A APPENDICES

A.1 CONTINUAL EARLY-EXIT NETWORKS WITH FINETUNING.

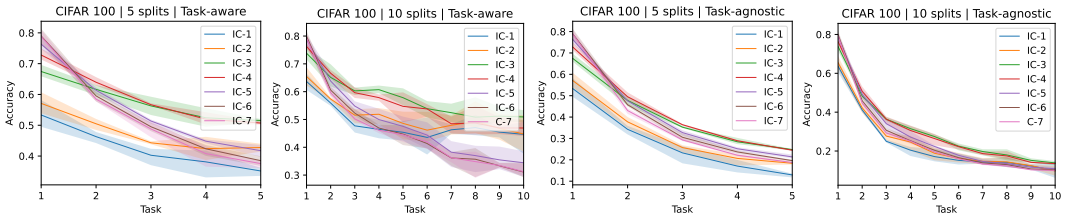
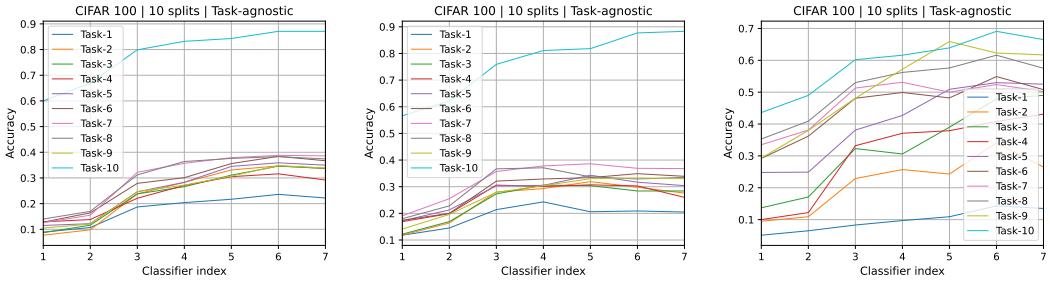


Figure 6: Finetuning without exemplars on CIFAR100 dataset.

In Figure 6, we present the performance of each classifier in a continually learned early-exit network using FT method at each task phase. In both task-incremental and class-incremental learning scenarios, the later classifiers exhibit more forgetting compared to the earlier ones, particularly noticeable in the last three classifiers.

A.2 RESULTS FOR MORE CL METHODS WITH EARLY-EXIT NETWORKS.



(a) FT-E (b) ER (c) BiC

Figure 7: CIL results of exemplar-based methods on CIFAR 100 (10-split).

Unlike the more conspicuous pattern shown by the exemplar-free method LwF, where early ICs perform significantly better for early task data, exemplar-based methods weaken a bit this phenomenon as shown in Figure 7. In these methods, early ICs maintain slightly higher or similar performance levels compared to later ICs for previous task data. Additionally, for FT-E and ER, there is a noticeable performance gap between new and old task data, whereas this distinction is absent in BiC due to the bias removal facilitated by its prediction rectification layer.

A.3 TASK-INCREMENTAL RESULTS

For the two exemplar-free methods considered in our experiments, when combined with an early-exit network for task incremental learning, both of them demonstrate significant performance improvements compared to training the standard network (only with the final classifier, referred to as 'NoEE' in the paper). Specifically, FT exhibits better performance than its corresponding NoEE version, even with a 75% increase in inference speed on both CIFAR100 and TinyImageNet datasets (refer to Table 3 and Table 4). Furthermore, achieving the best performance for FT with early-exit network requires only 50% of the inference computation cost, showing its efficiency for practical resource-limited applications. In the case of LwF, a 25% and 50% speed-up, compared to its corresponding NoEE, can be attained while maintaining the same level of performance in the 5-split and 10-split settings, respectively. As for the methods that store exemplars in memory, their performance at full inference capacity is similar to their corresponding NoEE's ones, except for FT-E, which shows similar performance with a 25% and 50% speed-up for the 5-split and 10-split settings on the CIFAR 100 dataset, and more than 25% speed-up on the TinyImageNet 10-split setting.

