Frontiers of Information Technology & Electronic Engineering
www.jzus.zju.edu.cn; engineering.cae.cn; www.springerlink.com
ISSN 2095-9184 (print); ISSN 2095-9230 (online)
E-mail: jzus@zju.edu.cn

# Supplementary materials for

Jiaqi GAO, Jingqi LI, Hongming SHAN, Yanyun QU, James Z. WANG, Fei-Yue WANG, Junping ZHANG, 2023. Forget less, count better: a domain-incremental self-distillation learning benchmark for lifelong crowd counting. *Front Inform Technol Electron Eng*, 24(2):187-202.
https://doi.org/10.1631/FITEE.2200380

## 1 Domain concept and gaps of different datasets

A domain $D$ consists of two components, a feature space $X$ and a marginal probability distribution $P(x)$ (i.e., $D = (X, P(x))$), according to the definition in Weiss et al. (2016). This implies that if two domains ($D_A$ and $D_B$) are different, they may have either different feature spaces ($X_A \neq X_B$) or different marginal probability distributions ($P(X_A) \neq P(X_B)$). In crowd counting tasks, on one hand, off-the-shelf datasets are captured from different cameras or different scenarios like streets, museums, or gyms, so the data distributions are different; on the other hand, from the Bayesian perspective (i.e., $P(x) = P(c)/P(c|x)$, where $c$ is the number of persons and $x$ is the given crowd image), the marginal probability distribution $P(x)$ of each dataset is different. $P(c|x)$ is our single learning model $f(\cdot)$, which has fixed capacity and maps from the input images $x$ to the estimated count number $c$. $P(c)$ represents the population density of each dataset, and varies from dataset to dataset, as shown in Fig. S1. Thus, there exist domain gaps among these crowd counting datasets with different $P(x)$'s. This is our further theoretical analysis of the core domain concept.

Specifically, the ShanghaiTech PartA dataset is collected from the Internet with the highly variant density distribution ranging from 33 to 3139 pedestrians per image. As a larger crowd counting dataset, the UCF-QNRF dataset includes 1535 images collected from several image search engines like Google Image Search and Flicker. In contrast to other available public datasets, the NWPU-Crowd dataset is a generally more extensive and more crowded dataset annotating heads from 0 to 20 033 per image, which first introduces negative samples like extremely-high-density images and images containing no person. As shown in Fig. S1, the ShanghaiTech PartB dataset contains fewer persons per image (123 persons on average) compared to the three other datasets (501, 815, and 418 persons on average) used for experiments in our study. Such a prominent domain shift problem motivates us to investigate the catastrophic forgetting and generalization issues in the lifelong crowd counting task in this study.

In the proposed lifelong crowd counting task, data come from non-stationary and changing distributions, which means $P_{t_i}(X, Y) \neq P_{t_j}(X, Y)$, where $t_i$ and $t_j$ represent different time-steps $t$, and $X$ and $Y$ are the crowd images and their corresponding ground truth density maps (labels), respectively. Different from typical crowd counting, the distribution shifting problem raises a challenge in lifelong crowd counting. At different time-steps, the marginal distribution of crowd images $X$ shifts among different datasets, while the generation of the ground truth remains unchanged, which is $P_{t_i}(X) \neq P_{t_j}(X)$ and $P_{t_i}(Y|X) = P_{t_j}(Y|X)$. To tackle the lifelong crowd counting task, we specifically present one unseen dataset (JHU-Crowd++) that focuses on the generalization of crowd counting, which is much larger than any of the existing datasets, including seen and unseen domains.

To construct the seen domains, we organize four popular crowd counting datasets, including ShanghaiTech PartA, ShanghaiTech PartB, UCF-QNRF, and NWPU-Crowd. The training set (7092 images) comes from the four different datasets' training sets. The current domain performance and forgetting degree are evaluated on the corresponding test sets from the four datasets. To illustrate the model generalization ability, we choose only the test set of the JHU-Crowd++ dataset (1600 images) as the unseen domain dataset,

because it has a more significant count span. For a fair comparison with other training paradigms, none of the images in the JHU-Crowd++ dataset are trained during the lifelong learning process.

