hard to find convincible targets. In YOLO, because the YOLO is not good at small object detection, in this case, the method performs worse than in other situations. In comparison with other methods, our method still achieves better performance in the accuracy (MSE) and stability (MAE).

Besides, the improvements of our method are not all from DJ situations. Our method also outperforms in other situations in all the metrics. In DJ, DE, and BE, the large-scale movements of passengers induce more severe occlusion and scale changing, which are hard for other methods. In our

TABLE II
MSE FOR VARIOUS SITUATIONS IN DIFFERENT COUNTING METHODS.

Situations	MSE						
	Detection based methods			Counting based methods			Ours
	Faster-RCNN [37]	YOLO [28]	SSD [29]	CSRNet [10]	SFCN [9]	SANet [8]	
DJ	15.16	18.78	9.24	12.62	25.55	13.84	3.71
DB	5.65	7.13	5.72	5.89	9.35	8.02	3.42
DE	7.68	13.33	4.60	4.97	13.32	10.79	3.03
BE	15.79	16.56	11.18	20.75	35.51	15.83	4.40
NP	1.73	0.89	0.93	3.69	4.37	20.86	0.73

TABLE III
MAE FOR VARIOUS SITUATIONS IN DIFFERENT COUNTING METHODS.

Situations	MAE						
	Detection based methods			Counting based methods			Ours
	Faster-RCNN [37]	YOLO [28]	SSD [29]	CSRNet [10]	SFCN [9]	SANet [8]	
DJ	14.85	18.60	8.65	11.32	24.88	13.18	2.90
DB	5.18	5.86	5.18	3.84	8.90	6.90	2.74
DE	6.77	11.56	3.91	4.32	12.02	9.15	2.49
BE	15.54	14.14	10.56	20.38	35.45	14.90	3.68
NP	1.48	0.52	0.48	3.56	4.25	20.81	0.42

method, the motion information and historical information are all considered, which lead to better representations in proposals. The results show that our method are more efficient in situations with massive movements.

In the view of generality, the MSE values in our method are about 4.0, and the MAE values are about 3.0. While, the other methods show a large gap between all the situations, and have higher volatile. In the application of reality, the values of metrics should be stable, and methods should be useful in all the possibilities. Thus, our method has a significant advantage in practical application, as well.

D. Subjective Performances

The results of Faster-RCNN, YOLO, SSD, and ours for the DJ, DB, DE, and BE situations are shown in Fig. 6. To visualize our proposal for comparison, we take the peaks position in counting module as the centers of boxes and the kernel scales as the box sizes.

In experiments, YOLO is not good at the small object, and the method tends to predict a large bounding box output, as shown in the second column of Fig. 6. Sometimes, even the counting number is correct, the passengers' locations are wrong, and the bounding boxes contain multiple people at the same time.

Moreover, due to the region proposal module, Faster-RCNN is sensitive to potential head images, which causes the method more likely to give false detection, as shown in the second column of Fig. 6. Some areas may similar to the occluded head and recognized by Faster-RCNN as targets. That makes the counting numbers via Faster-RCNN larger, which also causes errors in the NP situation.

SSD also has more detection results and counting value in the same frame, but SSD is better than Faster-RCNN for passengers counting in most situations. In some cases which contain severe occlusion disturbance and scale problems like in BE situation in the fourth raw of Fig. 6, the SSD method also has many duplicate results and does not locate the correct head position.

Our method is better in counting passengers for the compartment images. The proposed method can count heads in the various scale against scale variation and also shows better performance than other methods.

E. Discussions in Motion Supervised Multi-scale Network

1) Multi-scale network comparison: We fuse the features from $Scale1$, $Scale2$, and $Scale3$ for generating a proposal. In the combinations of features in different scales, the proposals generated from low-level layers, such as $Scale1$, $Scale2$, and their combinations are not accurate for counting in large scale heads. Since $Scale3$ contains a more semantic feature which is useful to identify human heads. The proposals generated from three scale levels features are best among all the combinations. Moreover, as shown in Fig. 7, with the addition of different scales, the proposal tends to be refined gradually. The proposals generated by $Scale1$ is raw and large. With the $Scale2$ added, the proposal becomes smaller and finer. Finally, with all scales utilized, the proposal is accurate and better to present targets.

