Supplementary Materials for "DR.VIC: Decomposition and Reasoning for Video Individual Counting"

Tao Han[1†], Lei Bai[2†], Junyu Gao[1], Qi Wang[1*], Wanli Ouyang[2]
[1] Northwestern Polytechnical University, Xi'an 710072, P.R. China
[2] The University of Sydney, SenseTime Computer Vision Group, Australia
hantao10200@mail.nwpu.edu.cn, {baisanshi, gjy3035, crabwq}@gmail.com,
wanli.ouyang@sydney.edu.au

The supplementary file provides more details for paper " DR.VIC: Decomposition and Reasoning for Video Individual Counting", including the following aspects.

1) **Extra Experiments**,
2) **Detailed Network Architectures**,
3) **Limitations**,
4) **Potential Negative Societal Impact**,
5) **More Visualization Results**.
6) **Video Demonstration**

1. Extra Experiments

1.1. MIAE and MOAE

In the experiment part, we propose the Mean Inflow/Outflow Absolute Error (MIAE/MOAE) to evaluate the inflow and outflow estimation performance in the pair-wise images. Their detailed formulations are:

$$MIAE@\tau = \frac{\sum_{i=1}^{K}\sum_{t=\tau}^{T_i-1}|\hat{N}_i^+(t) - N_i^+(t)|}{\sum_{i=1}^{K} T_i - \tau}, \quad MOAE@\tau = \frac{\sum_{i=1}^{K}\sum_{t=\tau}^{T_i-1}|\hat{N}_i^-(t) - N_i^-(t)|}{\sum_{i=1}^{K} T_i - \tau} \tag{1}$$

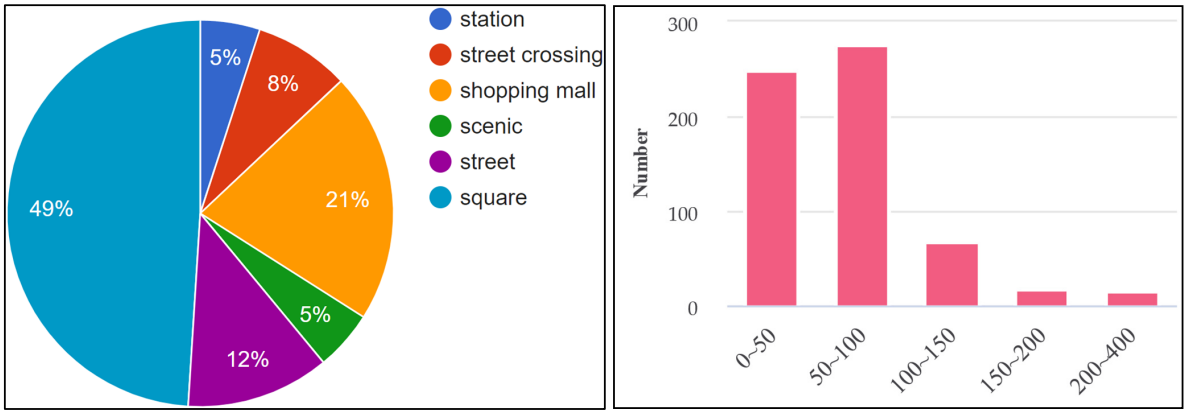
where $N_i^+(t)$ and $\hat{N}_i^+(t)$ represent the ground truth and predicted pedestrian inflow of frame $\mathbf{I}_t$ compared with frame $\mathbf{I}_{t-\tau}$, respectively. Similarly, $N_i^-(t)$ and $\hat{N}_i^-(t)$ are the ground truth and prediction for pedestrian outflow. T_i is the total number of frames of the i-th video. τ is the time interval (*i.e.* frame intervals) for sampling frames.

1.2. Scenes Categorization of SenseCrowd

To comprehensively evaluate the performance of our method on diverse scenes, we manually annotate the 634 videos in SenseCrowd with different scene labels from four perspectives, including:

1) Location: videos are classified into six categories by their locations. Fig. 1 a) shows the proportional distribution.
2) Density: Pedestrian density is a vital factor influencing the counting performance. Videos are divided into five classes according to the crowd density as shown in Fig. 1 b).

† Equal contribution.
* Corresponding author.



a) Location distribution b) Crowd flow density distribution

Figure 1. The pie chart of the captured video location and the statistical histogram of pedestrian counts on SenseCrowd dataset.

3) Time: 76% of videos are recorded in daylight and remaining 24% of videos at night.

4) Space: 77% of videos are captured in outdoor scenes and 23% in indoor scenes.

