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

Forget less, count better: a domain-incremental self-distillation learning benchmark for lifelong crowd counting*#

Jiaqi GAO†1, Jingqi LI[1], Hongming SHAN[2,3], Yanyun QU[4], James Z. WANG[5], Fei-Yue WANG†‡6, Junping ZHANG†‡1

[1]*Shanghai Key Laboratory of Intelligent Information Processing, School of Computer Science, Fudan University, Shanghai 200433, China*

[2]*Institute of Science and Technology for Brain-inspired Intelligence, Fudan University, Shanghai 200433, China*

[3]*Shanghai Center for Brain Science and Brain-inspired Technology, Shanghai 201210, China*

[4]*School of Information Science and Technology, Xiamen University, Xiamen 361005, China*

[5]*College of Information Sciences and Technology, the Pennsylvania State University, University Park, PA 16802, USA*

[6]*State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China*

†E-mail: jqgao20@fudan.edu.cn; feiyue.wang@ia.ac.cn; jpzhang@fudan.edu.cn

Received Sept. 7, 2022; Revision accepted Dec. 26, 2022; Crosschecked Jan. 16, 2023

Abstract: Crowd counting has important applications in public safety and pandemic control. A robust and practical crowd counting system has to be capable of continuously learning with the newly incoming domain data in real-world scenarios instead of fitting one domain only. Off-the-shelf methods have some drawbacks when handling multiple domains: (1) the models will achieve limited performance (even drop dramatically) among old domains after training images from new domains due to the discrepancies in intrinsic data distributions from various domains, which is called catastrophic forgetting; (2) the well-trained model in a specific domain achieves imperfect performance among other unseen domains because of domain shift; (3) it leads to linearly increasing storage overhead, either mixing all the data for training or simply training dozens of separate models for different domains when new ones are available. To overcome these issues, we investigate a new crowd counting task in incremental domain training setting called lifelong crowd counting. Its goal is to alleviate catastrophic forgetting and improve the generalization ability using a single model updated by the incremental domains. Specifically, we propose a self-distillation learning framework as a benchmark (forget less, count better, or FLCB) for lifelong crowd counting, which helps the model leverage previous meaningful knowledge in a sustainable manner for better crowd counting to mitigate the forgetting when new data arrive. A new quantitative metric, normalized Backward Transfer (nBwT), is developed to evaluate the forgetting degree of the model in the lifelong learning process. Extensive experimental results demonstrate the superiority of our proposed benchmark in achieving a low catastrophic forgetting degree and strong generalization ability.

Key words: Crowd counting; Knowledge distillation; Lifelong learning
https://doi.org/10.1631/FITEE.2200380 **CLC number:** TP391

‡ Corresponding authors

* Project supported by the National Natural Science Foundation of China (Nos. 62176059, 62101136, and U1811463), the Shanghai Municipal Science and Technology Major Project (No. 2018SHZDZX01), Zhangjiang Lab, the Shanghai Municipal of Science and Technology Project (No. 20JC1419500), the Shanghai Sailing Program (No. 21YF1402800), the Natural Science Foundation of Shanghai (No. 21ZR1403600), and the Shanghai Center for Brain Science and Brain-inspired Technology

Electronic supplementary materials: The online version of this article (https://doi.org/10.1631/FITEE.2200380) contains supplementary materials, which are available to authorized users

ORCID: Jiaqi GAO, https://orcid.org/0000-0003-0910-0801; Junping ZHANG, https://orcid.org/0000-0002-5924-3360

1 Introduction

Crowd counting is to predict the number of persons in an image or a video sequence. Accurate crowd counting for crowded scenes has important applications such as traffic control, preventing stampedes from occurring, and estimating participation in large public events like parades. For example, during a pandemic, authorities may need to maintain social distancing for public spaces to minimize the risk of infection. Thus, crowd counting systems are usually deployed in multiple diverse scenarios, such as malls, museums, squares, and public squares. For one site, the running system is expected to continually handle the non-stationary data with different densities, illumination, occlusion, and various head scales. For multiple sites, the system should also consider dozens of scenes and perspective information.

As data are increasingly produced and labeling is time-consuming, the new domain data available for training are usually collected and labeled incrementally. We may ask: how can we sustainably handle the crowd counting problem in multiple domains using a single model when the newly available domain data arrive? We try to find the best potential solution to this question from the following aspects.

Currently, most crowd counting approaches (Zhang YY et al., 2016; Sam et al., 2017; Sindagi and Patel, 2017; Cao et al., 2018; Li YH et al., 2018; Chen XY et al., 2019; Liu WZ et al., 2019; Ma et al., 2019, 2020; Tan et al., 2019; Bai et al., 2020; Jiang XH et al., 2020b; Tian et al., 2020; Song et al., 2021) concentrate on training an independent model for each single domain dataset. They heavily rely on the assumption that images from both the training set and test set are independent and identically distributed. Although producing promising counting performance in the corresponding domain, such a training strategy, as shown in Fig. 1a, has drawbacks in dealing with multiple and incremental new datasets, which are common in the real world, e.g., when limited labeled data from a new site are available before applying the model at the site. One drawback is that these separately trained models often have low generalization ability when dealing with new, unseen domain data due to the domain shift evidenced in Table 1. Another is that saving multiple different sets of trained parameters from distinct domains for inference is not economical when

Table 1 The mean absolute error (MAE) scores of our reproduced DM-Count (Wang BY et al., 2020) model separately trained in a single dataset and tested over other datasets, showing obvious performance drop due to domain discrepancy

Dataset	MAE			
	SHA	SHB	QNRF	NWPU
SHA	**59.7**	19.0	143.3	161.1
SHB	124.6	**7.0**	209.9	179.1
QNRF	69.6	14.0	**85.6**	124.8
NWPU	74.7	11.7	100.9	**88.4**

The datasets in the first column are used for training, and those in the second row for testing. The bold number indicates that the model achieves the best performance when the training and test datasets are from the same domain

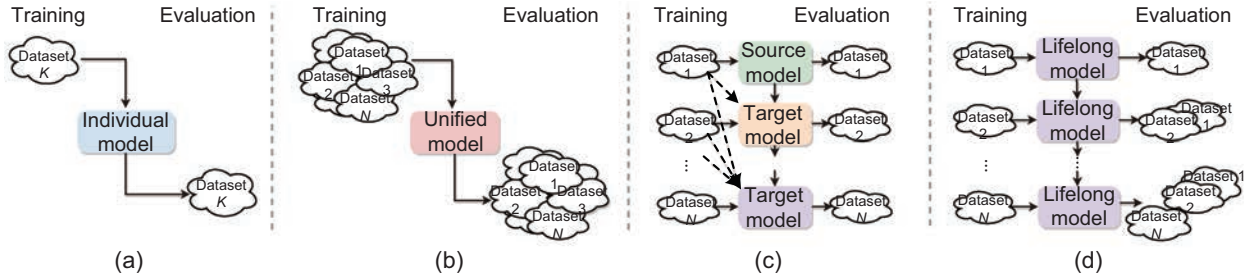

Fig. 1 The conceptual differences of four training paradigms: (a) directly training an individual model for each dataset; (b) training a unified model by mixing all datasets from different domains; (c) leveraging previous data or models to improve the performance on the target domain dataset; (d) ours: lifelong learning with incremental domains to improve the performance among all domains. In (c), the dashed lines indicate that the past domain data may be used repeatedly to improve the performance in the target domain dataset. In (d), our proposed FLCB (forget less, count better) model does not replay any previous domain data and evaluates all domain datasets at the training stage. Without storing previous domain data, FLCB itself can still sustainably handle the crowd counting problem among multiple domains, being updated by the new available domain dataset only

deploying them to hundreds of thousands of real-world sites. Training a shared and universal model from scratch by mixing all the data (also known as joint training) or sequential training for each newly incoming dataset may improve the performance on the unseen domains (Figs. 1b and 1c). Nevertheless, both paradigms still have some limitations. The joint training strategy (Ma et al., 2021; Yan et al., 2021) requires storing all training data from previous domains when the newly available data arrive, leading to lengthy training time and high storage overhead. Meanwhile, the sequential training strategy will dramatically deteriorate the model's performance among previous domains after training the new domain data, i.e., catastrophic forgetting.

To deal with the aforementioned forgetting, generalization, and storage overhead issues, inspired by the learning mechanism of mammals, we investigate a new task of crowd counting in this study, termed lifelong crowd counting, which can sustainably learn with the new domain data and concurrently alleviate catastrophic forgetting and performance drop among preceding domains under the domain-incremental training settings (Fig. 1d). Note that the goal of the proposed lifelong crowd counting task is different from that of previous cross- and multi-domain crowd counting tasks (Chen BH et al., 2021; Ma et al., 2021; Yan et al., 2021). During the whole lifelong learning process with incremental training data, the goal is to maximize the overall performance among all domains—previously trained, newly arriving, and unseen—instead of focusing only on the target domain performance. We consider the trade-off between the forgetting degree and the generalization ability of the models. In particular, we develop a novel benchmark of domain-incremental lifelong crowd counting with the help of knowledge self-distillation techniques. The proposed benchmark has both strong generalization ability on unseen domains and low forgetting degrees among seen domains. This enables the model to have sustainable counting capability when new data arrive in the future. In our experiments, we use four fruitful crowd counting backbones, CSRNet (Li YH et al., 2018), SFANet (Zhu L et al., 2019), DM-Count (Wang BY et al., 2020), and DKPNet (Chen BH et al., 2021), to illustrate the effectiveness and superiority of our proposed framework.

