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. *Frontiers of Information Technology & Electronic Engineering*, 24(2):187-202. https://doi.org/10.1631/FITEE.2200380

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

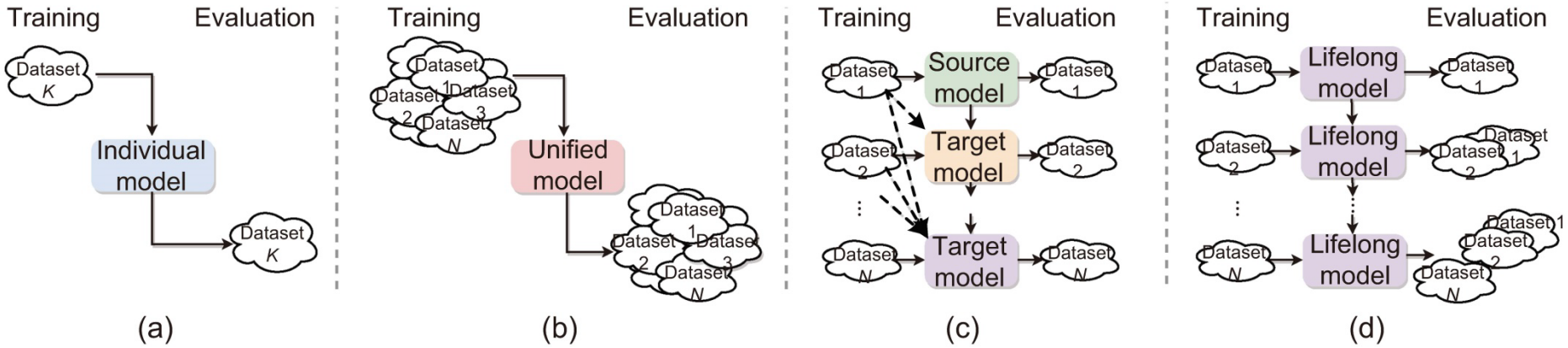
Key words: Crowd counting; Knowledge distillation; Lifelong learning

Corresponding author: Junping Zhang
E-mail: jpzhang@fdu.edu.cn
ORCID: https://orcid.org/0000-0002-5924-3360

Motivation

- A robust and practical crowd counting system is expected to have the ability of ***continuous learning*** with the ***newly incoming domain data in real-world scenarios***, instead of fitting one domain only.
- The well-trained model in a specific single domain achieves imperfect performance among other unseen domains, and the performance even ***degrades dramatically*** in previous seen domains due to domain shift.
- 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?

Motivation



Separate training (a) for each individual dataset may not generalize well to other datasets.

Joint training (b) needs linearly increasing storage overhead and training time to handle multiple domain crowd counting.

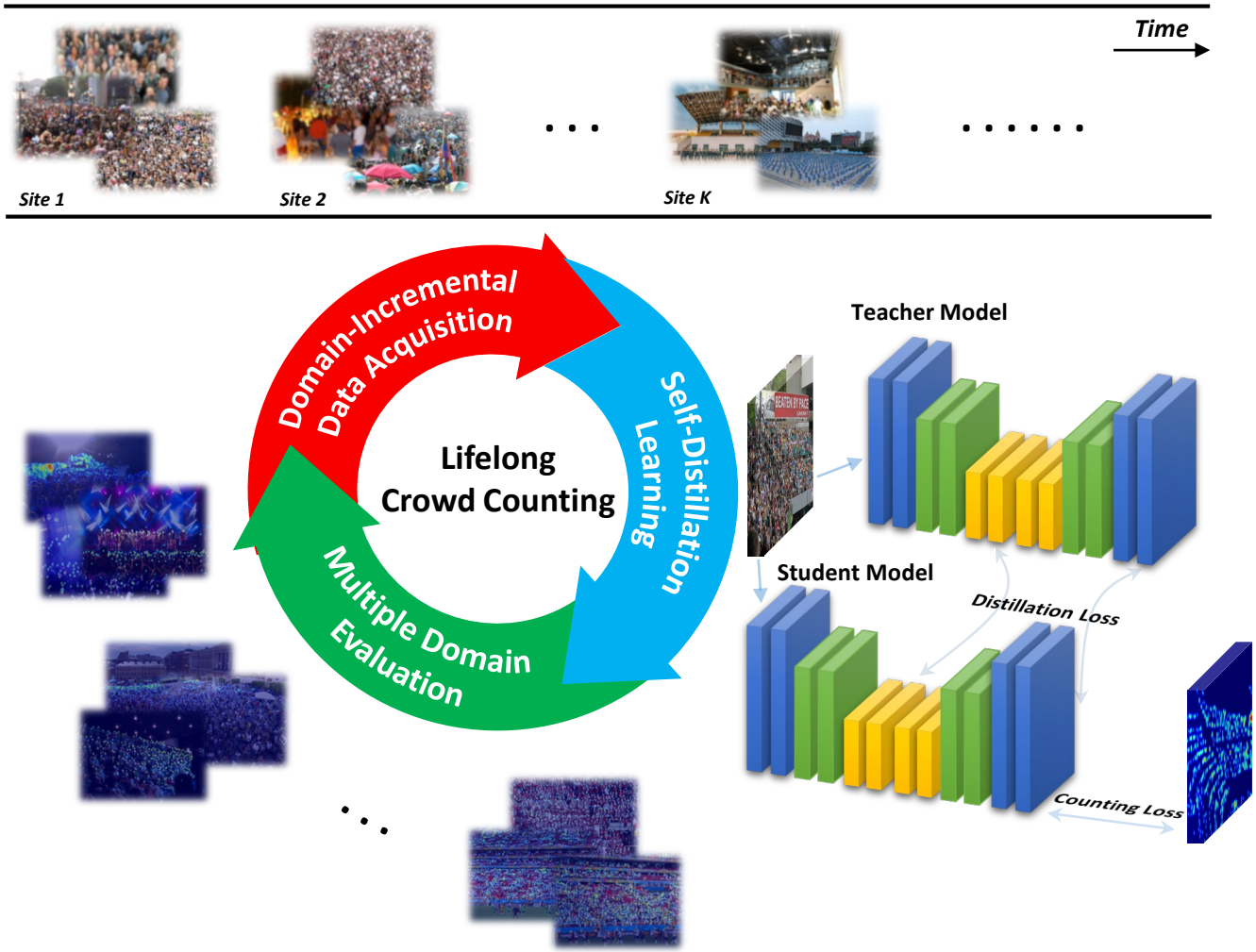
Sequential training (c) aims at improving the performance of the target domain, which may result in catastrophic forgetting among previous domains.