To choose the best combination of feature maps, we test the counting performance in the situation of DJ with different scale features. As shown in Fig. 8, we compare the performances of the average MSE and MAE values of our method

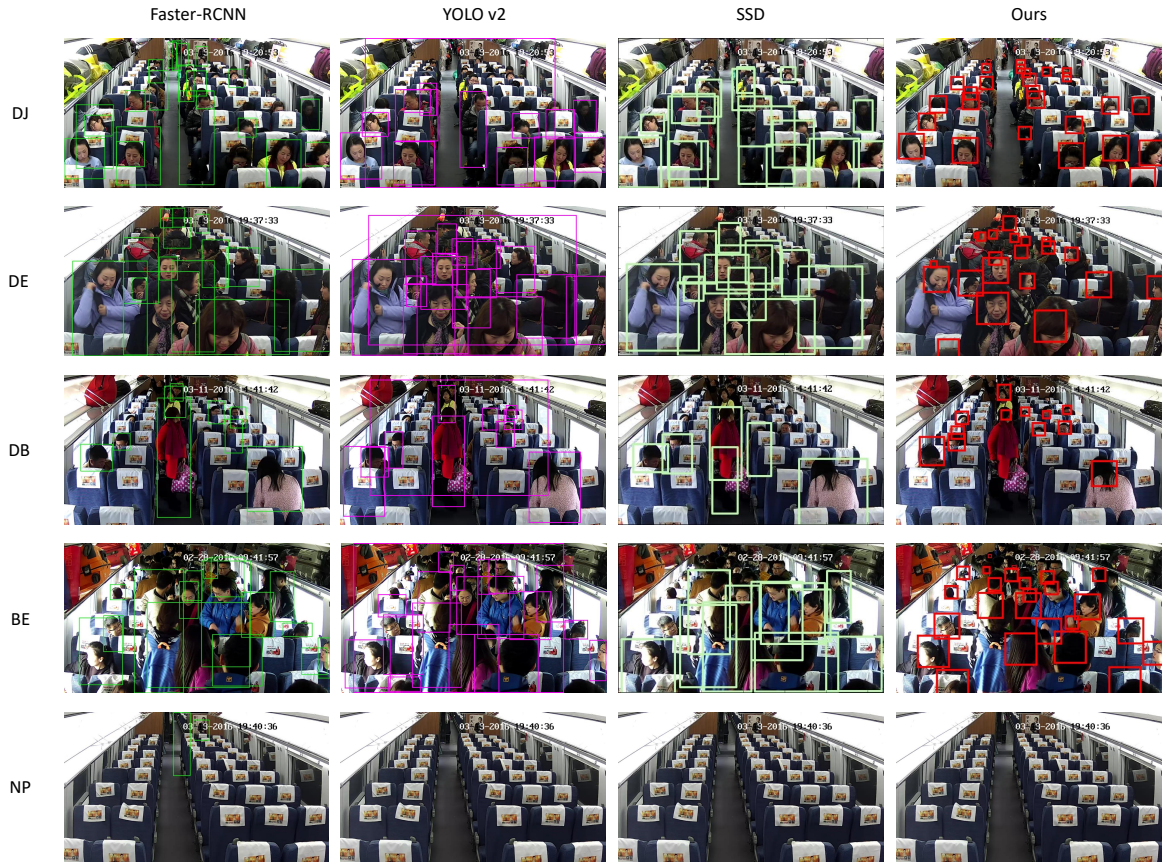


Fig. 6. Results of detection based methods and our methods. Our method is more robust for scale changing in every situation and more accurate in counting tasks.

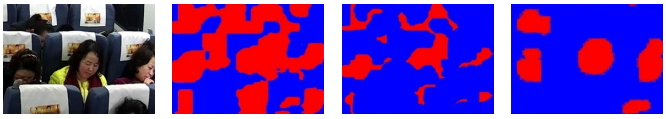


Fig. 7. The proposal generated from different features. (a) is the original frame. (b) is the proposal from Scale1 features. (c) is the proposal from $Scale1$ and $Scale2$ features. (d) is a proposal from features of all the scales. Only using features in all scales can generate the best proposal for counting.

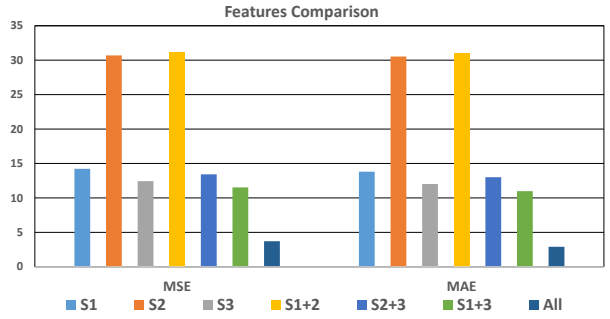
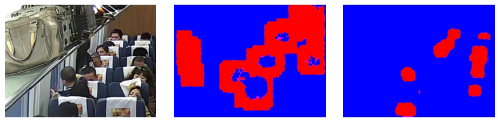


Fig. 8. The different scale features comparison. S1, S2, and S3 mean using only use Scale1 feature, Scale2 feature and Scale3 feature separately for proposal.

using various features extracted from CNN. In Fig., $S1$, $S2$, and $S3$ means only using $Scale1$ feature, $Scale2$ feature and $Scale3$ feature separately for proposal generation. $S1+2$, $S2+3$, $S1+3$, and All means using the combination of $Scale1$ and $Scale2$ features, $Scale2$ and $Scale3$ features, $Scale1$ and $Scale3$ features, and all the three scale features respectively.