1.3. Performance of Diverse Scenes on SenseCrowd

Based on the scene labels, we test the counting performance with different scenes and report the MAE in Tab. 1. DRNet outperforms the tracking-based methods [5, 6] and density-based methods (single image counting baselines and [3]) in all scenes with obvious gaps. Take the Square videos as an example, DRNet achieves 8.4 MAE, which is a very low estimation error among all tested methods. At the same time, other methods, such as directly sampling the density map for video individual counting, lead to very poor results.

Tab. 2 reports the nunmber of pedestrians on CroHD videos by directly counting the total identificaitons in the uploaded rtracking results [1]. We find that these figures are far more than the ground truth. Although we can get more reasonable results by searching an optimal sample rate when performing tracking, it still has a huge margin of error. Hence, the tracking-based methods is not robust and advisable when applying it to a long-time video for counting pedestrian number.

Methods	Location						Time		Space	
	L0	L1	L2	L3	L4	L5	Day	Night	Indoor	Outdoor
FairMOT [6]	23.4	25.9	38.8	61.6	32.7	51.7	27.3	35.6	27.7	34.9
HeadHunter-T [5]	22.8	28.6	46.9	61.6	29.0	56.8	29.2	32.8	31.7	29.5
SICC (sampling)	177.2	139.7	210.1	233.7	135.7	229.1	178.8	152.5	143.0	182.1
SICC (maximum)	26.6	26.9	26.2	26.8	34.6	40.4	29.7	24.1	27.0	28.9
LOI [3]	24.4	21.3	20.7	28.5	22.1	44.3	26.8	17.8	22.6	25.4
DRNet	**8.4**	**11.2**	**18.1**	**33.6**	**11.2**	**29.9**	**11.8**	**14.1**	**12.6**	**12.2**

Table 1. Video Individual Counting performance on SenseCrowd dataset measured by MAE. $L0 \sim L5$ represents six location categories: square, shopping mall, street crosses, scenic, street, and station, respectively.

Methods	CroHD11	CroHD12	CroHD13	CroHD14	CroHD15	MAE↓	MSE↓	MRAE(%)↓
	133	737	734	1040	321			
FairMOT [6]	366	3215	7011	2626	2337	2518.0	3230.3	428.1
HeadHunter-T [5]	307	2145	2556	1531	888	892.4	1085.8	166.0

Table 2. Video individual counting performance of thracking-based methods. The underline fonts represent the number of ground truth pedestrians. These results show that the existing tracking methods cannot be directly appplied to this new task.

2. Network Architectures

Tab. 3 elaborates the detailed network architecture of DRNet, including the image representation backbone and head descriptor extraction module, which is consisted a head localization branch and a descriptor generation branch.

Image Representation Backbone. Tab. 3 explains the VGG16 configurations in DRNet. In this table, "k(3,3)-c256-s1-BN-R" represents the convolutional or de-convolutional operation with kernel size of 3×3, output channels of 256, and stride size of 1. The "BN" and "R" mean that the Batch Normalization and ReLU layer are added to this convolutional layer. With the VGG-16 [4] backbone, we output three stages features and fuse them with the Feature Pyramid Networks (FPN) [2].

Head Localization Branch. Tab. 3 also explains the configurations for extracting head proposals. "ResBlock-c256-s1-BN-R" represents a residual CNN module with three convolution layers, 256 output channels, and stride size of 1. This branch finally outputs one-channel density map with the same shape as the input image.

Descriptor Generation Branch. Similarly, we use two Residual modules and two convolution layers to further refine the feature map for head descriptor generation, which receives 384-channel feature maps and produces a 256-channel feature maps.

Image Representation Backbone (VGG16)	
Stage1 (1/4)	conv1: [k(3,3)-c64-s1-BN-R] ... conv7: [k(3,3)-c256-s1-BN-R]
Stage2 (1/8)	... conv10: [k(3,3)-c512-s1-BN-R]
Stage3 (1/16)	... conv13: [k(3,3)-c512-s1-BN-R]
output channels: [Stage1:256, Stage2:512, Stage3:512]	
FPN Module (output channels: 576)	FPN Module (output channels: 384)
Head Descriptor Extraction	
Head Localization Branch	**Descriptor Generation Branch**
Dropout2d(0.2) ResBlock-c256-s1-BN-R ResBlock-c128-s1-BN-R deconv:k(2,2)-c64-s2-BN-R conv:k(3,3)-c32-s1-BN-R deconv:k(2,2)-c16-s2-BN-R conv:k(3,3)-c1-s1-R	Dropout2d(0.2) ResBlock-c384-s1-BN-R ResBlock-c256-s1-BN-R conv:k(3,3)-c256-s1-BN-R conv:k(3,3)-c256-s1

Table 3. Detailed network architecture of DRNet.

