CVPR SEATTLE
JUNE 14-19 2020 WASHINGTON

Semantic Drift Compensation for Class-Incremental Learning

Lu Yu[1,2], Bartłomiej Twardowski[1], Xialei Liu[1], Luis Herranz[1], Kai Wang[1],
Yongmei Cheng[2], Shangling Jui[3], Joost van de Weijer[1]
Computer Vision Center, Universitat Autonoma de Barcelona, Barcelona, Spain [1]
School of Automation, Northwestern Polytechnical University, Xi'an, China [2]
Huawei Kirin Solution, Shanghai, China [3]

Class-IL for Embeddings

Class-incremental learning (Class-IL) sequentially increases the number of classes to be classified. During training, the network has only access to data of one task at a time. The main problem of Class-IL is that networks suffer from **catastrophic forgetting** which is a drastic performance drop on previous tasks.

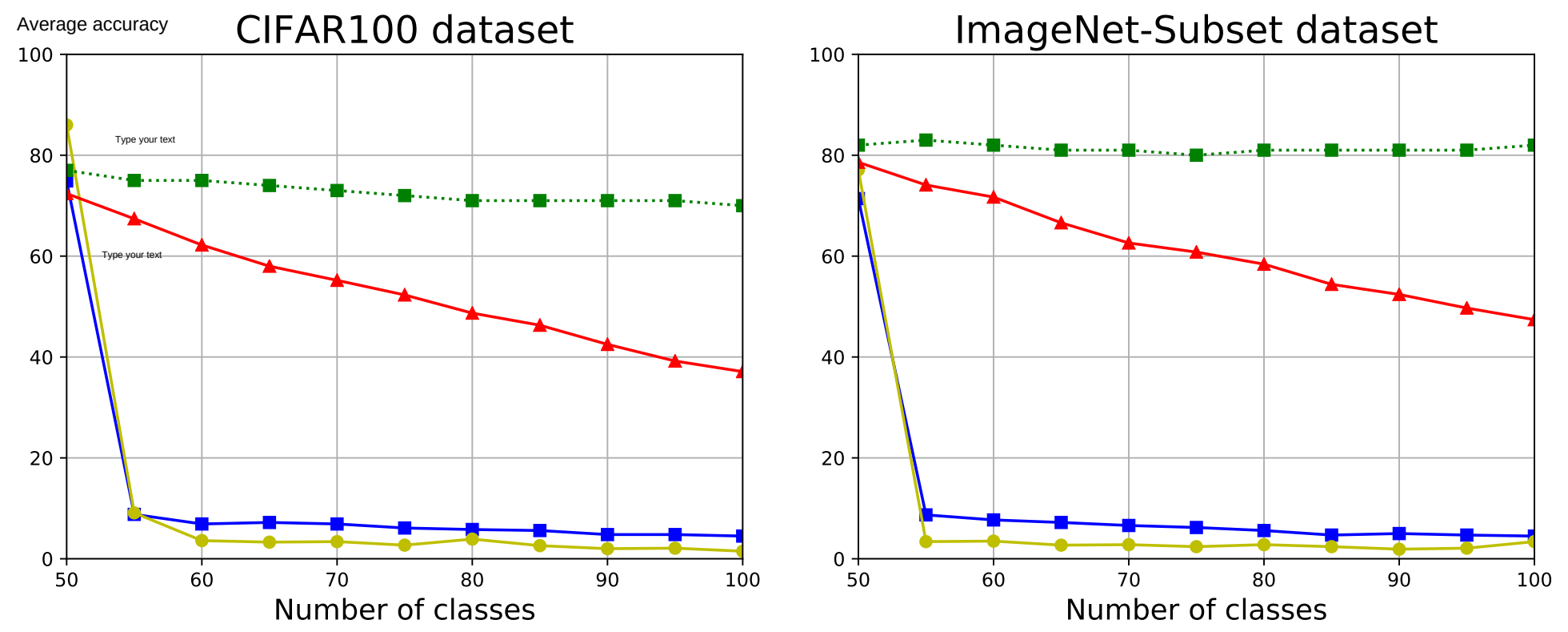
Contributions

- We show that embedding networks suffer significantly less from catastrophic forgetting than classification networks.
- Most existing literature on IL focuses on preventing drift, whereas our method aims to estimate the drift and compensate for it. Our method, SDC, can be combined with existing methods to improve results.
- We outperform existing non-exemplar methods and obtain competitive results compared to methods which store exemplars.