The contributions of this work can be summarized as follows:

1. To the best of our knowledge, this is the first work to investigate lifelong crowd counting by considering the catastrophic forgetting and generalization ability issues. Our method may serve as a benchmark for further research in the lifelong crowd counting community.

2. We design a balanced domain forgetting loss function (BDFLoss) to prevent the model from dramatically forgetting the previous knowledge when being trained on the newly arriving crowd counting dataset.

3. We propose a new quantitative metric, normalized Backward Transfer (nBwT) of lifelong crowd counting, to measure the forgetting degree of trained models among seen data domains. We treat the mean absolute error (MAE) as the criterion for evaluating model generalization on the unseen data domain.

4. Extensive experiments indicate that our proposed method has a lower degree of forgetting compared with sequential training and outperforms the joint training strategy on the unseen domain with a much lower MAE score and time and space complexity.

2 Related work

2.1 Crowd counting

Traditional detection- and regression-based methods extract handcrafted features such as scale invariant feature transform (SIFT) (Lowe, 1999) and histogram of oriented gradient (HoG) (Dalal and Triggs, 2005) to detect individual heads (Dalal and Triggs, 2005; Leibe et al., 2005; Tuzel et al., 2008; Dollar et al., 2012) or directly regress the count number (Chan and Vasconcelos, 2009). Nevertheless, these models cannot learn the spatial information of person distribution to make accurate predictions in highly congested scenes. Most of the latest crowd counting approaches are built upon deep learning methods to estimate a density map for a given image. Many researchers design various architectures like fully convolutional networks (Wang C et al., 2015; Zhang C et al., 2015), multi-column networks (Boominathan et al., 2016; Zhang YY et al., 2016; Sam et al., 2017; Sindagi and Patel, 2017), scale aggregation or scale pyramid networks (Cao

et al., 2018; Chen XY et al., 2019; Liu LB et al., 2019; Jiang XH et al., 2020b; Zhao et al., 2020; Song et al., 2021), and attention mechanisms (Guo et al., 2019; Liu N et al., 2019; Zhu L et al., 2019; Jiang XH et al., 2020a; Sindagi and Patel, 2020) to extract the multi-scale feature representations to deal with scale variation and non-uniform distribution issues. CSRNet (Li YH et al., 2018) points out the multi-scale feature redundancies among multi-branch architectures and proposes a new deeper single-column convolutional neural network (CNN) with dilated convolutions to capture different receptive fields. ADCNet (Bai et al., 2020) extends the discrete dilated ratio (integer value) to a continuous value to match the large-scale variation and self-correct the density map using the expectation-maximization (EM) algorithm. Local region modeling methods (Liu L et al., 2020; Jiang SQ et al., 2020) also help correct the local information. Most off-the-shelf crowd counting models focus on single domain learning. The models will be retrained when the new domain data arrive. In our study, we focus on using a single model to handle multiple incremental datasets for crowd counting.

2.2 Cross-/multi-domain learning

Many researchers exploit the cross-domain problems (Wang Q et al., 2019, 2022; Wu et al., 2021; Zou et al., 2021; Liu WZ et al., 2022) in crowd counting, including cross-scene (Zhang C et al., 2015), cross-view (Zhang Q et al., 2021), and cross-modal (Liu LB et al., 2021). The adversarial scoring network (Zou et al., 2021) is applied to adapt to the target domain from coarse to fine granularity. In addition, cross-domain features can be extracted by the message-passing mechanisms based on a graph neural network (Luo et al., 2020). A semantic extractor (Han et al., 2020) has been designed to capture the semantic consistency between the source domain and target domain to enhance the adapted model. A large synthetic dataset (GCC) (Wang Q et al., 2019) has been released to study the transferability from synthetic data to real-world data. Quite a few researchers (Shi et al., 2019; Xiong et al., 2019; Yang et al., 2020) investigated similar tasks like vehicle counting based on the same crowd counting architectures. Learning with multiple domains simultaneously (Chen BH et al., 2021; Ma et al., 2021; Yan et al., 2021) has also been preliminarily explored, and is required to mix all the data for training at the same time. DCANet (Yan et al., 2021) uses a channel-attention-guided multi-dilation module to assist the model in learning a domain-invariant representation, while DKPNet (Chen BH et al., 2021) propagates the domain-specific knowledge with the help of variational attention techniques. Ma et al. (2021) developed a scale alignment component to learn an adaptive rescaling factor for each image patch for better crowd counting. In reality, such cross-domain approaches need a careful alignment module design and place more emphasis on the target domain performance only, while the multi-domain learning methods require more storage overhead to save old domain data. These methods often achieve limited performance in previous (source) domains. In contrast, our proposed lifelong crowd counting task is based on training the domains incrementally (one by one) using a single model, alleviating catastrophic performance drop of the previous domains (forget less), and maintaining the overall performance in all domains (count better). The lifelong crowd counting system can mimic the biological brain to learn sustainably in its lifetime inspired by the learning mechanisms of mammals, i.e., integrating the new knowledge increasingly while maintaining previous memories.

2.3 Lifelong learning

Lifelong learning attempts to alleviate the catastrophic forgetting issues and enhance the model generalization ability when a system increasingly faces non-stationary data. The mainstream strategies are applied to image classification (Kirkpatrick et al., 2017; Lopez-Paz and Ranzato, 2017; Rebuffi et al., 2017; Li ZZ and Hoiem, 2018; Belouadah and Popescu, 2019) and numerical prediction tasks (He YJ and Sick, 2021), which can be categorized into four groups: model-growth approaches (Rusu et al., 2016), rehearsal-based techniques (Lopez-Paz and Ranzato, 2017; Rebuffi et al., 2017), regularization (Kirkpatrick et al., 2017; Rebuffi et al., 2017), and distillation mechanisms (Li ZZ and Hoiem, 2018). Specifically, the model-growth (e.g., product-based neural network (PNN) (Rusu et al., 2016)) and rehearsal-based methods (e.g., GEM (Lopez-Paz and Ranzato, 2017)) require more computational and memory costs because they either instantiate a new network or replay old data when learning new classes or tasks. LwF (Li ZZ and Hoiem, 2018) is a combi-

nation of the distillation networks and fine-tuning to boost the overall performance. However, the aforementioned classification-based lifelong learning approaches cannot migrate to the crowd counting task directly because counting is an open-set problem (Xiong et al., 2019) by nature, whose value ranges from zero to positive infinity in theory. Latent feature representations with general visual knowledge together with high-level semantic information at the output layer play a crucial role in such dense prediction tasks. Therefore, in this paper, we propose a simple yet effective self-distillation loss at both the feature level and the output level for lifelong crowd counting to alleviate catastrophic forgetting with a low time and space complexity.

3 Methodology

In this section, we will first introduce concrete formalized definitions of typical crowd counting and the proposed lifelong crowd counting. After that, we describe the details of our proposed domain-incremental self-distillation lifelong crowd counting benchmark including model architectures and the proposed loss function.

3.1 Problem formulation

3.1.1 Typical crowd counting

A typical crowd counting task can be regarded as a density map regression problem, training and validating in a single domain, as shown in Fig. 1a. Suppose that one dataset $D_{\mathcal{M}} = \langle X_{\mathcal{M}}, Y_{\mathcal{M}} \rangle$ contains $\mathcal{M}$ training images and the corresponding annotations. Then, a binary map $\boldsymbol{B}$ is easy to obtain given the coordinates of pedestrian heads per image, which can be formally defined as follows:

$$B_{(i,j)} = \begin{cases} 1, & \text{head center } (i,j), \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

The ground truth density map Y is generated by employing the Gaussian kernel G_σ to smooth the binary map:

$$Y = G_\sigma \circledast \boldsymbol{B}. \tag{2}$$

Here, $\circledast$ represents the convolution operation. Then, the typical crowd counting is transformed to regress the generated density maps. The pixel-level $\mathcal{L}_2$ loss is the most commonly used one to optimize the model $\mathcal{F}(\cdot;\theta)$ with parameter θ by minimizing the difference between predictions and ground truths:

$$\min_{\theta} \frac{1}{\mathcal{M}} \sum_{m=1}^{\mathcal{M}} \mathcal{L}_2(\mathcal{F}(X_m;\theta), Y_m) \ . \tag{3}$$

3.1.2 Lifelong crowd counting

We propose a new, challenging, yet practical crowd counting task, i.e., lifelong crowd counting, for investigating the catastrophic forgetting and model generalization problems in training domain-incremental datasets. Different from previous works that maintained good performance only in a single target domain, the lifelong crowd counting model could be sustainably optimized over the new incoming datasets to maximize the performance among all domains.