3. Limitations

While our work achieves promising video individual counting, it has two limitations:

Simplification: In this work, we propose the direction to simplify the video individual counting to inflow estimation between image pairs according to the observation in real world datasets. While effective, this direction can not handle the cases that people pop in and out of the scene in a very shot time (such as 1 second) and the cases that people re-enter the scene after a relative long period (e.g., 1 minute). Thus, this direction is not capable of generating 100% accurate counting.

Evaluation on very long videos: In the real world applications, counting on very long videos (e.g., 1 hour) are useful. However, our experiments does not cover this case due to the limitations in datasets. We will collect and label long videos for this task in the future work.

4. Potential negative societal impact

Employment impact: The development of intelligent systems will inevitably require fewer human resources. That means fewer people will be hired in some related fields. For example, the video individual counting discussed in this paper is usually done manually by some professional staffs. Once this technology is applied in practice, some security personnel and

management personnel may be affected as the job opportunity may be cut down. This technology can also be transferred to some traditional industries, such as commodity statistics, which will also affect the employment of some workers.

Environmental impact: The training of the model involved in this technology requires considerable electrical support, which would consume a certain amount of energy. We suggest using clean energy for decreasing the impact on environment.

5. More visualization results

Fig. 2 and Fig. 3 present more visualization samples on the SenseCrowd test set. Overall, these samples on a variety of scenes demonstrate that DRNet achieves promising inflow (and additional outflow) reasoning performance, which ensures the success of DRNet in video individual counting. However, we can still observe that some difficult scenes (e.g., high density, occlusion, person multi-view and scale variations *etc*.) have a lot of rooms for improvement, which also points out a direction for future research.

6. Video demonstration

We also make a video demo to showcase the performance of DRNet for Video Individual Counting, please check it if interested. Fig. 4 shows a screenshot of the demo, which demonstrates the pedestrian count performance with the time goes by. In the video clip, it can be found that DRNet can effectively count people for several minutes even in the poor light scene. Besides, we also notice that the accumulated error will make the prediction number go away from the ground truth number gradually, which also tell us the future concentration in this task is how to further eliminate the accumulated error. Overall, this demonstration shows the video individual counting technique would be possible to help improving the social management in the near future. For watching the complete video, please move to https://youtu.be/CIqexlvYT4g or https://www.bilibili.com/video/BV1cY411H7hr/.

References

[1] Motchallenge. [Online]. https://motchallenge.net. 2

[2] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *CVPR*, pages 2117–2125, 2017. 3

[3] James Munkres. Algorithms for the assignment and transportation problems. *Journal of the society for industrial and applied mathematics*, 5(1):32–38, 1957. 2

[4] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 3

[5] Ramana Sundararaman, Cedric De Almeida Braga, Eric Marchand, and Julien Pettre. Tracking pedestrian heads in dense crowd. In *CVPR*, pages 3865–3875, 2021. 2

[6] Yifu Zhang, Chunyu Wang, Xinggang Wang, Wenjun Zeng, and Wenyu Liu. Fairmot: On the fairness of detection and re-identification in multiple object tracking. *IJCV*, pages 1–19, 2021. 2

Figure 2. Visualization samples of SenseCrowd. The green and red circles in the image pairs denote matched and unmatched pedestrians, respectively. The red, blue, and green points in prediction results respectively denote correctly identified flows (inflow or outflow in the corresponding columns), missed flows, and over-counted flows, respectively.