Lifelong learning (d) can sustainably improve the model performance in all domains, and is investigated in this paper (forget less and count better, FLCB).

Main idea

- We propose a domain-incremental self-distillation learning benchmark to handle the multiple domain crowd counting problem from the perspective of lifelong learning, investigating the catastrophic forgetting and generalization issues.
- We design a balanced domain forgetting loss function (BDFLoss) to prevent the model from dramatically forgetting the previous meaningful knowledge when being trained on the newly arriving data.
- We propose a new quantitative metric, normalized Backward Transfer (nBwT), to measure the forgetting degree in the whole lifelong crowd counting process.

Framework

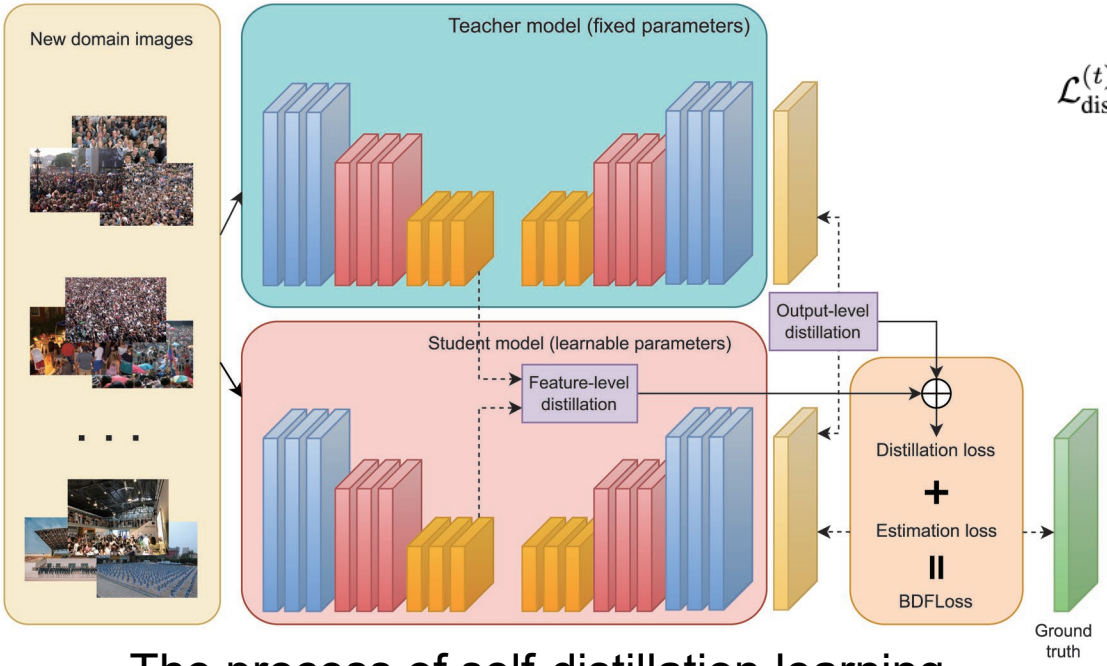


Overview of the domain-incremental self-distillation learning benchmark (forget less, count better, FLCB)

Method

The model trained at the last step is used as a teacher model, whose parameters are fixed to ***distill old meaningful knowledge*** to the current model.