Method	5-split					10-split				
	Speed-up ↑					Speed-up ↑				
	NoEE	$\sim 0\%$	$\sim 25\%$	$\sim 50\%$	$\sim 75\%$	**NoEE**	$\sim 0\%$	$\sim 25\%$	$\sim 50\%$	$\sim 75\%$
JT	**85.30**	83.64	83.56	78.73	56.59	**90.53**	89.30	89.30	87.06	68.26
FT	36.12	**51.73**	52.26	52.08	39.57	32.34	**50.91**	51.40	52.05	47.29
LwF	74.18	**75.73**	75.58	71.82	54.88	72.53	**78.12**	78.04	76.18	63.69
FT-E	69.40	**71.71**	71.51	66.65	48.35	74.98	**77.30**	77.24	74.72	58.91
ER	**66.61**	65.01	64.87	60.91	46.82	74.05	**74.09**	74.07	72.07	58.57
BiC	**78.31**	77.91	77.79	73.19	54.02	81.72	**81.77**	81.73	79.29	64.54
LUCIR	**76.07**	74.57	73.63	64.20	45.91	**78.73**	76.94	76.42	70.52	54.71
iCaRL	**78.11**	77.35	77.18	72.36	54.60	81.65	**81.88**	81.79	79.49	65.30

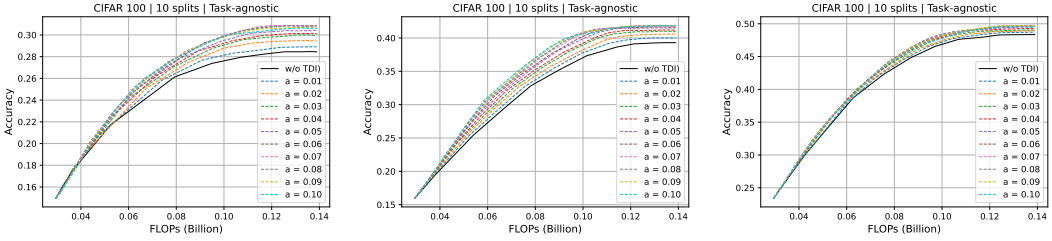
Table 3: TAW results on CIFAR100.

Method	5-split					10-split				
	Speed-up ↑					Speed-up ↑				
	NoEE	$\sim 0\%$	$\sim 25\%$	$\sim 50\%$	$\sim 75\%$	**NoEE**	$\sim 0\%$	$\sim 25\%$	$\sim 50\%$	$\sim 75\%$
JT	65.94	**66.36**	64.96	54.99	39.16	**74.19**	73.04	72.58	66.50	52.75
FT	27.69	**36.58**	36.73	33.77	26.03	23.85	**34.82**	35.25	34.09	26.82
LwF	57.85	**60.32**	58.94	50.96	38.49	51.06	**61.87**	61.38	57.69	47.44
ER	38.47	**39.23**	38.39	34.72	29.08	43.92	**46.25**	45.79	43.36	36.67
FT-E	44.06	**44.81**	43.39	37.92	29.48	46.91	**48.92**	48.25	44.31	35.89
BiC	**58.27**	58.22	57.15	50.15	38.99	59.36	**61.11**	60.41	54.33	42.34
LUCIR	**60.56**	57.25	53.96	43.99	32.05	**62.20**	60.65	57.97	48.80	37.24
iCaRL	**57.93**	57.69	56.25	49.14	38.54	56.46	**59.80**	58.98	53.25	42.80

Table 4: TAW results on TinyImageNet200.

One notable phenomenon is that LwF achieves comparable performance to the exemplar-based methods on the CIFAR 100 dataset at different levels of speed-up, and even outperforms them on the TinyImageNet dataset.

A.4 HYPERPARAMETERS OF TDI.



(a) LwF (b) FT-E (c) BiC

Figure 8: Results of TDI when using different hyperparameter setting for α.

As depicted in Figure 8, we illustrate the impact of the hyperparameter of α in our proposed TDI for various CL methods. We experimented with different values for α within the range of $[0.01, 0.1]$, all of which show performance improvement. Thus, we just opted for a uniform value of $\alpha = 0.05$ for all methods considered in this paper, which has already yielded enhanced performance. Nevertheless, more sophisticated strategies can be explored in future work.