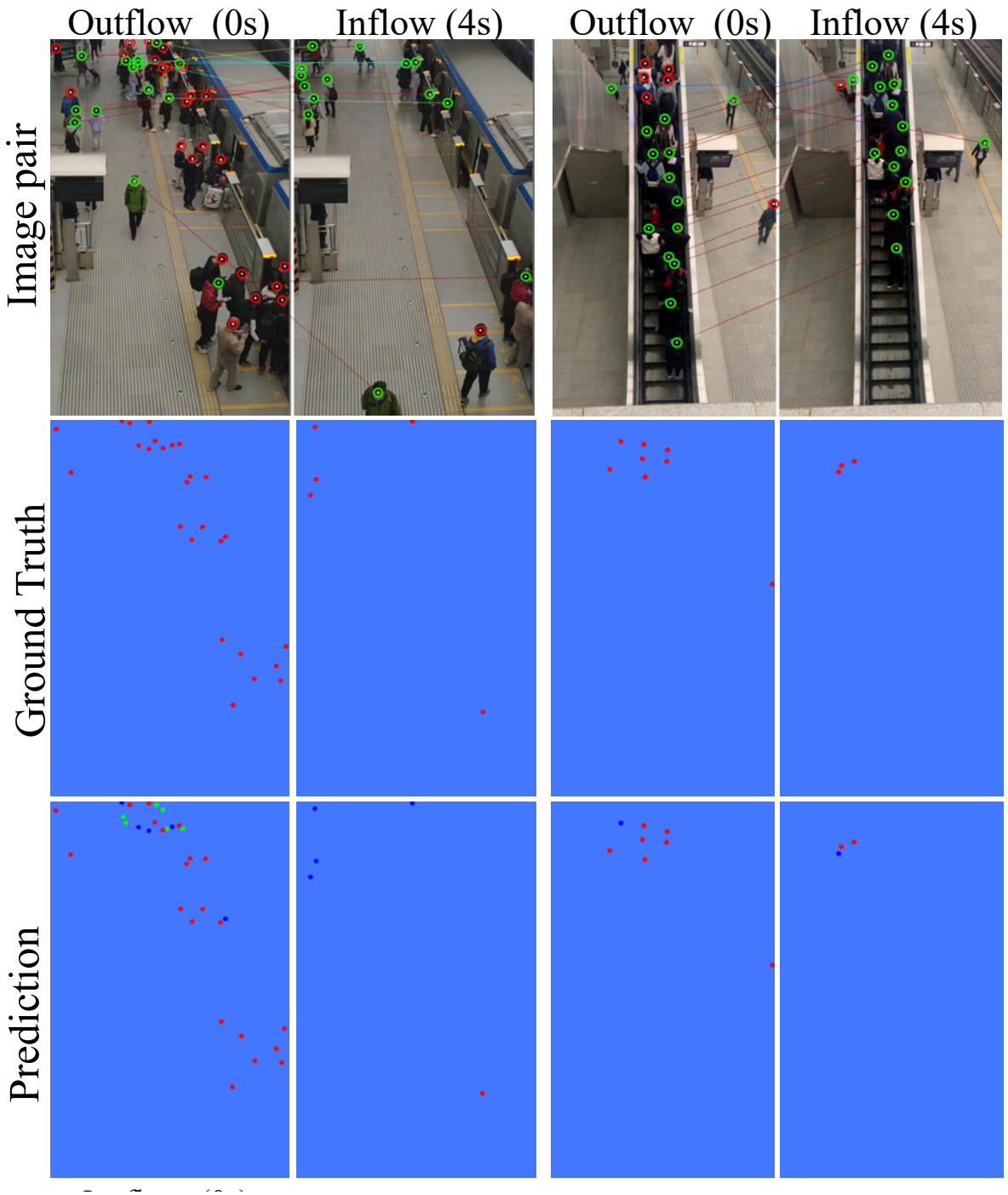


Figure 3. Visualization samples of SenseCrowd. The green and red circles in the image pairs denote matched and unmatched pedestrians, respectively. The red, blue, and green points in prediction results respectively denote correctly identified flows (inflow or outflow in the corresponding columns), missed flows, and over-counted flows, respectively.

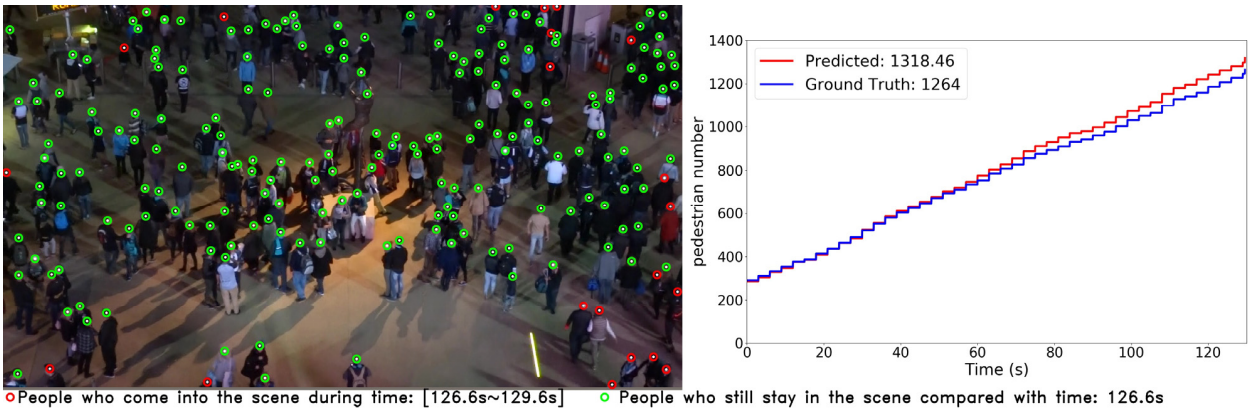


Figure 4. The screenshots of the video demonstration. Left: red circles depict the people who entered the scene from 3 seconds ago to now. green circles represent the people who still stay in the scene compared with the frame at 3 seconds ago. Right: red curve and blue curve represent the predicted and ground truth accumulated pedestrian count (including the initial crowd count and the later inflow-crowd) from time 0, respectively.