For convenience, we first define some key notations as follows and introduce the details of the lifelong crowd counting process. A sequence of $\mathcal{N}$ domain datasets $\{\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_{\mathcal{N}}\}$ is prepared to train the lifelong crowd counter $\mathcal{G}^*(\cdot;\psi)$ with parameters ψ one by one. $X^{(t)}_{\mathcal{M}_t}$ and $Y^{(t)}_{\mathcal{M}_t}$ are the training images and corresponding ground truth density maps with $\mathcal{M}_t$ samples from the t^{th} domain $\mathcal{D}_t$, respectively. Here, we assume that different datasets are coming from different domains with their own distinct data distributions, i.e., $p(X^{(i)}) \neq p(X^{(j)}),\ i \neq j$, because they are normally captured from different cameras or different scenarios like streets, museums, and gymnasiums. The model is initially trained from scratch over the first domain and then trained and optimized by the rest of the other datasets sequentially. The optimal object ψ^* is defined as follows:

$$\arg\min_{\psi} \sum_{t=1}^{\mathcal{N}} \mathbb{E}_{(X^{(t)}_{\mathcal{M}_t}, Y^{(t)}_{\mathcal{M}_t})}[\mathcal{L}(\mathcal{G}^{(t)}(X^{(t)}_{\mathcal{M}_t};\psi), Y^{(t)}_{\mathcal{M}_t})], \tag{4}$$

where $\mathcal{G}^{(t)}(\cdot;\psi)$ represents the t^{th} model for training the t^{th} dataset $X^{(t)}_{\mathcal{M}_t}$ with $\mathcal{M}_t$ samples. The ultimate model is expected to achieve decent performance among seen and unseen domains. What deserves to be pointed out is that lifelong crowd counting is distinct from cross-domain tasks with different optimization objectives, as well as the training settings. In lifelong crowd counting, the goal is to maximize the performance on both seen and unseen domains instead of maximizing the target domain performance only. Specifically, when the

training data from previous domains are absent or unavailable, lifelong crowd counters could still work efficiently because they are trained and updated only by the newly arriving domain dataset one after another.

3.2 Overview of our proposed framework

Our proposed framework focuses on tackling the catastrophic forgetting and generalization issues under the circumstances of domain-incremental training settings. In this study, we simply regard different crowd counting datasets as different domains because the statistics (mean and variance) of person count are different. The detailed explanations of the domain concept can be seen in the supplementary materials. To be more specific, we propose a novel domain-incremental self-distillation lifelong crowd counting benchmark for sustainable learning with newly arriving data and without an obvious performance drop among previous domains. The key factor is how to effectively leverage the previously learned meaningful knowledge when training over the data from a new domain for better crowd counting. Inspired by the knowledge distillation technique, we expect to use a well-trained model among old domains (teacher model) to guide the currently optimized model with new domain data (student model) to mitigate performance drop among previous domains, considering that the old data may be unavailable. The overview of our proposed framework is illustrated in Fig. 2. We design a self-distillation mechanism plugged into both feature- and output-level layers of the network to constrain the output distribution similarities between the teacher and student models, which can reuse the learned knowledge when facing the new domain data without storing or training the old data repeatedly. Details will be given in Section 3.3. The ultimate model is expected to be deployed to an arbitrary domain to estimate the person count.

For better understanding, the overall training pipeline is described in detail as shown in Algorithm 1. A queue Q collects $\mathcal{N}$ increasingly arriving datasets from different domains to be trained one by one. First, we initialize the first model $\mathcal{G}^{(1)}(\cdot;\psi)$ by training the first available dataset $\mathcal{D}_1$ in queue Q. Another queue P is prepared for future evaluation, receiving the test set popped from Q. After that, the model will be trained and optimized by the subsequent datasets from $\mathcal{D}_2$ to $\mathcal{D}_{\mathcal{N}}$, repeating the following main steps until queue Q is empty:

1. Pop the t^{th} dataset $\mathcal{D}_t$ from queue Q for

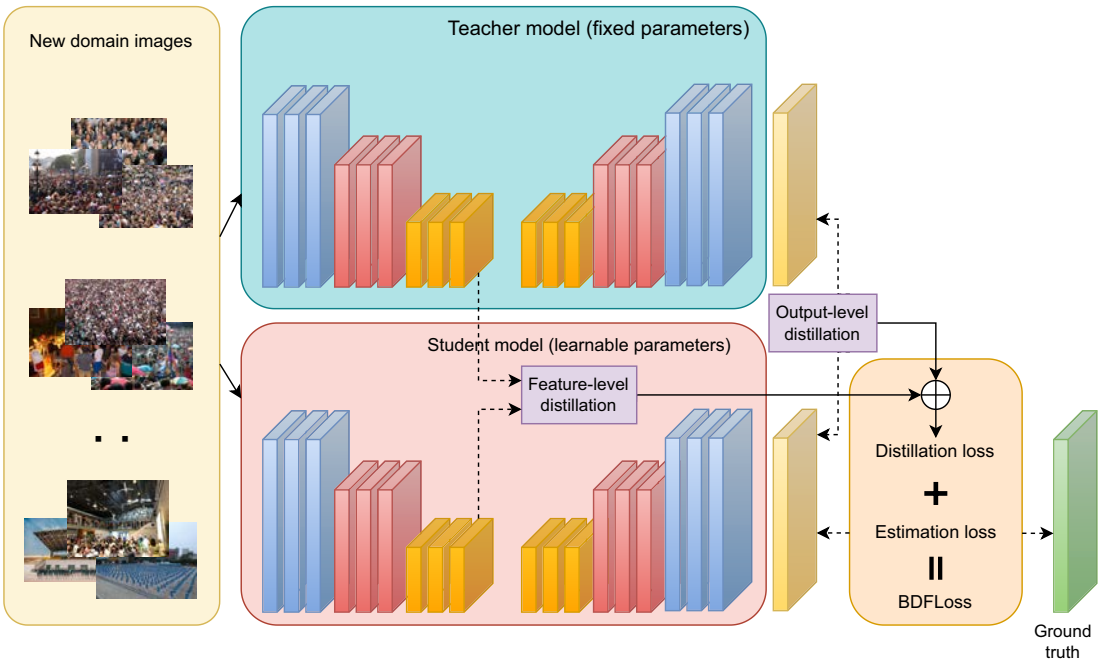

Fig. 2 Overall architecture of our proposed domain-incremental self-distillation learning benchmark (FLCB)

training.

2. Copy the parameters of the last well-trained model $\mathcal{G}^{(t-1)}$ to model $\mathcal{F}(\cdot;\theta)$ as a teacher network for distillation.

3. Train the current t^{th} model $\mathcal{G}^{(t)}(\cdot;\psi)$ over the t^{th} dataset $\mathcal{D}_t$ via the self-distillation loss we propose.

4. Push the t^{th} dataset $\mathcal{D}_t$ into queue P for evaluation when the model converges.

Note that the parameters θ of model $\mathcal{F}(\cdot;\theta)$ are frozen during the lifelong training process. The fixed model is regarded as a teacher network to guide the current student network $\mathcal{G}^{(t)}(\cdot;\psi)$ with learnable parameters ψ to remember old meaningful knowledge for better crowd counting. Eventually, we obtain the final model with the best parameters ψ^*, which can continue to be trained using our proposed framework when the newly coming labeled data are ready in the future. Because we do need to store any previously seen training data to be replayed to train our model, the time and space complexities are approximately $O(\mathcal{N})$ and $\Omega(\mathcal{M})$, respectively, superior to $O(\mathcal{N}^2)$ and $\Omega(\mathcal{N}\times\mathcal{M})$ of joint training. $\mathcal{M}$ is the maximum of $\mathcal{M}_i$. Although the distillation mechanism is required to save an additional model, its storage overhead is negligible compared to storing the entire dataset for retraining.

Algorithm 1 FLCB training pipeline

Notations:
$X^{(t)}_{\mathcal{M}_t}$: the t^{th} training dataset with $\mathcal{M}_t$ samples.
$Y^{(t)}_{\mathcal{M}_t}$: the corresponding density maps of $X^{(t)}_{\mathcal{M}_t}$.
$\mathcal{M}_1, \mathcal{M}_2, ..., \mathcal{M}_{\mathcal{N}}$: samples of each dataset.
P : a queue containing previously seen datasets.
Q : a queue containing future unseen datasets.
$\mathcal{F}^{(t)}(\cdot;\theta)$: teacher model with fixed parameters θ at the t^{th} step.
$\mathcal{G}^{(t)}(\cdot;\psi)$: student model with updated parameters ψ at the t^{th} step.
Input: $\{\mathcal{D}_1, \mathcal{D}_2, ..., \mathcal{D}_{\mathcal{N}}\}$: a sequence of $\mathcal{N}$ domain datasets, $\mathcal{D}_i = \langle X^{(i)}_{\mathcal{M}_i}, Y^{(i)}_{\mathcal{M}_i}\rangle$.
Output: the optimal model parameters ψ^*.