Softmax versus Embedding

Metric networks (trained with e.g. triplet loss) can be used to perform classification. Each class is represented by the class mean, also called *prototype*. Then classification is done by applying a Nearest Class Mean Classifier (NCM). The advantage of metric learning:

- New classes can be naturally added without any architectural changes for metric learning networks;

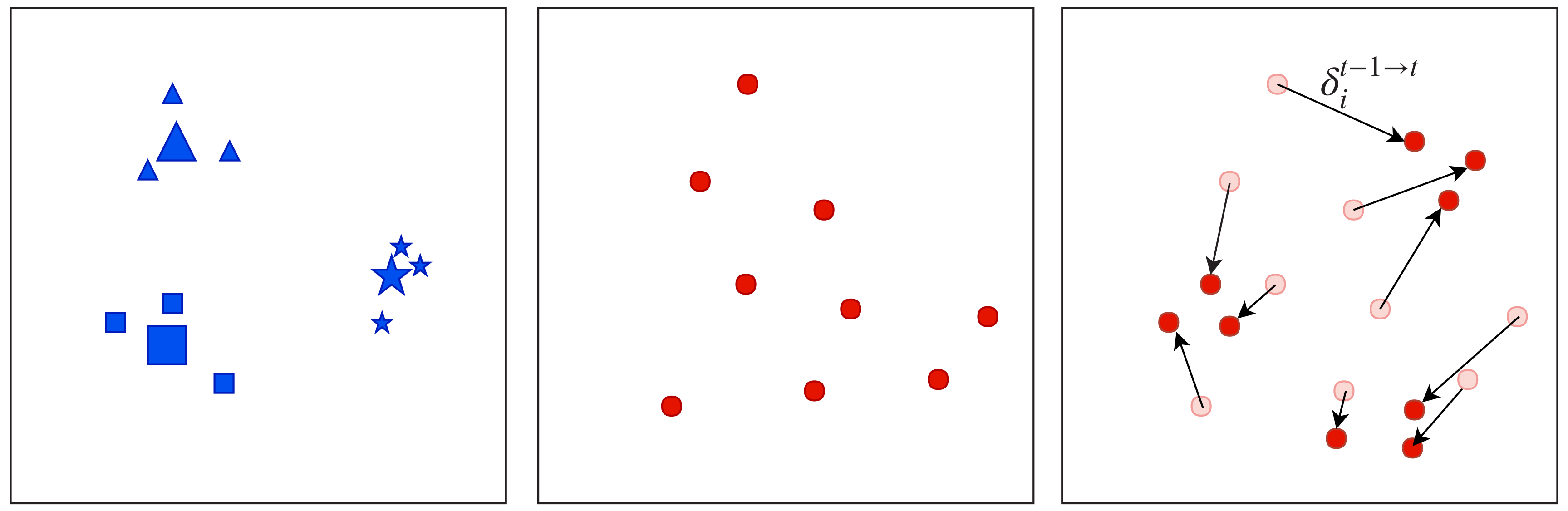


Methods incrementally trained with finetuning: joint training, classification network, embedding network.

Conclusion: Embeddings learned with metric losses suffer significantly less from catastrophic forgetting.

Semantic Drift Compensation

Illustration of SDC: (a) Data and prototypes of three classes of task 1 after training task 1. (b) Data of task 2 after training task 1. (c) Drift of data of task 2 while training task 2. This results in a sparse vector field of drift vectors. (d) This vector field is used to approximate the drift of the prototypes of task 1.