$Scale1$ and $Scale2$ features contain some fine-textured information of head images, which are useful in small head detection. $Scale3$ has more semantic information, which is essential in detecting head and generating a proposal. In our experiment, the $Scale2$ features are useful when combining with semantic information in $Scale3$, but it's hard to produce a correct proposal only with $Scale2$ features individually or the combination of $Scale1$ and $Scale2$. When we use $Scale1$ features for proposal, with the detailed texture information, the proposals tend to be coarse and large, which could be captured by kernels in the counting method and produce counting results primarily. While, when with $Scale2$ features, the proposals tend to be finer and smaller. As shown in Fig. 7(c), the finer proposal may bring more noise for the counting method, but better in representing targets. On the other side, in the experiment, $Scale2$ features offer some sort of compliment to $Scale1$. The over-fitting always appears when we try to generate a proposal only with $S2$ or $S1+S2$ features. Moreover, when combining all the features of three scales, the proposals become stable and reliable. All these facts indicate that texture information in $Scale1$ and $Scale2$ and semantic information in $Scale3$ are all necessary for counting.



(a) (b) (c)

Fig. 9. Comparison with or without motion supervision. (a) is the raw image. (b) and (c) are from the method training without and with motion supervision.

2) Motion supervision: As shown in Fig. 9, considering the occlusion and scale changing in images, proposals with motion supervised focus on the passengers' head better. On the contrary, without motion information, the network tends to misrecognize the objects as passengers and give much larger proposals.

Numerically, in MSE and MAE metric, the inaccurate proposals bring the values to more than 20 in the DJ and the BE situations, which is unusable in passengers' counting. In overall, as shown in Fig.7 and Fig.9, our methods can handle the complicated occlusion and scale changing in the passengers counting task. The proposals generated from representation methods efficiently capture the targets.

F. Discussion in Spatially-temporally Enhanced Counting Method

To verify the efficiency of the proposed module, experiments of utilizing spatial and temporal enhancement are operated and discussed in this part.

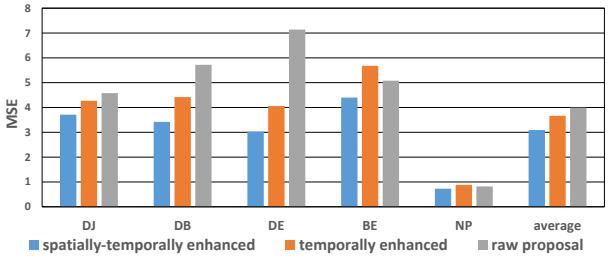


Fig. 10. MSE comparison of the passenger counting with spatial and temporal enhancements. MSE can reflect the stability of the system. The enhancements effectively keep the proposals more stable and reliable for counting.

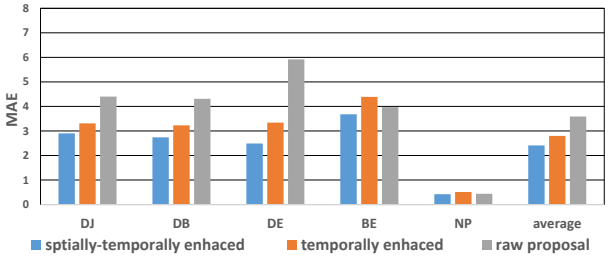


Fig. 11. MAE comparison of the passenger counting with spatial and temporal enhancement. MAE measures absolute accuracy in counting. Though proposed enhanced methods, the proposals are better to represent targets and boost performances in counting.

As shown in Fig. 10 and 11, our motion supervised multi-scale network is already able to offer counting proposal. With the enhancement, the proposal can be better in all situations. Since the more intense movement of passengers brings more severe scale changing, in the situations of DB, DE, and BE, the improvements become more apparent. If there is excessive movement in a single frame, the faulty proposals may connect to or adhere to others and form a broader range of mistakes. That makes the counting number plunge and produces larger MSE and MAE.

Further, the performances are improved a lot when utilizing temporal information. The average MAE and MSE values decrease by about 22.03% and 8.30%, respectively, compared to the original proposal. Moreover, with the spatial information, the MAE and MSE values decrease by about 32.81% and 22.63%, respectively.