$P \leftarrow \varnothing$
$Q \leftarrow \{\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_{\mathcal{N}}\}$
$\langle X^{(1)}_{\mathcal{M}_1}, Y^{(1)}_{\mathcal{M}_1}\rangle \leftarrow Q$.top()
Q.pop()
Train $\mathcal{G}^{(1)}(X^{(1)}_{\mathcal{M}_1};\psi)$
$\psi^* \leftarrow \arg\min_{\psi} \mathcal{L}^{(1)}_{\text{count}}(\cdot;\psi)$
P.push($X^{(1)}$)
for $t = 2, 3, \ldots, \mathcal{N}$ **do**
  $\mathcal{F}^{(t-1)}(\cdot;\theta) \leftarrow \mathcal{G}^{(t)}(\cdot;\psi^*)$
  $\langle X^{(t)}_{\mathcal{M}_t}, Y^{(t)}_{\mathcal{M}_t}\rangle \leftarrow Q$.top()
  Train $\mathcal{G}^{(t)}(X^{(t)}_{\mathcal{M}_t};\psi)$
  $\psi^* \leftarrow \arg\min_{\psi} \mathcal{L}^{(t)}_{\text{count}}(\cdot;\psi) + \lambda\mathcal{L}^{(t)}_{\text{distill}}(\cdot;\theta,\psi)$
  P.push($\langle X^{(t)}_{\mathcal{M}_t}, Y^{(t)}_{\mathcal{M}_t}\rangle$)
  Q.pop()
  Test all seen datasets in P with $\mathcal{G}^{(t)}(\cdot;\psi^*)$
end for
Return ψ^*
// Time complexity: $O(\mathcal{N})$
// Space complexity: $\Omega(\mathcal{M})$
// $\mathcal{M} = \max\{\mathcal{M}_i | i = 1, 2, ..., \mathcal{N}\}$

3.3 Balanced domain forgetting loss

To balance the model plasticity (the ability to learn new data) and stability (the ability to remember previous knowledge), we propose a novel balanced domain forgetting loss function, i.e., BDFLoss, consisting of mainly counting loss and self-distillation loss. We integrate the optimal transport loss in our basic $\mathcal{L}_1$ counting loss in this study because it has tighter generalization error bounds (Wang BY et al., 2020). $\mathcal{L}_1$ counting loss is defined as follows:

$$\mathcal{L}_1(Y,\hat{Y}) = \frac{1}{\mathcal{M}}\sum_{i=1}^{\mathcal{M}} |Y_i - \hat{Y}_i|, \tag{5}$$

where $\mathcal{L}_1(\cdot,\cdot)$ loss computes the difference between the predicted and actual counts.

The optimal transport loss $\mathcal{L}_{\text{OT}}$ is used to minimize the distribution discrepancy between the predicted density maps and the point-annotated binary maps, defined as follows:

$$\begin{aligned}\mathcal{L}_{\text{OT}}(Y,\hat{Y}) &= \mathcal{W}_c\left(\tfrac{Y}{||Y||_1}, \tfrac{\hat{Y}}{||\hat{Y}||_1}; \mathcal{C}\right)\\ &= \left\langle \alpha^*, \tfrac{Y}{||Y||_1}\right\rangle + \left\langle \beta^*, \tfrac{\hat{Y}}{||\hat{Y}||_1}\right\rangle,\end{aligned} \tag{6}$$

where $\mathcal{W}_c(\mu, v; \mathcal{C})$ is the optimal transport loss with the transport cost $\mathcal{C}$. It aims at minimizing the cost to transform one probability distribution μ to another v. $\mathcal{C}$ is defined as the quadratic transport cost here. α^* and β^* are the optimal solutions to its dual problem:

$$\max_{\alpha,\beta} \langle\alpha,\mu\rangle + \langle\beta,v\rangle \quad \text{s.t.} \quad \alpha_i + \beta_j \le \mathcal{C}_{ij}, \forall i,j. \tag{7}$$

To improve the approximation of the low-density regions of images, we embed a normalized regularization item $\mathcal{L}_{\text{r}}$, defined as follows:

$$\mathcal{L}_{\text{r}}(Y,\hat{Y}) = \frac{1}{\mathcal{M}}\sum_{i=1}^{\mathcal{M}} \frac{1}{2}\left\| \tfrac{Y_i}{||Y_i||_1} - \tfrac{\hat{Y}_i}{||\hat{Y}_i||_1}\right\|_1. \tag{8}$$

Thus, the total count loss is made up of the three aforementioned loss functions with two hyper-parameters, η and γ, which are set to 0.1 and 0.01, respectively, in our experiments.

$$\mathcal{L}_{\text{count}} = \mathcal{L}_1 + \eta\mathcal{L}_{\text{OT}} + \gamma\mathcal{L}_{\text{r}}. \tag{9}$$

When training to the t^{th} domain, the performance among previous domains may degrade dramatically, i.e., catastrophic forgetting, if no constraints are imposed. The self-distillation loss $\mathcal{L}_{\text{distill}}$ is designed to help the model forget less and count better during the lifelong learning process. To be more specific, we regard the current training model $\mathcal{G}^{(t)}(\cdot)$ as the student model, which can be guided by the teacher model $\mathcal{G}^{(t-1)}(\cdot)$ well-trained at the previous step (Fig. 2). The student model is not expected to forget some previously learned knowledge when training in the new domain. Normally, the deep layers of a CNN with a large receptive field contain task-specific and high-level semantic information, while the intermediate layers include general visual knowledge. They are mutually beneficial and complementary, and assist the model in remembering the helpful knowledge learned previously, during the lifelong crowd counting process. Thus, we deploy the self-distillation loss at both the feature level and the output level when the t^{th} new domain dataset arrives for training.

$$\begin{aligned}\mathcal{L}_{\text{distill}}^{(t)} = &\frac{1}{\mathcal{M}_t}\sum_{i=1}^{\mathcal{M}_t}\Big(\underbrace{||\mathcal{G}^{(t-1)}(X_i^{(t)}) - \mathcal{G}^{(t)}(X_i^{(t)})||^2}_{\text{output-level distillation}} \\ &+ \underbrace{||\mathcal{H}^{(t-1)}(X_i^{(t)}) - \mathcal{H}^{(t)}(X_i^{(t)})||^2}_{\text{feature-level distillation}}\Big),\end{aligned} \tag{10}$$

where $\mathcal{H}(\cdot)$ denotes the feature extractor of model $\mathcal{G}(\cdot)$. Since the similarity metric is not our crucial research point in this study, we just choose the L_2 loss for simplicity. To sum up, the BDFLoss is made up of these two components within the hyper-parameter λ:

$$\mathcal{L}_{\text{BDF}} = \mathcal{L}_{\text{count}} + \lambda\mathcal{L}_{\text{distill}}, \tag{11}$$

where λ is applicable as a trade-off between model plasticity and stability. It is the same as vanilla sequential fine-tuning when λ is equal to 0.

3.4 Model architectures

Our proposed domain-incremental self-distillation lifelong crowd counting benchmark is model-agnostic. To illustrate its effectiveness, we integrate it into several state-of-the-art crowd counting backbone models without the bells and whistles, CSRNet (Li YH et al., 2018), SFANet (Zhu L et al., 2019), DM-Count (Wang BY et al., 2020), and DKPNet (Chen BH et al., 2021). Because the attention map supervision of SFANet may introduce some biases in the experimental comparisons and the source code of DKPNet is not released, we make the following modifications in our experiments. A small improvement of SFANet is that we enable the network to learn the attention map adaptively based on training images without generating additional attention maps for supervision. We modify DKPNet-baseline in our experiments because we focus only on investigating the effectiveness of our proposed framework in forgetting and generalization under different model capacities.

4 Experiment settings

In this section, we will briefly introduce four datasets used in our experiments, the training settings, and some hyper-parameter selections.

4.1 Datasets

We train and evaluate our model in the public crowd counting datasets, i.e., ShanghaiTech PartA (Zhang YY et al., 2016), ShanghaiTech PartB (Zhang YY et al., 2016), UCF-QNRF (Idrees et al., 2018), NWPU-Crowd (Wang Q et al., 2021), and JHU-Crowd++ (Sindagi et al., 2019) (Table 2). To illustrate the generalization of different training paradigms, we have to select one of them as the unseen dataset that could never be trained during the domain-incremental lifelong learning process. In our experiments, we take the JHU-Crowd++ dataset as an unseen one because it has a variety of diverse scenarios and unconstrained environmental conditions (Sindagi et al., 2019). The synthetic dataset GCC (Wang Q et al., 2019) is also used to analyze the synthetic-to-real generalization performance under the lifelong crowd counting settings.

4.2 Implementation details

We strictly follow the same basic image preprocessing settings as in most recent literature (Li YH et al., 2018; Ma et al., 2019; Zhu L et al., 2019; Wang

BY et al., 2020). The crop size is 256×256 for SHA, and 512×512 for SHB, QNRF, and NWPU datasets. To generate the density map as ground truth, we just adopt the fixed Gaussian kernel whose variance σ is set to 15 for all datasets. Several useful augmentations like random horizontal flipping with a probability of 0.5 and normalization are applied to those images before training. The hyper-parameter λ in the loss function is set to 0.5 to achieve a trade-off between model plasticity and stability. We use the fixed learning rate of 1×10^{-5}, a simple weight decay of 5×10^{-4}, and an Adam optimizer in all our experiments. We use the PyTorch framework and NVIDIA GeForce RTX 3090 GPU workstation.