Computation of the Semantic Drift

- **Drift of the current data**

$$\boldsymbol{\delta}_i^{t-1\to t} = \mathbf{z}_i^t - \mathbf{z}_i^{t-1}\ ,\ y_i \in C^t, \tag{1}$$

- **Approximation of semantic drift**

$$\hat{\Delta}_{c^s}^{t-1\to t} = \frac{\sum_i \left[y_i \in C^t\right] w_i \boldsymbol{\delta}_i^{t-1\to t}}{\sum_i \left[y_i \in C^t\right] w_i}, \tag{2}$$

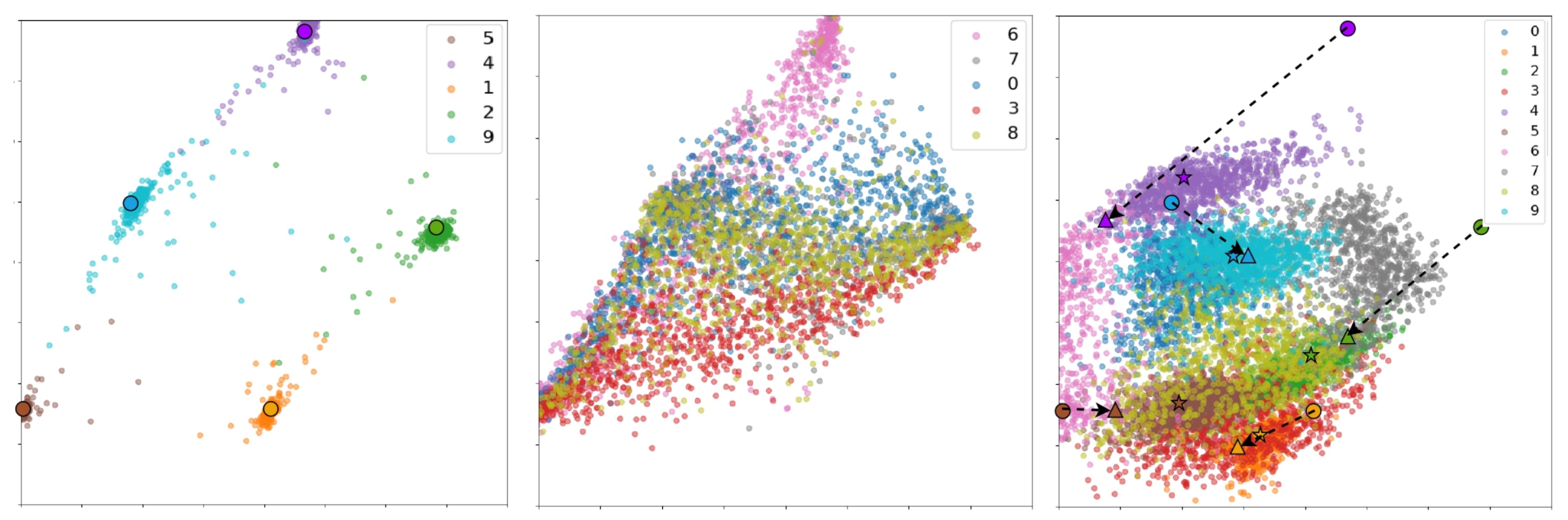
with

$$w_i = e^{-\frac{\left\|\mathbf{z}_i^{t-1} - \boldsymbol{\mu}_{c^s}^{t-1}\right\|^2}{2\sigma^2}}] \tag{3}$$

For all data points in task t we monitor the semantic drift during the training of task t. This is then used to compute the semantic drift of all previously learned prototypes by assigning a weight to the drift vectors according to their distance to the prototypes.