## 2 Effect of different training orders

1. Forgetting degree analysis

We compare the results of our proposed framework with different training orders and the corresponding baseline models in Table S1. The results clearly show that our method can mitigate the forgetting phenomenon in the lifelong crowd counting process compared with the vanilla sequential training strategy under the circumstances of different training orders.

2. Generalization analysis

To avoid the generalization performance improvement caused by a particular training order, we conduct the same experiments with different training orders of four benchmark datasets. As shown in Table S2, the model still achieves outstanding performance on the unseen JHU-Crowd++ dataset in contrast with the

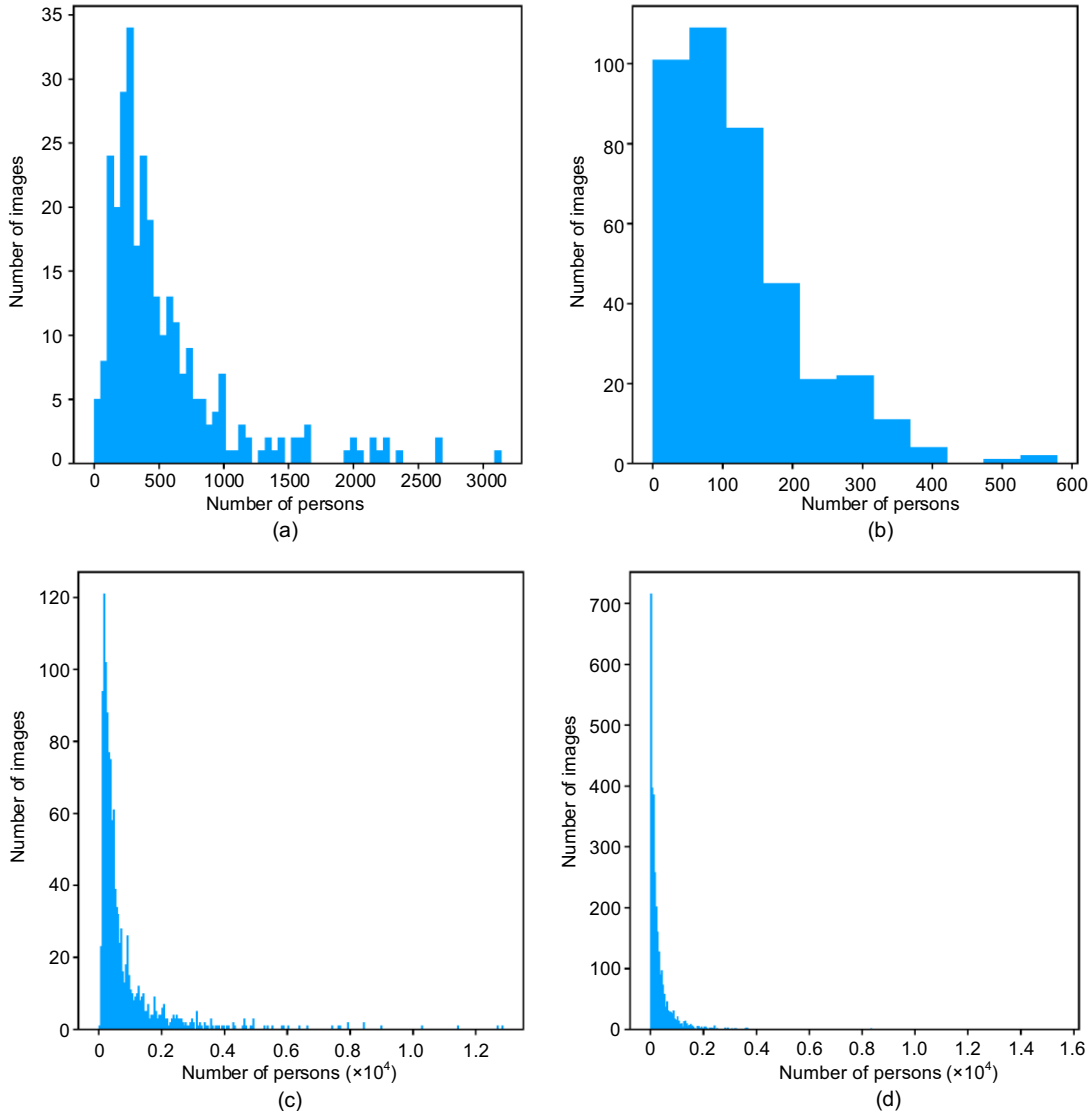


**Fig. S1 Data distributions of four benchmark datasets: (a) ShanghaiTech PartA; (b) ShanghaiTech PartB; (c) UCF-QNRF; (d) NWPU-Crowd**

**Table S1 Forgetting degree comparison results with different training orders**

| Training order | MAE | | | | RMSE | | | | mMAE | mRMSE | nBwT |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | SHA | QNRF | SHB | NWPU | SHA | QNRF | SHB | NWPU | | | |
| SHA→QNRF→SHB→NWPU(BASELINE) | 92.9 | 100.1 | 7.7 | 90.0 | 157.8 | 179.4 | 12.4 | 393.6 | 72.7 | 185.8 | 0.371 |
| SHA→QNRF→SHB→NWPU(FLCB) | 68.8 | 84.3 | 7.8 | 76.6 | 113.9 | 160.1 | 12.2 | 364.2 | **59.4** | **162.6** | **−0.010** |
| NWPU→QNRF→SHA→SHB(BASELINE) | 124.9 | 240.1 | 7.4 | 218.2 | 229.0 | 435.8 | 12.5 | 826.5 | 147.7 | 376.0 | 1.576 |