4.3 Evaluation metrics

The catastrophic forgetting phenomenon often exists in domain-incremental learning. To evaluate how much old knowledge on earth the model forgets in the previous domains and make a fair comparison with other methods, we propose a new metric,

Table 2 The number of images used to train models on different datasets

Dataset	Number of raw/training samples	Number of testing samples	Number of persons per image		
			Minimum	Maximum	Average
ShanghaiTech PartA	300/300	182	33	3139	501
ShanghaiTech PartB	400/400	316	9	578	123
UCF-QNRF	1201/1201	334	49	12 865	815
NWPU-Crowd	3609/3609	1500	0	20 033	418
JHU-Crowd++	2772/0	1600	0	25 791	346
GCC	15 212*		0	3995	501

* The total number of training and testing samples

Table 3 The results with different domain-incremental lifelong learning methods

Model	MAE					RMSE					mMAE	mRMSE
	SHA	QNRF	SHB	NWPU	JHU (unseen)	SHA	QNRF	SHB	NWPU	JHU (unseen)		
LwF* (Li ZZ and Hoiem, 2018)	62.3	81.4	11.5	90.8	90.4	104.4	133.4	18.2	395.2	298.2	61.5	162.8
EwC* (Kirkpatrick et al., 2017)	64.9	88.5	10.2	84.2	85.9	117.2	171.7	17.6	377.7	294.1	62.0	171.1
FLCB (Ours)	68.8	84.3	7.8	76.6	**84.8**	113.9	160.1	12.2	364.2	**264.8**	**59.4**	**162.6**

* represents our reproduced results of modified approaches. The bold number indicates the best performance

Table 4 Quantitative results with different paradigms to compare the forgetting degree and overall performance

Model	Method	MAE				RMSE				mMAE	mRMSE	nBwT	#params. ($\times 10^7$)	MACs ($\times 10^{10}$)
		SHA	QNRF	SHB	NWPU	SHA	QNRF	SHB	NWPU					
CSRNet (Li YH et al., 2018)	BASELINE	98.4	123.9	13.4	114.5	168.1	225.3	19.1	456.5	87.6	217.3	0.424	1.626	2.707
	LwF*	71.5	107.4	11.3	123.3	122.4	198.9	16.7	520.3	**78.4**	214.6	−0.042		
	FLCB	66.6	112.5	13.0	121.4	100.4	198.6	22.0	473.2	**78.4**	**198.6**	**−0.102**		
	JOINT	64.0	109.0	14.0	124.8	100.6	199.7	18.6	499.4	78.0	204.6	–		
SFANet (Zhu L et al., 2019)	BASELINE	85.4	112.6	14.8	106.9	141.3	200.7	18.1	463.7	79.9	206.0	0.545	1.702	2.728
	LwF*	75.0	101.3	11.5	108.3	128.5	177.2	19.0	450.0	74.0	193.7	−0.002		
	FLCB	69.4	103.7	12.7	108.8	110.9	176.6	20.9	445.0	**73.7**	**188.4**	**−0.097**		
	JOINT	77.7	136.8	14.0	127.8	124.0	236.3	17.3	458.5	89.1	209.0	–		
DM-Count (Wang BY et al., 2020)	BASELINE	76.0	94.1	9.6	108.3	122.2	154.1	17.5	481.4	72.0	193.8	0.176	2.150	2.699
	LwF*	74.6	90.2	9.4	86.9	124.1	164.9	14.9	375.4	65.3	169.8	0.049		
	FLCB	69.2	95.4	9.7	83.6	113.2	166.0	15.6	370.8	**64.5**	**166.4**	**−0.013**		
	JOINT	78.2	86.7	7.9	88.5	129.3	153.3	13.0	393.8	65.3	172.4	–		
DKPNet (Chen BH et al., 2021)	BASELINE	92.9	100.1	7.7	90.0	157.8	179.4	12.4	393.6	72.7	185.8	0.371	1.328	1.038
	LwF*	62.3	81.4	11.5	104.4	133.4	18.2	90.8	395.2	61.5	162.8	−0.009		
	FLCB	68.8	84.3	7.8	76.6	113.9	160.1	12.2	364.2	**59.4**	**162.6**	**−0.010**		
	JOINT	65.0	86.0	8.4	81.2	108.5	163.3	13.2	357.7	60.2	160.7	–		

We take sequential training as our BASELINE and joint training as JOINT for reference. FLCB is our proposed method.
* represents our reproduced results of modified approaches. The bold number indicates the best performance among the lifelong learning methods

called normalized Backward Transfer (nBwT). With the help of nBwT, the total forgetfulness over t incremental domains could be measured to determine whether the model is equipped with the sustainable learning ability. The normalization operation we introduce in nBwT could eliminate the potential negative impact because of the different learning difficulties in different domains.

$$\mathrm{nBwT}_t = \frac{1}{t-1}\sum_{i=1}^{t-1}\frac{e_{t,i}-e_{i,i}}{e_{i,i}}\ ,\ t=2,3,...,\mathcal{N},\quad (12)$$

where $e_{t,i}$ is the test MAE score of the i^{th} dataset when obtaining the optimal model on the t^{th} dataset, and $i<t$. nBwT_t is the accumulation of the forgetting performance among all previous $t-1$ domain datasets. The non-zero divisor $e_{i,i}$ is a normalization factor. The larger the nBwT value is, the greater the model forgetting degree is. A value smaller than 0 indicates that the model has attained a positive performance improvement among previously trained datasets. The theoretical lower bound of nBwT_t is $-\frac{1}{t-1}$ when $e_{t,i}$ equals zero.

Furthermore, we propose two reasonable and impartial criteria, i.e., mMAE and mRMSE, the respective means of MAE and root mean square error (RMSE) in $\mathcal{N}$ datasets, to evaluate roughly the overall counting precision of the lifelong crowd counting task:

$$\mathrm{mMAE}=\frac{1}{\mathcal{N}}\sum_{i=1}^{\mathcal{N}}\frac{1}{\mathcal{M}_i}\sum_{j=1}^{\mathcal{M}_i}|\hat{Y_j}-Y_j|,\quad (13)$$

$$\mathrm{mRMSE}=\frac{1}{\mathcal{N}}\sum_{i=1}^{\mathcal{N}}\sqrt{\frac{1}{\mathcal{M}_i}\sum_{j=1}^{\mathcal{M}_i}||\hat{Y_j}-Y_j||^2},\quad (14)$$

where $\mathcal{M}_i$ denotes the number of images from the i^{th} test set. $\hat{Y_j}$ and Y_j are the predicted count and actual count of the j^{th} image, respectively. mMAE and mRMSE reduce to standard MAE and RMSE respectively when $\mathcal{N}$ is equal to 1.

In addition, we still use the standard MAE score on the unseen JHU-Crowd++ dataset to compare the model generalization within different training strategies.

5 Experimental results

In this section, we first evaluate the overall performance and generalization ability of our proposed FLCB framework by comparison with two classical continual learning approaches (Kirkpatrick et al., 2017; Li ZZ and Hoiem, 2018) (Table 3). Then, we demonstrate the difference between FLCB and three other learning strategies, especially for analyzing their respective forgetting degrees among the trained datasets (SHA, SHB, QNRF, and NWPU), and their generalization abilities on the unseen dataset (JHU-Crowd++). The synthetic-to-real experiments are also conducted considering the data privacy issues and some ethical policies.

Table 5 Forgetting performance in the intermediate process of lifelong crowd counting among four models with FLCB

Method (FLCB)	Model	MAE				RMSE				mMAE	mRMSE	nBwT
		SHA	QNRF	SHB	NWPU	SHA	QNRF	SHB	NWPU			
SHA→QNRF	CSRNet	73.9	121.8	–	–	111.7	225.3	–	–	97.9	168.5	0.068
SHA→QNRF	SFANet	73.4	111.3	–	–	114.4	200.4	–	–	92.4	157.4	0.225
SHA→QNRF	DM-Count	65.2	84.8	–	–	117.2	149.0	–	–	75.0	133.1	**0.058**
SHA→QNRF	DKPNet	62.1	82.9	–	–	103.9	149.7	–	–	**72.5**	**126.8**	0.078
SHA→QNRF→SHB	CSRNet	73.9	121.8	16.1	–	111.7	225.3	29.9	–	70.6	122.3	0.034
SHA→QNRF→SHB	SFANet	73.4	111.3	20.5	–	114.4	200.4	31.5	–	68.4	115.4	0.113
SHA→QNRF→SHB	DM-Count	65.2	84.8	13.6	–	117.2	149.0	25.6	–	54.3	97.3	0.029
SHA→QNRF→SHB	DKPNet	63.5	86.4	10.3	–	109.6	147.5	17.3	–	**53.4**	**91.5**	**−0.014**
SHA→QNRF→SHB→NWPU	CSRNet	66.6	112.5	13.0	121.4	100.4	198.6	22.0	473.2	78.4	198.6	**−0.102**
SHA→QNRF→SHB→NWPU	SFANet	69.4	103.7	12.7	108.8	110.9	176.6	20.9	445.0	73.7	188.4	**−0.097**
SHA→QNRF→SHB→NWPU	DM-Count	69.2	95.4	9.7	83.6	113.2	166.0	15.6	370.8	64.5	166.4	**−0.013**
SHA→QNRF→SHB→NWPU	DKPNet	68.8	84.3	7.8	76.6	113.9	160.1	12.2	364.2	**59.4**	**162.6**	**−0.010**

The data underlined are all less than zero, which means that our proposed FLCB method has a positive effect on the overall performance of past domains

5.1 Analysis of catastrophic forgetting

As shown in Table 3, we reproduce two of the classical lifelong learning methods and modify them to adapt to our crowd counting task, because most lifelong learning methods focus on the classification task, while crowd counting is a regression-like task. The average performances in past domains and unseen domains of our proposed FLCB method all surpass those of LwF and EwC approaches. We compare the quantitative results between the baselines and our proposed method based on four benchmark models. The results in Table 4 demonstrate that our method can remarkably alleviate the catastrophic forgetting phenomenon on all models with the lowest mMAE, mRMSE, and nBwT (i.e., forgetting degree) under the domain-incremental training settings. We also report the model parameters and the Multiply-ACcumulate operations (MACs) for each benchmark model. The forgetting degree in the intermediate process is detailed in Table 5. The results imply that the model will forget less and count better when more labeled datasets are involved in the lifelong learning process. This indicates that our framework can remember the old yet meaningful knowledge from the last well-trained model when handling the new domain dataset.