Moreover, in BE situation, the single temporal enhancement may result in worse performance, as shown in experiments. Because the BE situation has the most intense activity of passenger movement, the proposal from the last frame is different from current proposal. Addition of the historical knowledge would disturb the prediction of a new proposal in this case. However, benefiting from the spatial enhancement, the instantaneous information of passengers' states are given. The influence of dramatic movements can be modified and reduced by spatial prior, which leads to a better performance in final proposals.

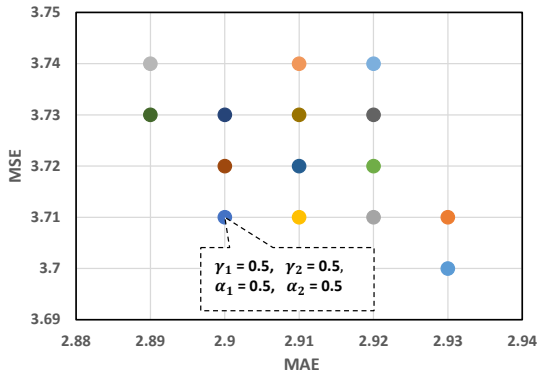


Fig. 12. The results of grid search for comparing hyper-parameters γ_1, γ_2, α_1 and α_2. All parameters are tested from 0.1 to 0.9 with the step size 0.2, and the overall number of settings in experiments is 625. The searching is operated in DJ condition, and we rank all the settings with the sum of MSE and MAE value. In this figure, we draw the lowest 20 settings. Some settings have the same results in MAE and MSE(e.g. $\gamma_1 = 0.7$, $\gamma_2 = 0.5$, $\alpha_1 = 0.7$, $\alpha_2 = 0.7$ and $\gamma_1 = 0.7$, $\gamma_2 = 0.5$, $\alpha_1 = 0.7$, $\alpha_2 = 0.9$ have the same results of MSE=3.74, MAE=2.92), so the number of the points in figure is less than 20.

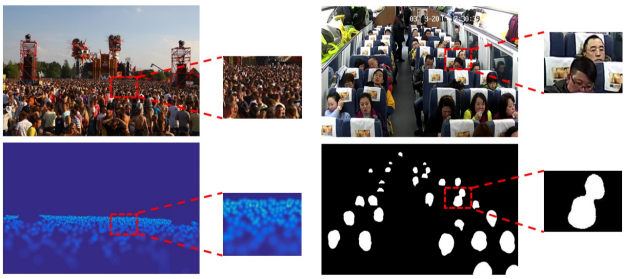
We also compare the hyper-parameters γ_1, γ_2, α_1 and α_2 in this part, as shown in Fig. 12. We find that the sum of MSE and MAE of the best 20 settings are all in range of 6.62 to 6.66. After the comprehensive consideration of MSE and MAE, We choose the setting of $\gamma_1 = 0.5$, $\gamma_2 = 0.5$, $\alpha_1 = 0.5$, $\alpha_2 = 0.5$.

G. Discussion in Crowd Counting Datasets

Our main contribution is solving the counting problem for passengers in compartments. The scene is specific and has many differences to the common crowd dataset, which makes the counting of passengers are hard to be handled by other

methods. The density, occlusion condition, and the ranges of scales are different as shown in Figure. 1.

In detail, there are three differences in data.



(a) (b)

Fig. 13. We give some examples for typical crowd counting dataset (a) and the passengers counting data (b). Though both datasets have problems in the multiple scales, occlusion, etc., the problems are different in datasets.

(a) (b)

Fig. 14. We provide from examples from Shanghaitech[1] dataset. Since the factors like position, angle, resolution, etc. are different in each image in typical crowd counting dataset, the variation and range of head scales in images are various. However, the compartment videos are taken from the same cameras with the same settings, which indicates the same range of head scale, as shown in Figure. 1 and Additional Figure. 2.

(1). The conditions of scales are different. As shown in Fig. 13, we take the same size of the area from the normalized images from Shanghaitech[3] (a) and passengers counting dataset (b). In Fig 13(a), the overall value is 6256 and the value in parted box is 216. In Fig 13(b), the overall value is 30 and the value in parted box is only 2. Compared with the images of compartments, the density in common crowd datasets is significantly larger.

(2). The condition of occlusion is different. As shown in Fig. 13, though occlusion is severe in (b), the edge of each image of the head is more explicit. However, the density of common crowd counting is enormous. It is hard to locate all the people in images.

(3). The ranges of scales are different. Due to the same angle of cameras with the indoor condition in passengers counting dataset, as shown in Fig. 1 and 13, all the images in dataset contain the same range and the similar distribution of scales. However, in common dataset, every images have different scale condition as shown in Fig. 14. Our contributions of Spatially-temporally enhanced counting method and Partial proposal training method are hard to be reflected.