Visualization on MNIST

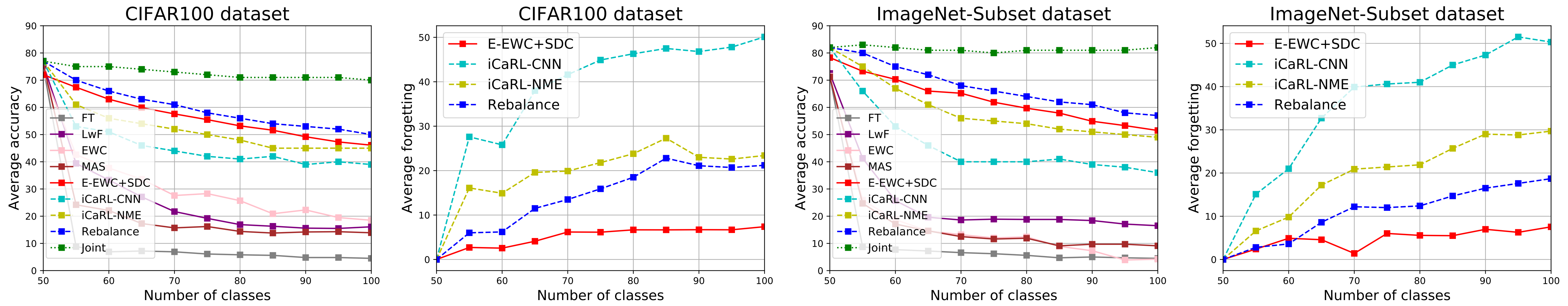
SDC with E-FT: the saved prototypes of the previous task(indicated by circle) are corrected by SDC to new positions(indicated by triangle).



We can see that the approximated drift vectors improve the locations of the prototypes to be closer to the correct positions (indicated by star). As a result, the accuracy of the overall method remains higher while training new tasks.

Experiments on CIFAR100 and ImageNet-Subset

Comparison of average incremental accuracy and average forgetting with eleven-task setting on CIFAR100 and ImageNet-Subset dataset. Our E-EWC+SDC beats all the other non-exemplar based methods.