5.2 Effect of hyper-parameter λ

The proposed balanced domain forgetting loss (BDFLoss) is composed of optimal transport counting loss and self-distillation loss. The hyper-parameter λ plays a dominant role in our proposed BDFLoss to control how much previously learned meaningful knowledge should be retrained when learning on new domain data. In other words, λ is a trade-off between model plasticity and stability. The greater the value of λ is, the more attention should be paid to leveraging the distilled knowledge. If λ is equal to 0, it degenerates to the vanilla sequential training without any constraint of previous knowledge. We just empirically choose $\lambda = 0.5$ to conduct our main experiments in this study. In this subsection, we also investigate whether different λ values will have a visible effect on forgetting. The extensive results demonstrate that $\lambda = 0.5$ is a reasonable choice (Table 6).

5.3 Analysis of model generalization

5.3.1 Real-to-real generalization

To build a robust model for better crowd counting, we expect that the model can obtain acceptable performance among unseen domains, because

Table 6 Forgetting degree comparison results with different hyper-parameters λ's

Method (FLCB)	λ	MAE				RMSE				nBwT
		SHA	QNRF	SHB	NWPU	SHA	QNRF	SHB	NWPU	
SHA→QNRF	0.1	62.2	77.2	–	–	104.7	137.5	–	–	0.080
SHA→QNRF	0.5	62.1	82.9	–	–	103.9	149.7	–	–	**0.078**
SHA→QNRF	1.0	62.5	81.2	–	–	108.4	145.3	–	–	0.085
SHA→QNRF→SHB	0.1	62.2	77.2	11.0	–	104.7	137.5	19.8	–	**0.040**
SHA→QNRF→SHB	0.5	63.5	86.4	10.3	–	109.6	147.5	17.3	–	0.072
SHA→QNRF→SHB	1.0	62.5	81.2	10.7	–	108.4	145.3	20.1	–	0.043
SHA→QNRF→SHB→NWPU	0.1	65.5	92.5	8.7	84.4	111.4	181.8	14.7	410.1	0.042
SHA→QNRF→SHB→NWPU	0.5	68.8	84.3	7.8	76.6	113.9	160.1	12.2	364.2	**−0.010**
SHA→QNRF→SHB→NWPU	1.0	67.0	84.8	11.0	80.0	112.4	181.1	18.3	354.9	0.079

The model used is DKPNet. The bold number indicates the lowest forgetting degree

Table 7 Generalization comparison of different training strategies on the unseen JHU-Crowd++ dataset

Model	MAE				RMSE			
	CSRNet	SFANet	DM-Count	DKPNet	CSRNet	SFANet	DM-Count	DKPNet
JOINT	103.2	115.5	96.3	89.8	320.0	347.6	320.3	318.7
LwF*	101.6	107.7	94.6	90.4	322.3	312.3	**296.0**	298.2
FLCB	**92.9**	**102.2**	**82.5**	**84.8**	**305.1**	**311.3**	298.5	**264.8**

* represents our reproduced results of modified approaches. The bold number indicates the best performance

labeling crowd images is extremely expensive and time-consuming in the real world. After the ultimate models converge, we test them directly on the unseen JHU-Crowd++ dataset (Table 7). Note that the images from JHU-Crowd++ are never trained during the process of lifelong learning. Our proposed FLCB can achieve lower prediction errors in terms of MAE and RMSE over the unseen dataset, indicating a stronger generalization ability compared with the joint training strategy. Furthermore, taking DKP-Net as an example, we delve into the ablation study of different layers for distillation in the intermediate lifelong learning process. Every time the training of a new incoming dataset is finished, the model will be evaluated on the unseen dataset. The results, shown in Table 8, illustrate that its performance is boosted progressively with incremental data from different domains. It is also indicated that the model can count better on the unseen domain under the mutually complementary interaction of both feature- and output-level distillation. Training in different orders may achieve fluctuating performance in unseen domains. We present the results in the supplementary materials because they could be related to curriculum learning, which is not our main focus in this study.

5.3.2 Synthetic-to-real generalization

Considering data privacy and some ethical policies (i.e., the real-world training images may be unobtainable), we conduct the training with the same lifelong settings on the synthetic crowd dataset (GCC) (Wang Q et al., 2019) and investigate the generalization on the unseen real-world dataset (ShanghaiTech PartB). The GCC dataset is collected from the GTA5 game environment, containing 15 212 synthetic images with diverse scenes. The synthetic dataset can provide precise but not time-consuming annotations for training. We split the GCC synthetic dataset into four subsets to mock the same lifelong training settings. The forgetting phenomenon among incremental synthetic subsets is still analyzed (Table 9), as well as the generalization performance on the unseen dataset. After obtaining the ultimate model, our FLCB benchmark achieves the lowest mMAE, mRMSE, and nBwT among previously seen datasets and decent performance on the unseen real-world dataset. Furthermore, the generalization experimental results (Table 10) verify the superiority of our proposed benchmark.

Table 10 The test MAE and RMSE scores on the unseen ShanghaiTech PartB dataset after training synthetic GCC subsets

Method	MAE	RMSE
JOINT	22.8	30.6
CycleGAN (Zhu JY et al., 2017)	25.4	39.7
SE CycleGAN (Wang Q et al., 2019)	19.9	28.3
FLCB	**16.1**	**25.0**

The bold number indicates the best performance

In summary, our proposed lifelong crowd count-

Table 8 Generalization comparison on the unseen JHU-Crowd++ dataset with self-distillation at different levels during the entire lifelong learning process

Distillation		MAE			RMSE		
Feature	Output	A→Q	A→Q→B	A→Q→B→N	A→Q	A→Q→B	A→Q→B→N
✓		102.6	93.2	87.1	341.0	324.4	298.5
	✓	106.7	102.3	90.4	345.9	354.8	298.2
✓	✓	**96.2**	**90.5**	**84.8**	**327.8**	**313.0**	**264.8**

A, Q, B, and N are the abbreviations for the names of four datasets SHA, QNRF, SHB, NWPU, respectively. The bold number indicates the best performance

Table 9 Experimental results of DKPNet with the synthetic-to-real training settings

Method	MAE				RMSE				mMAE	mRMSE	nBwT
	GCC-1	GCC-2	GCC-3	GCC-4	GCC-1	GCC-2	GCC-3	GCC-4			
BASELINE	55.4	34.7	18.5	35.6	131.3	82.8	53.3	74.9	36.1	85.6	1.130
LwF*	42.8	37.7	16.5	35.1	104.5	108.5	43.1	70.6	33.0	81.7	0.378
FLCB	40.0	35.1	14.6	41.7	95.4	100.5	34.5	82.5	**32.8**	**78.2**	**0.192**

* represents our reproduced results of modified approaches. The bold number indicates the best performance

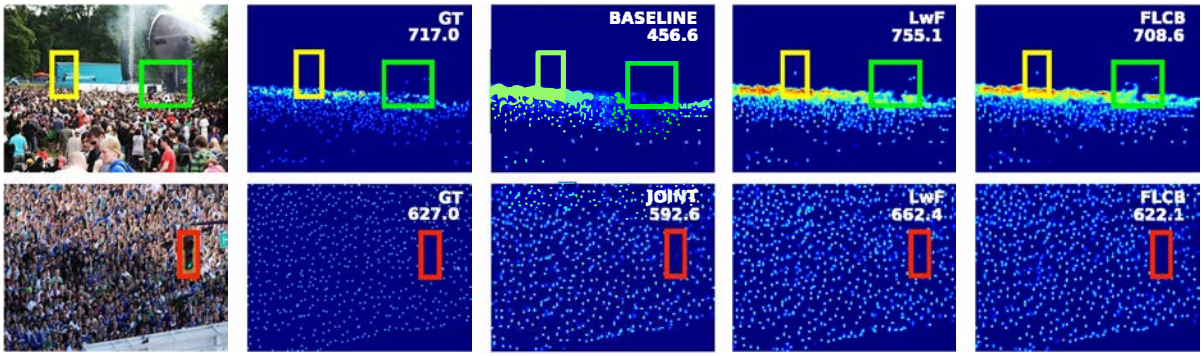

Fig. 3 The visualization results of different training paradigms. The top row shows the predictions and compares the forgetting degree on the first training dataset (SHA), while the bottom row illustrates the predictions and compares the generalization ability on the unseen dataset (JHU) (red: FLCB can correctly discriminate the non-human objects like traffic lights; green: FLCB may be affected by background noise such as loudspeakers; yellow: FLCB may not handle well the missing annotations, which is not the key research point in our work). References to color refer to the online version of this figure

ing benchmark FLCB can help the crowd counters forget less and count better to sustainably handle multiple-domain crowd counting using a single model, which indicates that it has potential to tackle more complicated scenes in the future.