Moreover, in our proposed pipeline, we take advantage of the motion information in the video which are lack in typical crowd counting datasets. Thus, the contributions in motion supervision and spatially-temporally enhanced counting method can not be reflected. Considering the aforementioned differences, especially in the range of scales, the partial proposal training method is also infeasible as well. So, we only compare the multi-scale network in crowd counting dataset as follow:

TABLE IV
THE RESULTS OF OUR MULTI-SCALE NETWORK IN CROWD COUNTING DATASETS. THE SH A, SH B AND UCF INDICATE THE SHANGHAITECH A, SHANGHAITECH B AND UCF_CC_50 DATASETS.

Methods	SH A [17]		SH B [17]		UCF [16]	
Metrics	MAE	MSE	MAE	MSE	MAE	MSE
MCNN [15]	181.8	277.7	32.0	49.8	377.6	509.1
CSRNet [10]	68.2	115.0	10.6	16.0	266.1	397.5
SFCN [9]	64.8	107.5	7.6	13.0	214.2	318.2
SANet [8]	67.0	115.0	10.6	16.0	258.4	334.9
Our network	132.6	193.3	17.8	21.7	541.5	703.4

Among the crowd datasets, our multi-scale network works better in Shanghaitech B [17], because the Shanghaitech B [17] has the relatively stable position of the camera and the ranges of scale in images are similar. On the contrary, the images in Shanghaitech A [17] and UCF_CC_50 [16] have different angles and positions of the camera, which lead to the range of head images are various in each image. These problems deviate from the original aims of our design. Moreover, the UCF_CC_50 [16] dataset only contains 50 images, and we use 40 images for training. The limited data is not beneficial to network training as well. The problems of crowd counting datasets are far from our passengers counting. The passenger counting is a unique problem and contains a high value in both sides of researches and applications. In this work, we focus on proposing a novel pipeline to solve the problem and a dataset to boost the researches in passengers counting.

H. Discussion in Partial Proposal Training

The proposed training method leads the tiny multi-scale network to operate in the embedded devices. In the specific value of the method, the overall network without the backbone is 6.6M, and enable to train in 680M of memory with batch size 24. The total training epoch is 75, and the overall procedure on board is about 10 hours. With the proposed method, all the procedures can be operated within a limited condition of hardware devices. All this fact makes our methods practicable and feasible in limited equipment in the compartment environment.

V. CONCLUSION

In this paper, we focus on the passengers counting problems in the railway compartment and offer a benchmark in both datasets and methods. We present a complete pipeline for giving head proposals, counting numbers, and operate the system in the limited hardware condition for passenger counting problems. The proposed motion supervised representation provides reliable and robust proposals against scale changing and occlusion. The spatial and temporally enhanced counting offers accurate counting result from proposals. With the given partial proposal training method, the methods can operate and solve the problem in the actual scene of railway compartments. The advantages of our approach are particularly significant in

this scene, and our method achieves better performance than the other methods in all situations in datasets.

REFERENCES

[1] P. Korshunov and W. T. Ooi, "Video quality for face detection, recognition, and tracking," ACM Trans. Multimedia Comput. Commun. Appl., vol. 7, no. 3, pp. 14:1–14:21, Sep. 2011. [Online]. Available: http://doi.acm.org/10.1145/2000486.2000488

[2] X. Yang, T. Zhang, and C. Xu, "Semantic feature mining for video event understanding," ACM Trans. Multimedia Comput. Commun. Appl., vol. 12, no. 4, pp. 55:1–55:22, Aug. 2016. [Online]. Available: http://doi.acm.org/10.1145/2962719

[3] J. M. Grant and P. J. Flynn, "Crowd scene understanding from video: A survey," ACM Trans. Multimedia Comput. Commun. Appl., vol. 13, no. 2, pp. 19:1–19:23, Mar. 2017. [Online]. Available: http://doi.acm.org/10.1145/3052930

[4] X. Li, M. Chen, and Q. Wang, "Measuring collectiveness via refined topological similarity," ACM Trans. Multimedia Comput. Commun. Appl., vol. 12, no. 2, pp. 34:1–34:22, Mar. 2016. [Online]. Available: http://doi.acm.org/10.1145/2854000

[5] H.-B. Zhang, B. Zhong, Q. Lei, J.-X. Du, J. Peng, D. Chen, and X. Ke, "Sparse representation-based semi-supervised regression for people counting," ACM Trans. Multimedia Comput. Commun. Appl., vol. 13, no. 4, pp. 47:1–47:17, Aug. 2017. [Online]. Available: http://doi.acm.org/10.1145/3106156

[6] S. Yang, P. Luo, C. C. Loy, and X. Tang, "Wider face: A face detection benchmark," in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[7] T. Lin, M. Maire, S. J. Belongie, L. D. Bourdev, R. B. Girshick, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: common objects in context," CoRR, vol. abs/1405.0312, 2014. [Online]. Available: http://arxiv.org/abs/1405.0312

[8] 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 (ECCV), 2018, pp. 734–750.