| NWPU→QNRF→SHA→SHB(FLCB) | 62.3 | 78.8 | 10.7 | 94.8 | 108.0 | 138.6 | 20.2 | 417.5 | **61.7** | **171.1** | **0.043** |
| QNRF→SHA→SHB→NWPU(BASELINE) | 87.1 | 107.1 | 10.1 | 100.1 | 162.6 | 212.8 | 16.1 | 462.1 | 76.1 | 213.4 | 0.432 |
| QNRF→SHA→SHB→NWPU(FLCB) | 61.3 | 84.2 | 10.3 | 83.9 | 104.8 | 149.7 | 17.9 | 377.8 | **59.9** | **162.6** | **−0.001** |

The bold number indicates the best performance with the same training order

**Table S2 Generalization comparison results with different training orders on the unseen JHU-Crowd++ dataset**

| Training mode | MAE | RMSE |
|---|---|---|
| ShanghaiTech PartA | 106.0 | 338.3 |
| ShanghaiTech PartB | 154.4 | 530.2 |
| UCF-QNRF | 97.8 | 315.9 |
| NWPU-Crowd | 94.5 | 323.4 |
| JOINT | 89.8 | 318.7 |
| NWPU→QNRF→SHA→SHB | **87.2** | **287.5** |
| SHA→QNRF→SHB→NWPU | **84.8** | **264.8** |
| QNRF→SHA→SHB→NWPU | **83.4** | **264.8** |

The bold numbers indicate the performances of lifelong crowd counting models with different training orders

single-domain training settings and joint training strategy. Compared with joint training and individual training, the results verify the effectiveness of our proposed domain-incremental self-distillation learning framework in consistently strengthening the model generalization ability with different training orders.

## References

Weiss K, Khoshgoftaar TM, Wang DD, 2016. A survey of transfer learning. *J Big Data*, 3(1):9. https://doi.org/10.1186/s40537-016-0043-6