5.4 Visualization results

To make a more qualitative comparison, we visualize the prediction density maps under different training strategies. As illustrated in Fig. 3, we discover that the sequential training methods achieve terrible performance among old domains after training images from a new domain. Our proposed lifelong crowd counting benchmark can estimate crowd density on both seen and unseen datasets more accurately and outperforms other training paradigms.

5.5 Discussions

5.5.1 Limitations

In this paper, we attempt to develop a single model to handle the incremental datasets from different domains for better lifelong crowd counting. Judging from both quantitative and qualitative results, our proposed FLCB does well in achieving a trade-off performance from all domain datasets compared with other methods. However, there are still some limitations that may drive future research directions in lifelong crowd counting. On one hand, according to the visualization results, our proposed FLCB method seems to have difficulty in dealing with the missing annotations (yellow bounding boxes) and background noises (green bounding boxes), like the loudspeaker box in Fig. 3. On the other hand, we do not integrate any replay-based strategies into our experiments considering the training time and storage overhead. Efficient data sampling strategies and replay-based approaches may boost lifelong crowd counting, which deserves to be investigated in the future.

5.5.2 Lifelong learning vs. self-supervised learning

We would like to discuss lifelong learning and self-supervised learning from a pretraining perspective. They share something in common that is expected to lay the foundation for artificial general intelligence. Recent literature (Caron et al., 2020; Chen T et al., 2020; Grill et al., 2020; He KM et al., 2020; Niu et al., 2020, 2022; Huang et al., 2022; Niu and Wang, 2022a, 2022b) shows the power of self-supervised learning as a novel pretraining paradigm to empower multiple downstream tasks. To an extent, lifelong learning could be regarded as a kind of pretraining method, because it learns the shared knowledge and general representations to boost performance. However, lifelong learning usually requires labeled data for training to enhance model capacity, whereas self-supervised learning does not. From our perspectives, both types of learning could provide a good pretrained network or initialization for the training of other domain datasets or downstream tasks, and lifelong learning may empower self-supervised learning in the future.

6 Conclusions

We propose a domain-incremental self-distillation learning benchmark for lifelong crowd counting to deal with the catastrophic forgetting and model generalization issues using a single model when training new datasets from different domains one after another. With the help of the BDFLoss function that we have designed, the model can forget less and count better during the entire lifelong crowd counting process. Additionally, our proposed metric nBwT can be used to measure the forgetting degree in future lifelong crowd counting models. Extensive experiments demonstrate that our proposed benchmark has a lower forgetting degree over the sequential training baseline and a stronger generalization ability compared with the joint training strategy. Our proposed method is a simple yet effective way to sustainably handle the crowd counting problem among multiple domains using a single model with limited storage overhead when the newly available domain data arrive. It can be incorporated into any existing backbone as a plug-and-play training strategy for better crowd counting in the real world. Although our work considers crowd counting, the proposed framework has the potential to be applied in other regression-related image or video tasks.

Contributors

Jiaqi GAO designed the research and drafted the paper. Jingqi LI contributed ideas for experiments and analysis. Jingqi LI, Hongming SHAN, Yanyun QU, James Z. WANG, Fei-Yue WANG, and Junping ZHANG helped organize and revised the paper. Jiaqi GAO, Hongming SHAN, and Junping ZHANG finalized the paper.

Compliance with ethics guidelines

Jiaqi GAO, Jingqi LI, Hongming SHAN, Yanyun QU, James Z. WANG, Fei-Yue WANG, and Junping ZHANG declare that they have no conflict of interest.

Data availability

The data that support the findings of this study are available from the corresponding authors upon reasonable request.

References

Bai S, He ZQ, Qiao Y, et al., 2020. Adaptive dilated network with self-correction supervision for counting. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.4594-4603. https://doi.org/10.1109/CVPR42600.2020.00465

Belouadah E, Popescu A, 2019. IL2M: class incremental learning with dual memory. Proc IEEE/CVF Int Conf on Computer Vision, p.583-592. https://doi.org/10.1109/ICCV.2019.00067

Boominathan L, Kruthiventi SSS, Babu RV, 2016. CrowdNet: a deep convolutional network for dense crowd counting. Proc 24th ACM Int Conf on Multimedia, p.640-644. https://doi.org/10.1145/2964284.2967300

Cao XK, Wang ZP, Zhao YY, et al., 2018. Scale aggregation network for accurate and efficient crowd counting. Proc 15th European Conf on Computer Vision, p.734-750. https://doi.org/10.1007/978-3-030-01228-1_45

Caron M, Misra I, Mairal J, et al., 2020. Unsupervised learning of visual features by contrasting cluster assignments. Proc 34th Int Conf on Neural Information Processing Systems, p.9912-9924.

Chan AB, Vasconcelos N, 2009. Bayesian Poisson regression for crowd counting. Proc 12th IEEE Int Conf on Computer Vision, p.545-551. https://doi.org/10.1109/ICCV.2009.5459191

Chen BH, Yan ZY, Li K, et al., 2021. Variational attention: propagating domain-specific knowledge for multi-domain learning in crowd counting. Proc IEEE/CVF Int Conf on Computer Vision, p.16065-16075. https://doi.org/10.1109/ICCV48922.2021.01576

Chen T, Kornblith S, Norouzi M, et al., 2020. A simple framework for contrastive learning of visual representations. Proc 37th Int Conf on Machine Learning, p.1597-1607.

Chen XY, Bin YR, Sang N, et al., 2019. Scale pyramid network for crowd counting. Proc IEEE Winter Conf on Applications of Computer Vision, p.1941-1950. https://doi.org/10.1109/WACV.2019.00211

Dalal N, Triggs B, 2005. Histograms of oriented gradients for human detection. Proc IEEE Computer Society Conf on Computer Vision and Pattern Recognition, p.886-893. https://doi.org/10.1109/CVPR.2005.177

Dollar P, Wojek C, Schiele B, et al., 2012. Pedestrian detection: an evaluation of the state of the art. *IEEE Trans Patt Anal Mach Intell*, 34(4):743-761. https://doi.org/10.1109/TPAMI.2011.155

Grill JB, Strub F, Altché F, et al., 2020. Bootstrap your own latent a new approach to self-supervised learning. Proc 34th Int Conf on Neural Information Processing Systems, p.21271-21284.

Guo D, Li K, Zha ZJ, et al., 2019. DADNet: dilated-attention-deformable ConvNet for crowd counting. Proc 27th ACM Int Conf on Multimedia, p.1823-1832. https://doi.org/10.1145/3343031.3350881

Han T, Gao JY, Yuan Y, et al., 2020. Focus on semantic consistency for cross-domain crowd understanding. Proc IEEE Int Conf on Acoustics, Speech and Signal Processing, p.1848-1852. https://doi.org/10.1109/ICASSP40776.2020.9054768

He KM, Fan HQ, Wu YX, et al., 2020. Momentum contrast for unsupervised visual representation learning. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.9729-9738. https://doi.org/10.1109/CVPR42600.2020.00975

He YJ, Sick B, 2021. CLeaR: an adaptive continual learning framework for regression tasks. *AI Persp*, 3(1):2. https://doi.org/10.1186/S42467-021-00009-8

Huang ZZ, Chen J, Zhang JP, et al., 2022. Learning representation for clustering via prototype scattering and positive sampling. *IEEE Trans Patt Anal Mach Intell*, early access. https://doi.org/10.1109/TPAMI.2022.3216454

Idrees H, Tayyab M, Athrey K, et al., 2018. Composition loss for counting, density map estimation and localization in dense crowds. Proc 15th European Conf on Computer Vision, p.532-546. https://doi.org/10.1007/978-3-030-01216-8_33

Jiang SQ, Lu XB, Lei YJ, et al., 2020. Mask-aware networks for crowd counting. *IEEE Trans Circ Syst Video Technol*, 30(9):3119-3129. https://doi.org/10.1109/TCSVT.2019.2934989

Jiang XH, Zhang L, Xu ML, et al., 2020a. Attention scaling for crowd counting. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.4706-4715. https://doi.org/10.1109/CVPR42600.2020.00476

Jiang XH, Zhang L, Lv P, et al., 2020b. Learning multi-level density maps for crowd counting. *IEEE Trans Neur Netw Learn Syst*, 31(8):2705-2715. https://doi.org/10.1109/TNNLS.2019.2933920

Kirkpatrick J, Pascanu R, Rabinowitz N, et al., 2017. Overcoming catastrophic forgetting in neural networks. *PNAS*, 114(13):3521-3526. https://doi.org/10.1073/pnas.1611835114

Leibe B, Seemann E, Schiele B, 2005. Pedestrian detection in crowded scenes. Proc IEEE/CVF Computer Society Conf on Computer Vision and Pattern Recognition, p.878-885. https://doi.org/10.1109/CVPR.2005.272

Li YH, Zhang XF, Chen DM, 2018. CSRNet: dilated convolutional neural networks for understanding the highly congested scenes. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.1091-1100. https://doi.org/10.1109/CVPR.2018.00120

Li ZZ, Hoiem D, 2018. Learning without forgetting. *IEEE Trans Patt Anal Mach Intell*, 40(12):2935-2947. https://doi.org/10.1109/TPAMI.2017.2773081

Liu L, Lu H, Xiong HP, et al., 2020. Counting objects by blockwise classification. *IEEE Trans Circ Syst Video Technol*, 30(10):3513-3527. https://doi.org/10.1109/TCSVT.2019.2942970

Liu LB, Qiu ZL, Li GB, et al., 2019. Crowd counting with deep structured scale integration network. Proc IEEE/CVF Int Conf on Computer Vision, p.1774-1783. https://doi.org/10.1109/ICCV.2019.00186

Liu LB, Chen JQ, Wu HF, et al., 2021. Cross-modal collaborative representation learning and a large-scale RGBT benchmark for crowd counting. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.4823-4833. https://doi.org/10.1109/CVPR46437.2021.00479

Liu N, Long YC, Zou CQ, et al., 2019. ADCrowdNet: an attention-injective deformable convolutional network for crowd understanding. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.3225-3234. https://doi.org/10.1109/CVPR.2019.00334

Liu WZ, Salzmann M, Fua P, 2019. Context-aware crowd counting. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.5099-5108. https://doi.org/10.1109/CVPR.2019.00524

Liu WZ, Durasov N, Fua P, 2022. Leveraging self-supervision for cross-domain crowd counting. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.5341-5352. https://doi.org/10.1109/CVPR52688.2022.00527

Lopez-Paz D, Ranzato M, 2017. Gradient episodic memory for continual learning. Proc 31st Int Conf on Neural Information Processing Systems, p.6467-6476.