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

[10] 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.

[11] X. Hu, J. Deng, J. Zhao, W. Hu, E. C.-H. Ngai, R. Wang, J. Shen, M. Liang, X. Li, V. C. M. Leung, and Y.-K. Kwok, "Safedj: A crowd-cloud codesign approach to situation-aware music delivery for drivers," ACM Trans. Multimedia Comput. Commun. Appl., vol. 12, no. 1s, pp. 21:1–21:24, Oct. 2015. [Online]. Available: http://doi.acm.org/10.1145/2808201

[12] D. Kang, Z. Ma, and A. B. Chan, "Beyond counting: Comparisons of density maps for crowd analysis tasks - counting, detection, and tracking," CoRR, vol. abs/1705.10118, 2017. [Online]. Available: http://arxiv.org/abs/1705.10118

[13] D. Kang and A. B. Chan, "Crowd counting by adaptively fusing predictions from an image pyramid," in British Machine Vision Conference 2018, BMVC 2018, Northumbria University, Newcastle, UK, September 3-6, 2018, 2018, p. 89. [Online]. Available: http://bmvc2018.org/contents/papers/0283.pdf

[14] Z. Zou, X. Su, X. Qu, and P. Zhou, "Da-net: Learning the fine-grained density distribution with deformation aggregation network," IEEE Access, vol. 6, pp. 60 745–60 756, 2018. [Online]. Available: https://doi.org/10.1109/ACCESS.2018.2875495

[15] Y. Zhang, D. Zhou, S. Chen, S. Gao, and Y. Ma, "Single-image crowd counting via multi-column convolutional neural network," in 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016, 2016, pp. 589–597. [Online]. Available: https://doi.org/10.1109/CVPR.2016.70

[16] H. Idrees, I. Saleemi, C. Seibert, and M. Shah, "Multi-source multi-scale counting in extremely dense crowd images," in 2013 IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, June 23-28, 2013, 2013, pp. 2547–2554. [Online]. Available: https://doi.org/10.1109/CVPR.2013.329

[17] Y. Zhang, D. Zhou, S. Chen, S. Gao, and Y. Ma, "Single-image crowd counting via multi-column convolutional neural network," in 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016, 2016, pp. 589–597. [Online]. Available: https://doi.org/10.1109/CVPR.2016.70

[18] C. Zhang, H. Li, X. Wang, and X. Yang, "Cross-scene crowd counting via deep convolutional neural networks," in IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015, 2015, pp. 833–841. [Online]. Available: https://doi.org/10.1109/CVPR.2015.7298684

[19] F. Dai, H. Liu, Y. Ma, J. Cao, Q. Zhao, and Y. Zhang, "Dense scale network for crowd counting," CoRR, vol. abs/1906.09707, 2019. [Online]. Available: http://arxiv.org/abs/1906.09707

[20] D. B. Sam, S. V. Peri, M. N. S., A. Kamath, and R. V. Babu, "Locate, size and count: Accurately resolving people in dense crowds via detection," CoRR, vol. abs/1906.07538, 2019. [Online]. Available: http://arxiv.org/abs/1906.07538

[21] C. Shang, H. Ai, and B. Bai, "End-to-end crowd counting via joint learning local and global count," in 2016 IEEE International Conference on Image Processing, ICIP 2016, Phoenix, AZ, USA, September 25-28, 2016, 2016, pp. 1215–1219. [Online]. Available: https://doi.org/10.1109/ICIP.2016.7532551

[22] L. Boominathan, S. S. S. Kruthiventi, and R. V. Babu, "Crowdnet: A deep convolutional network for dense crowd counting," in Proceedings of the 2016 ACM Conference on Multimedia Conference, MM 2016, Amsterdam, The Netherlands, October 15-19, 2016, 2016, pp. 640–644. [Online]. Available: https://doi.org/10.1145/2964284.2967300

[23] D. B. Sam, S. Surya, and R. V. Babu, "Switching convolutional neural network for crowd counting," in 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017, 2017, pp. 4031–4039. [Online]. Available: http://doi.ieeecomputersociety.org/10.1109/CVPR.2017.429

[24] S. Kumagai, K. Hotta, and T. Kurita, "Mixture of counting cnns: Adaptive integration of cnns specialized to specific appearance for crowd counting," CoRR, vol. abs/1703.09393, 2017. [Online]. Available: http://arxiv.org/abs/1703.09393