A.5 NETWORK ARCHITECTURE

Table 5 and Table 6 display the details of our implemented early-exit ResNet 32 for CIFAR 100 and TinyImageNet datasets, respectively. It's worth noting that there is a difference in resolution between CIFAR 100 and TinyImageNet. Consequently, we have adapted the size of the fully connected layers for the internal classifiers accordingly.

LAYER	DOWNSAMPLE	OUTPUT SHAPE
Input	-	$32 \times 32 \times 3$
Conv1×1	-	$32 \times 32 \times 16$
BN	-	$32 \times 32 \times 16$
ReLU	-	$32 \times 32 \times 16$
ResBlk	-	$32 \times 32 \times 16$
ResBlk	-	$32 \times 32 \times 16$
ResBlk	-	$32 \times 32 \times 16$
FR (early exit)	MixPool	$4 \times 4 \times 16$
FC (early exit)	-	NumOfClasses
ResBlk	-	$32 \times 32 \times 16$
ResBlk	-	$32 \times 32 \times 16$
FR (early exit)	MixPool	$4 \times 4 \times 16$
FC (early exit)	-	NumOfClasses
ResBlk	stride=2	$16 \times 16 \times 32$
ResBlk	-	$16 \times 16 \times 32$
FR (early exit)	MixPool	$4 \times 4 \times 32$
FC (early exit)	-	NumOfClasses
ResBlk	-	$16 \times 16 \times 32$
ResBlk	-	$16 \times 16 \times 32$
FR (early exit)	MixPool	$4 \times 4 \times 32$
FC (early exit)	-	NumOfClasses
ResBlk	stride=2	$8 \times 8 \times 64$
ResBlk	-	$8 \times 8 \times 64$
FR (early exit)	MixPool	$4 \times 4 \times 64$
FC (early exit)	-	NumOfClasses
ResBlk	-	$8 \times 8 \times 64$
ResBlk	-	$8 \times 8 \times 64$
FR (early exit)	MixPool	$4 \times 4 \times 64$
FC (early exit)	-	NumOfClasses
ResBlk	-	$8 \times 8 \times 64$
ResBlk	-	$8 \times 8 \times 64$
-	AvgPool	$1 \times 1 \times 64$
FC (Final)	-	NumOfClasses

Table 5: Early-exit ResNet32 for the CIFAR100 dataset.

LAYER	DOWNSAMPLE	OUTPUT SHAPE
Input	-	$64 \times 64 \times 3$
Conv1×1	-	$64 \times 64 \times 16$
BN	-	$64 \times 64 \times 16$
ReLU	-	$64 \times 64 \times 16$
ResBlk	-	$64 \times 64 \times 16$
ResBlk	-	$64 \times 64 \times 16$
ResBlk	-	$64 \times 64 \times 16$
FR (early exit)	MixPool	$8 \times 8 \times 16$
FC (early exit)	-	NumOfClasses
ResBlk	-	$64 \times 64 \times 16$
ResBlk	-	$64 \times 64 \times 16$
FR (early exit)	MixPool	$8 \times 8 \times 16$
FC (early exit)	-	NumOfClasses
ResBlk	stride=2	$32 \times 32 \times 32$
ResBlk	-	$32 \times 32 \times 32$
FR (early exit)	MixPool	$8 \times 8 \times 32$
FC (early exit)	-	NumOfClasses
ResBlk	-	$32 \times 32 \times 32$
ResBlk	-	$32 \times 32 \times 32$
FR (early exit)	MixPool	$8 \times 8 \times 32$
FC (early exit)	-	NumOfClasses
ResBlk	stride=2	$16 \times 16 \times 64$
ResBlk	-	$16 \times 16 \times 64$
FR (early exit)	MixPool	$8 \times 8 \times 64$
FC (early exit)	-	NumOfClasses
ResBlk	-	$16 \times 16 \times 64$
ResBlk	-	$16 \times 16 \times 64$
FR (early exit)	MixPool	$8 \times 8 \times 64$
FC (early exit)	-	NumOfClasses
ResBlk	-	$16 \times 16 \times 64$
ResBlk	-	$16 \times 16 \times 64$
-	AvgPool	$8 \times 8 \times 64$
FC (Final)	-	NumOfClasses

Table 6: Early-exit ResNet32 for the TinyImageNet dataset.