- Self-distillation loss consists of ***feature-*** and ***output-level*** knowledge distillation.
- Counting loss is composed of basic L_1 loss, optimal transport loss, and the regularization item.



The process of self-distillation learning

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

$$\mathcal{L}_{\text{OT}}(\mathrm{Y},\hat{\mathrm{Y}}) = \mathcal{W}_c\left(\frac{\mathrm{Y}}{||\mathrm{Y}||_1},\frac{\hat{\mathrm{Y}}}{||\hat{\mathrm{Y}}||_1};\mathcal{C}\right) = \left\langle\alpha^*,\frac{\mathrm{Y}}{||\mathrm{Y}||_1}\right\rangle + \left\langle\beta^*,\frac{\hat{\mathrm{Y}}}{||\hat{\mathrm{Y}}||_1}\right\rangle$$

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

Evaluation metric

We propose **normalized Backward Transfer (nBwT)** to evaluate the forgetting degree of lifelong crowd counting models:

$$\mathrm{nBwT}_t = \frac{1}{t-1}\sum_{i=1}^{t-1}\frac{e_{t,i}-e_{i,i}}{e_{i,i}}\ ,\ t=2,\ldots,\mathcal{N}$$

- $e_{t,i}$ is the test estimation errors 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 zero 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.

Major results

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

Model	Method	MAE				RMSE				mMAE	mRMSE	nBwT	#params.	MACs
		SHA	QNRF	SHB	NWPU	SHA	QNRF	SHB	NWPU				($\times 10^7$)	($\times 10^{10}$)
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

Major results

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 has a positive effect on the overall performance of among past domains

Visualization results

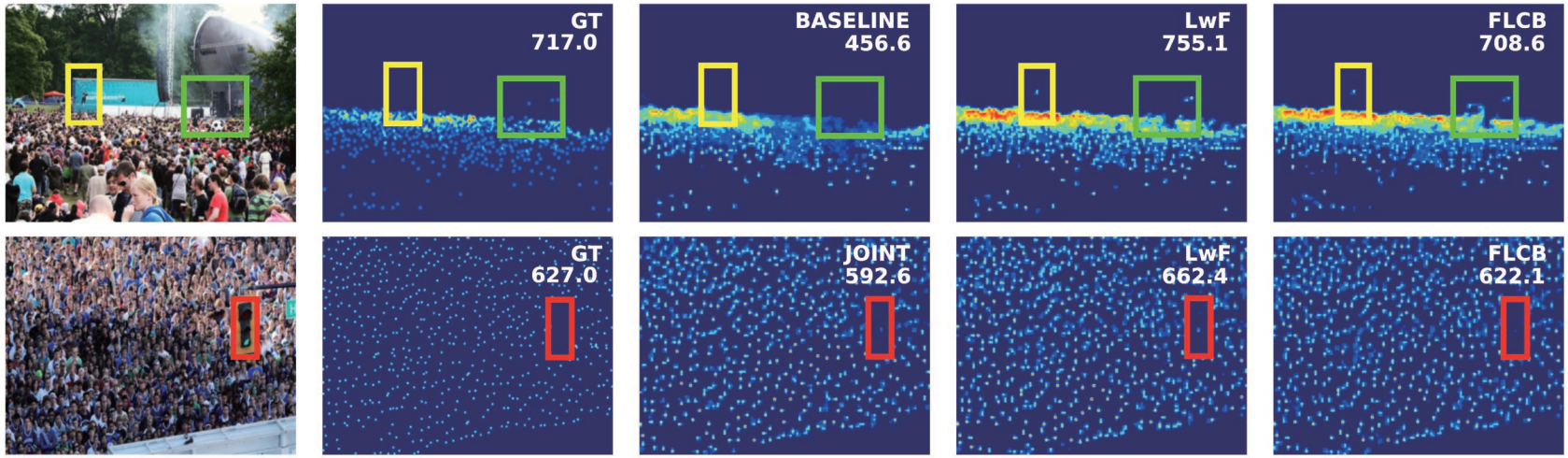


Fig. 3 The visualization results of different training paradigms. The top row shows the predictions and compares the forgetting degree on the first training dataset (SHA), while the bottom row illustrates the predictions and compares the generalization ability on the unseen dataset (JHU) (red: FLCB can 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

Conclusions

- The proposed FLCB method has a ***lower forgetting degree*** compared with sequential training and ***generalizes well among unseen data*** compared with the joint training strategy.
- With the help of ***BDFLoss*** that we have designed, the model can effectively forget less and count better during the entire lifelong crowd counting process.
- FLCB can be incorporated into any existing backbone as a ***plug-and-play training strategy*** for better crowd counting in the real world, and can serve as a promising benchmark for future lifelong crowd counting research.