[25] M. Marsden, K. McGuinness, S. Little, and N. E. O'Connor, "Fully convolutional crowd counting on highly congested scenes," in Proceedings of the 12th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2017) - Volume 5: VISAPP, Porto, Portugal, February 27 - March 1, 2017., 2017, pp. 27–33. [Online]. Available: https://doi.org/10.5220/0006097300270033

[26] V. A. Sindagi and V. M. Patel, "Cnn-based cascaded multi-task learning of high-level prior and density estimation for crowd counting," in 14th IEEE International Conference on Advanced Video and Signal Based Surveillance, AVSS 2017, Lecce, Italy, August 29 - September 1, 2017, 2017, pp. 1–6. [Online]. Available: https://doi.org/10.1109/AVSS.2017.8078491

[27] S. Ren, K. He, R. B. Girshick, and J. Sun, "Faster R-CNN: towards real-time object detection with region proposal networks," in Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada, 2015, pp. 91–99. [Online]. Available: http://papers.nips.cc/paper/5638-faster-r-cnn-towards-real-time-object-detection-with-region-proposal-networks

[28] J. Redmon, S. K. Divvala, R. B. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016, 2016, pp. 779–788. [Online]. Available: https://doi.org/10.1109/CVPR.2016.91

[29] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. E. Reed, C. Fu, and A. C. Berg, "SSD: single shot multibox detector," in Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part I, 2016, pp. 21–37. [Online]. Available: https://doi.org/10.1007/978-3-319-46448-0_2

[30] B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds., Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VII, ser. Lecture Notes in Computer Science, vol. 9911. Springer, 2016. [Online]. Available: https://doi.org/10.1007/978-3-319-46478-7

[31] I. H. Laradji, N. Rostamzadeh, P. O. Pinheiro, D. Vázquez, and M. W. Schmidt, "Where are the blobs: Counting by localization with point supervision," in Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part II, 2018, pp. 560–576. [Online]. Available: https://doi.org/10.1007/978-3-030-01216-8_34

[32] V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds., Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part VII, ser. Lecture Notes in Computer Science, vol. 11211. Springer, 2018. [Online]. Available: https://doi.org/10.1007/978-3-030-01234-2

[33] Z. Wang, Z. Xiao, K. Xie, Q. Qiu, X. Zhen, and X. Cao, "In defense of single-column networks for crowd counting," CoRR, vol. abs/1808.06133, 2018. [Online]. Available: http://arxiv.org/abs/1808.06133

[34] J. R. R. Uijlings, K. E. A. van de Sande, T. Gevers, and A. W. M. Smeulders, "Selective search for object recognition," International Journal of Computer Vision, vol. 104, no. 2, pp. 154–171, 2013. [Online]. Available: https://doi.org/10.1007/s11263-013-0620-5

[35] A. B. Chan and N. Vasconcelos, "Bayesian poisson regression for crowd counting," in IEEE 12th International Conference on Computer Vision, ICCV 2009, Kyoto, Japan, September 27 - October 4, 2009, 2009, pp. 545–551. [Online]. Available: https://doi.org/10.1109/ICCV.2009.5459191

[36] V. Pham, T. Kozakaya, O. Yamaguchi, and R. Okada, "COUNT forest: Co-voting uncertain number of targets using random forest for crowd density estimation," in 2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015, 2015, pp. 3253–3261. [Online]. Available: https://doi.org/10.1109/ICCV.2015.372

[37] R. B. Girshick, "Fast R-CNN," in 2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015, 2015, pp. 1440–1448. [Online]. Available: https://doi.org/10.1109/ICCV.2015.169

[38] R. B. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," CoRR, vol. abs/1311.2524, 2013. [Online]. Available: http://arxiv.org/abs/1311.2524

[39] A. Haar, "Der massbegriff in der theorie der kontinuierlichen gruppen," The Annals of Mathematics, vol. 34, no. 1, p. 147, Jan. 1933. [Online]. Available: https://doi.org/10.2307/1968346

[40] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778, 2016.

[41] J. E. Gentle, Matrix Algebra: Theory, Computations, and Applications in Statistics, 1st ed. Springer Publishing Company, Incorporated, 2007.