Lowe DG, 1999. Object recognition from local scale-invariant features. Proc 7th IEEE Int Conf on Computer Vision, p.1150-1157. https://doi.org/10.1109/ICCV.1999.790410

Luo A, Yang F, Li X, et al., 2020. Hybrid graph neural networks for crowd counting. Proc 34th AAAI Conf on Artificial Intelligence, p.11693-11700. https://doi.org/10.1609/aaai.v34i07.6839

Ma ZH, Wei X, Hong XP, et al., 2019. Bayesian loss for crowd count estimation with point supervision. Proc IEEE/CVF Int Conf on Computer Vision, p.6142-6151. https://doi.org/10.1109/ICCV.2019.00624

Ma ZH, Wei X, Hong XP, et al., 2020. Learning scales from points: a scale-aware probabilistic model for crowd counting. Proc 28th ACM Int Conf on Multimedia, p.220-228. https://doi.org/10.1145/3394171.3413642

Ma ZH, Hong XP, Wei X, et al., 2021. Towards a universal model for cross-dataset crowd counting. Proc IEEE/CVF Int Conf on Computer Vision, p.3205-3214. https://doi.org/10.1109/ICCV48922.2021.00319

Niu C, Wang G, 2022a. Self-supervised representation learning with MUlti-Segmental Informational Coding (MUSIC). https://arxiv.org/abs/2206.06461

Niu C, Wang G, 2022b. Unsupervised contrastive learning based transformer for lung nodule detection. *Phys Med Biol*, 67(20):204001. https://doi.org/10.1088/1361-6560/ac92ba

Niu C, Li MZ, Fan FL, et al., 2020. Suppression of correlated noise with similarity-based unsupervised deep learning. https://arxiv.org/abs/2011.03384

Niu C, Shan HM, Wang G, 2022. SPICE: semantic pseudo-labeling for image clustering. *IEEE Trans Image Process*, 31:7264-7278. https://doi.org/10.1109/TIP.2022.3221290

Rebuffi SA, Kolesnikov A, Sperl G, et al., 2017. iCaRL: incremental classifier and representation learning. Proc IEEE Conf on Computer Vision and Pattern Recognition, p.2001-2010. https://doi.org/10.1109/CVPR.2017.587

Rusu AA, Rabinowitz NC, Desjardins G, et al., 2016. Progressive neural networks. https://arxiv.org/abs/1606.04671

Sam DB, Surya S, Babu RV, 2017. Switching convolutional neural network for crowd counting. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.5744-5752. https://doi.org/10.1109/CVPR.2017.429

Shi ZL, Mettes P, Snoek C, 2019. Counting with focus for free. Proc IEEE/CVF Int Conf on Computer Vision, p.4200-4209. https://doi.org/10.1109/ICCV.2019.00430

Sindagi VA, Patel VM, 2017. Generating high-quality crowd density maps using contextual pyramid CNNs. Proc IEEE Int Conf on Computer Vision, p.1861-1870. https://doi.org/10.1109/ICCV.2017.206

Sindagi VA, Patel VM, 2020. HA-CCN: hierarchical attention-based crowd counting network. *IEEE Trans Image Process*, 29:323-335. https://doi.org/10.1109/TIP.2019.2928634

Sindagi V, Yasarla R, Patel V, 2019. Pushing the frontiers of unconstrained crowd counting: new dataset and benchmark method. Proc IEEE/CVF Int Conf on Computer Vision, p.1221-1231. https://doi.org/10.1109/ICCV.2019.00131

Song QY, Wang CA, Wang YB, et al., 2021. To choose or to fuse? Scale selection for crowd counting. Proc 35th AAAI Conf on Artificial Intelligence, p.2576-2583. https://doi.org/10.1609/aaai.v35i3.16360

Tan X, Tao C, Ren TW, et al., 2019. Crowd counting via multi-layer regression. Proc 27th ACM Int Conf on Multimedia, p.1907-1915. https://doi.org/10.1145/3343031.3350914

Tian YK, Lei YM, Zhang JP, et al., 2020. PaDNet: pan-density crowd counting. *IEEE Trans Image Process*, 29:2714-2727. https://doi.org/10.1109/TIP.2019.2952083

Tuzel O, Porikli F, Meer P, 2008. Pedestrian detection via classification on Riemannian manifolds. *IEEE Trans Patt Anal Mach Intell*, 30(10):1713-1727. https://doi.org/10.1109/TPAMI.2008.75

Wang BY, Liu HD, Samaras D, et al., 2020. Distribution matching for crowd counting. Proc 34th Int Conf on Neural Information Processing Systems, p.1595-1607.

Wang C, Zhang H, Yang L, et al., 2015. Deep people counting in extremely dense crowds. Proc 23rd ACM Int Conf on Multimedia, p.1299-1302. https://doi.org/10.1145/2733373.2806337

Wang Q, Gao JY, Lin W, et al., 2019. Learning from synthetic data for crowd counting in the wild. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.8198-8207. https://doi.org/10.1109/CVPR.2019.00839

Wang Q, Gao JY, Lin W, et al., 2021. NWPU-crowd: a large-scale benchmark for crowd counting and localization. *IEEE Trans Patt Anal Mach Intell*, 43(6):2141-2149. https://doi.org/10.1109/TPAMI.2020.3013269

Wang Q, Han T, Gao JY, et al., 2022. Neuron linear transformation: modeling the domain shift for crowd counting. *IEEE Trans Neur Netw Learn Syst*, 33(8):3238-3250. https://doi.org/10.1109/TNNLS.2021.3051371

Wu QQ, Wan J, Chan AB, 2021. Dynamic momentum adaptation for zero-shot cross-domain crowd counting. Proc 29th ACM Int Conf on Multimedia, p.658-666. https://doi.org/10.1145/3474085.3475230

Xiong HP, Lu H, Liu CX, et al., 2019. From open set to closed set: counting objects by spatial divide-and-conquer. Proc IEEE/CVF Int Conf on Computer Vision, p.8362-8371. https://doi.org/10.1109/ICCV.2019.00845

Yan ZY, Li PY, Wang B, et al., 2021. Towards learning multi-domain crowd counting. *IEEE Trans Circ Syst Video Technol*, early access. https://doi.org/10.1109/TCSVT.2021.3137593

Yang YF, Li GR, Wu Z, et al., 2020. Reverse perspective network for perspective-aware object counting. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.4374-4383. https://doi.org/10.1109/CVPR42600.2020.00443

Zhang C, Li HS, Wang XG, et al., 2015. Cross-scene crowd counting via deep convolutional neural networks. Proc IEEE Conf on Computer Vision and Pattern Recognition, p.833-841. https://doi.org/10.1109/CVPR.2015.7298684

Zhang Q, Lin W, Chan AB, 2021. Cross-view cross-scene multi-view crowd counting. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.557-567. https://doi.org/10.1109/CVPR46437.2021.00062

Zhang YY, Zhou DS, Chen SQ, et al., 2016. Single-image crowd counting via multi-column convolutional neural network. Proc IEEE Conf on Computer Vision and Pattern Recognition, p.589-597. https://doi.org/10.1109/CVPR.2016.70

Zhao MM, Zhang CY, Zhang J, et al., 2020. Scale-aware crowd counting via depth-embedded convolutional neural networks. *IEEE Trans Circ Syst Video Technol*, 30(10):3651-3662. https://doi.org/10.1109/TCSVT.2019.2943010

Zhu JY, Park T, Isola P, et al., 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. Proc IEEE Int Conf on Computer Vision, p.2223-2232. https://doi.org/10.1109/ICCV.2017.244

Zhu L, Zhao ZJ, Lu C, et al., 2019. Dual path multi-scale fusion networks with attention for crowd counting. https://arxiv.org/abs/1902.01115

Zou ZK, Qu XY, Zhou P, et al., 2021. Coarse to fine: domain adaptive crowd counting via adversarial scoring network. Proc 29th ACM Int Conf on Multimedia, p.2185-2194. https://doi.org/10.1145/3474085.3475377

List of supplementary materials

1 Domain concept and gaps of different datasets

2 Effect of different training orders

Fig. S1 Data distributions of four benchmark datasets

Table S1 Forgetting degree comparison results with different training orders

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