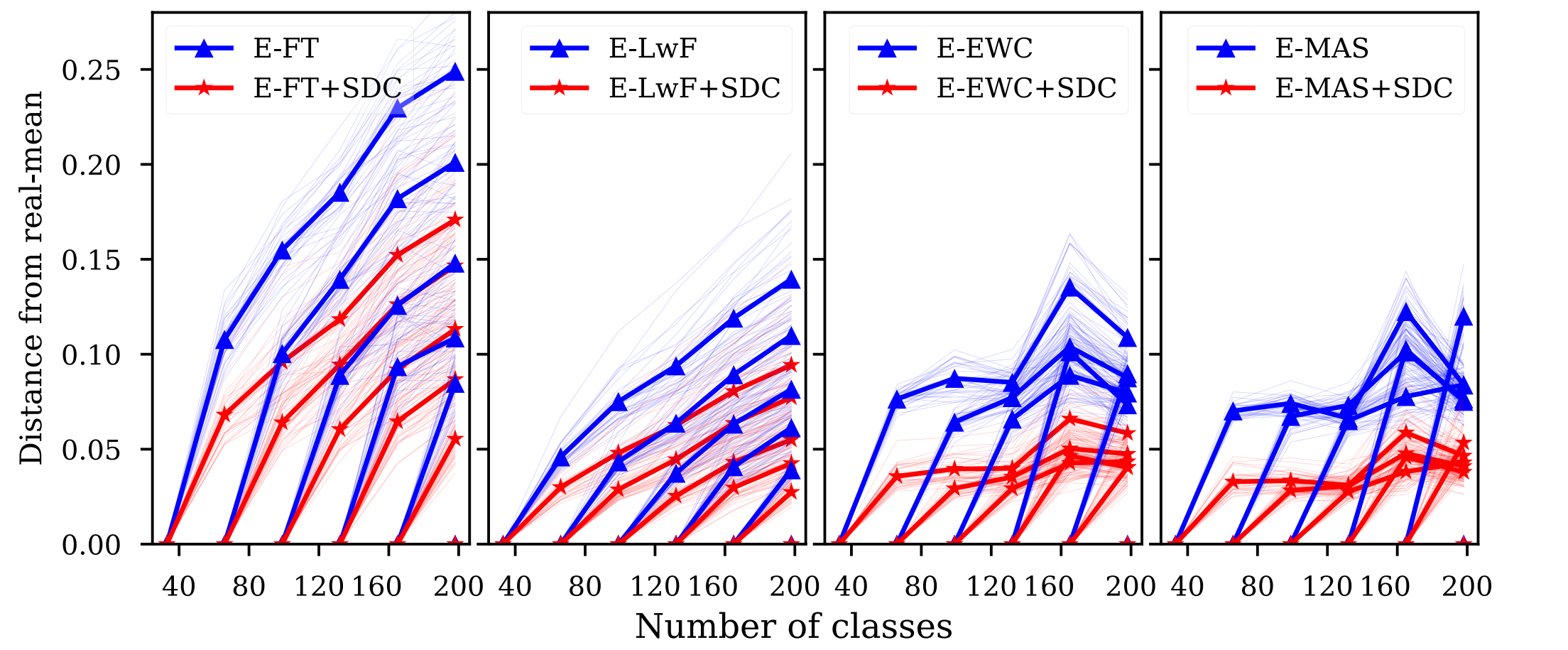
Comparison to State-of-the-Art Methods

1. On fine-grained datasets.

	CUB-200-2011						Flowers-102					
	T1	T2	T3	T4	T5	T6	T1	T2	T3	T4	T5	T6
E-Pre	78.5	69.1	62.1	58.1	54.7	52.1	90.9	77.5	77.7	76.1	75.2	73.6
E-Fix	84.1	70.6	61.7	56.9	53.5	50.3	98.2	83.6	82.8	80.1	78.4	76.9
FT	79.7	34.7	23.3	17.5	12.6	11.4	99.1	43.9	32.2	24.2	18.8	15.3
E-FT	84.1	73.6	62.5	54.2	43.0	37.4	98.2	76.0	59.3	50.2	42.4	29.1
E-FT+SDC	84.1	**75.5**	**69.5**	**63.6**	**57.5**	**49.3**	98.2	**85.5**	**74.1**	**61.9**	**49.8**	**35.3**
LwF	79.7	54.8	40.8	33.7	27.0	23.6	99.1	69.7	67.4	60.0	49.9	46.6
E-LwF	84.1	74.0	64.8	60.0	55.5	51.4	98.2	85.3	81.6	77.2	69.3	63.5
E-LwF+SDC	84.1	**74.4**	**65.9**	**61.3**	**57.3**	**52.7**	98.2	**86.1**	**82.2**	**79.6**	**74.7**	**69.7**
EWC	79.7	43.4	26.6	20.0	15.5	12.6	99.1	65.2	40.9	33.8	23.7	22.1
E-EWC	84.1	73.6	65.0	61.6	55.0	54.2	98.2	86.2	84.9	82.9	80.9	79.6
E-EWC+SDC	84.1	**74.8**	**67.4**	**62.8**	**58.2**	**56.4**	98.2	**87.6**	**86.9**	**86.0**	**84.2**	**83.9**
MAS	79.7	49.4	37.8	31.4	25.0	22.3	99.1	71.1	61.3	57.9	52.1	44.8
E-MAS	84.1	**72.5**	65.1	60.4	54.7	51.9	98.2	82.9	79.1	76.6	73.9	70.9
E-MAS+SDC	84.1	71.9	**65.3**	**61.1**	**57.3**	**54.4**	98.2	**83.1**	**80.7**	**78.8**	**76.8**	**76.0**

- SDC improves the results of all methods.
- The best overall accuracy with SDC outperform pre-trained model (E-Pre) and the model fixed after training the 1th task (E-Fix).
- E-LwF, E-EWC and E-MAS outperform E-FT.