[42] C. Sammut and G. I. Webb, Eds., Encyclopedia of Machine Learning and Data Mining. Springer, 2017. [Online]. Available: https://doi.org/10.1007/978-1-4899-7687-1

[43] G. Bianchi and I. Tinnirello, "Kalman filter estimation of the number of competing terminals in an IEEE 802.11 network," in Proceedings IEEE INFOCOM 2003, The 22nd Annual Joint Conference of the IEEE Computer and Communications Societies, San Franciso, CA, USA, March 30 - April 3, 2003, 2003, pp. 844–852. [Online]. Available: https://doi.org/10.1109/INFCOM.2003.1208922

[44] P. Mukhopadhyay and B. B. Chaudhuri, "A survey of hough transform," Pattern Recognition, vol. 48, no. 3, pp. 993–1010, 2015. [Online]. Available: https://doi.org/10.1016/j.patcog.2014.08.027

[45] N. Kanopoulos, N. Vasanthavada, and R. L. Baker, "Design of an image edge detection filter using the sobel operator," IEEE Journal of solid-state circuits, vol. 23, no. 2, pp. 358–367, 1988.

[46] X. Jin and J. Han, K-Medoids Clustering. Boston, MA: Springer US, 2010, pp. 564–565. [Online]. Available: https://doi.org/10.1007/978-0-387-30164-8426

[47] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," International Journal of Computer Vision, vol. 88, no. 2, pp. 303–338, Jun 2010. [Online]. Available: https://doi.org/10.1007/s11263-009-0275-4

[48] 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, pp. 1862–1878, 2019.

[49] J. Gao, Q. Wang, and X. Li, "Pcc net: Perspective crowd counting via spatial convolutional network," IEEE Transactions on Circuits and Systems for Video Technology, 2019.

[50] H. Zheng, Z. Lin, J. Cen, Z. Wu, and Y. Zhao, "Cross-line pedestrian counting based on spatially-consistent two-stage local crowd density estimation and accumulation," IEEE Transactions on Circuits and Systems for Video Technology, vol. 29, pp. 787–799, 2018.

[51] B. Sheng, C. Shen, G. Lin, J. Li, W. Yang, and C. Sun, "Crowd counting via weighted vlad on a dense attribute feature map," IEEE Transactions on Circuits and Systems for Video Technology, vol. 28, pp. 1788–1797, 2018.

[52] L. Zhang, M. Shi, and Q. Chen, "Crowd counting via scale-adaptive convolutional neural network," 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 1113–1121, 2018.

[53] S. Huang, X. Li, Z. Zhang, F. Wu, S. Gao, R. Ji, and J. Han, "Body structure aware deep crowd counting," IEEE Transactions on Image Processing, vol. 27, pp. 1049–1059, 2018.

[54] Z. Shen, Y. Xu, B. Ni, M. Wang, J. Hu, and X. Yang, "Crowd counting via adversarial cross-scale consistency pursuit," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 5245–5254.

[55] Z. Ye and Y. Peng, "Sequential cross-modal hashing learning via multi-scale correlation mining," ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), vol. 15, no. 4, pp. 1–20, 2019.

[56] Y. Peng and C.-W. Ngo, "Clip-based similarity measure for query-dependent clip retrieval and video summarization," IEEE Transactions on Circuits and Systems for Video Technology, vol. 16, no. 5, pp. 612–627, 2006.

[57] Z. Ye and Y. Peng, "Multi-scale correlation for sequential cross-modal hashing learning," in Proceedings of the 26th ACM international conference on Multimedia, 2018, pp. 852–860.

[58] H. Wang, P. Wang, and X. Qian, "Mpnet: An end-to-end deep neural network for object detection in surveillance video," IEEE Access, vol. 6, pp. 30 296–30 308, 2018.

Yuanzhi Liang received the B.S. degree from Lanzhou University, Lanzhou, China, in 2017. He is currently working towards the Master degree at School of software, Xi'an jiaotong University, Xi'an China, 710000. He is now a student at SMILES Laboratory, Xi'an Jiaotong University. His current research interests include crowd counting and detection.

Xueming Qian (M'09) received the B.S. and M.S. degrees from the Xi'an University of Technology, Xi'an, China, in 1999 and 2004, respectively, and the Ph.D. degree in electronics and Information engineering from Xi'an Jiaotong University, Xi'an, China, in 2008. From 2011 to 2014, he was an Associate Professor with Xi'an Jiaotong University, where he is currently a Full Professor and the Director of SMILES Lab. He was a Visiting Scholar with Microsoft Research Asia, Beijing, China, from 2010 to 2011. His current research interests include social media big data mining and search. Prof. Qian was a recipient of the Microsoft Fellowship in 2006 and the Outstanding Doctoral Dissertations of Xi'an Jiaotong University and Shaanxi Province in 2010 and 2011, respectively.

Li Zhu is a professor in School of Software, Xi'an Jiao-tong University. He received his M.S. and Ph.D. degrees from Xi'an Jiaotong University in 1995 and 2000, respectively. He received his B.S. degree from Northwestern Polytechnical University in 1989. His main research interests include multimedia processing & communication, parallel computing and networking.