2. Impact of SDC on the distance between real-mean and prototypes for CUB dataset over tasks.

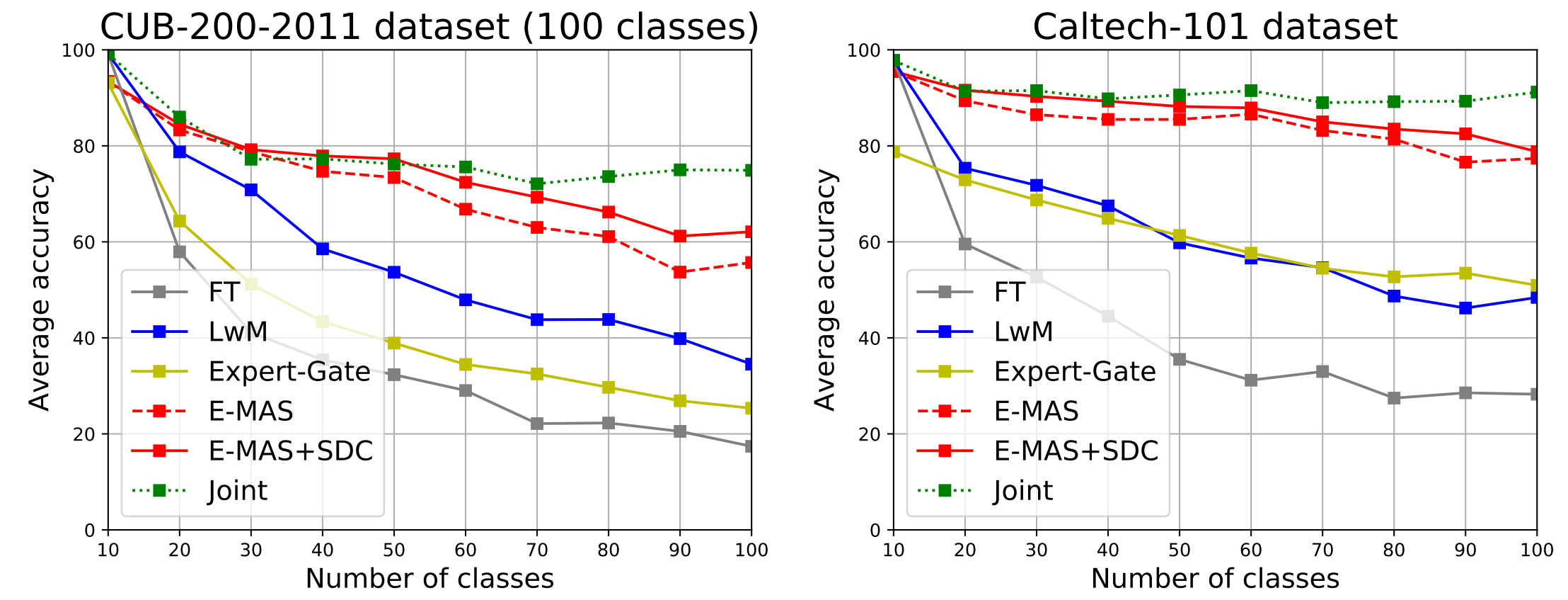


We measure the average distance between the real class-mean and the prototypes (before and after application of SDC). Each line represents a single class. Bold lines represent the mean value of all classes.

The graph confirms that SDC correctly compensates for part of the drift of the prototypes.

Comparison to State-of-the-Art Methods

Comparison of ten-task on CUB-200-2011 (100 classes) and Caltech-101. We obtain a clear superiority on both datasets.


CUB-200-2011 dataset (100 classes)
Average accuracy
FT
LwM
Expert-Gate
E-MAS
E-MAS+SDC
Joint
Number of classes
Caltech-101 dataset
Average accuracy
FT
LwM
Expert-Gate
E-MAS
E-MAS+SDC
Joint
Number of classes