Supplementary File: Are Exemplar-Based Class Incremental Learning Models Victim of Black-box Poison Attacks?

Neeresh Kumar Perla Md Iqbal Hossain Afia Sajeeda Ming Shao*
University of Massachusetts Dartmouth, Dartmouth, MA, USA
{nperla, mhossain10, asajeeda, mshao}@umassd.edu

1. Recap

Experiments are conducted on two dataset setups: (1) non-overlapped dataset and (2) overlapping dataset. In the non-overlapped dataset setup, the target model is trained on the CIFAR-100 [6] dataset consisting of 100 classes. These 100 classes are split into 10 tasks, each task containing 10 classes. The surrogate model is trained on ImageNet ILSVRC2012 [8] dataset, focusing on classes relevant to each task t of the target model. In the overlapped dataset setup, both target and surrogate models are trained on the CIFAR-100 dataset. Six target models — MEMO [17], DER [13], FOSTER [11], iCaRL [7], WA [15], and Replay [16] — are used in the experiments, along with four surrogate models — ResNet18 [3], VGG19 [9], DenseNet121 [4], and EfficientNet [10] — and five distinct adversarial attacks — MI-FGSM [1], DST-TI-FGSM [14], TI-MI-FGSM [2], ILA [5], and FIA [12] as shown in Figures 1, 2, 3, 4, and 5, for non-overlapped dataset and Figures 6, 7, 8, 9, and 10 for overlapped dataset.

2. Transferability and Forgetting

Transferability and Forgetting values for each experiment are shown in Tables 1 and 2. The transferability rate assesses the proportion of adversarial examples that are transferred to and deceived the target model and can be expressed as:

$$\text{Transferability Rate} \leftarrow \frac{|\{\forall x \in A_{\text{adv}} : \phi(x) \neq y\}|}{|A_{\text{adv}}|}. \quad (1)$$

Forgetting is calculated as the mean difference between the initial and final accuracies of each task and can be expressed as:

$$\text{Forgetting} \leftarrow \frac{1}{N}\sum_{i=1}^{N}(\text{Acc}_i^{\text{initial}} - \text{Acc}_i^{\text{final}}), \quad (2)$$

where N is the number of tasks. The mean forgetting value provides an overall measure of how much the model's performance has degraded from the original test accuracy.

3. Results of Non-overlapped Dataset

Figures 1, 2, 3, 4, and 5 illustrate the degradation in test accuracies for each task under MIFGSM [1], DST-TI-FGSM [14], TI-MI-FGSM [2], ILA [5], and FIA [12], respectively, using a non-overlapped dataset. MEMO, DER, and Replay were found to be highly vulnerable to adversarial samples generated by any surrogate model using our approach. As seen in the figures, the degradation in their test accuracies begins at early tasks, demonstrating their vulnerability. Conversely, iCARL, WA, and FOSTER show some resistance in early tasks, but they ultimately succumb to adversarial attacks in later tasks.

Table 1 presents the forgetting and transferability rates for each attack setting using the non-overlapping dataset. Transferability is the ratio of adversarial samples transferred to and deceived by the target model. Forgetting indicates the extent of information the model forgets after the attack. Notably, we observe some negative values in forgetting, indicating that while the target models were resistant in early tasks, they became vulnerable after a certain number of tasks.

Figure 11 visualizes the endpoint accuracy of clean and poisoned target models. The endpoint accuracy measures the last test accuracy of the target model, including data from all classes. A red line with the percentage drop shows the difference in test accuracies.

4. Results of Overlapped Dataset

Figures 6, 7, 8, 9, and 10 represent degradation in test accuracies for each task under an attack using an overlapped dataset. Among the five target models, MEMO, DER, and Replay were found to be highly vulnerable to adversarial samples generated by any surrogate model. Conversely, iCaRL, WA, and FOSTER show high resistance to adversarial samples generated using overlapped dataset compared to

*This research work is supported in part by the Marine and Undersea Technology (MUST) Research Program at the University of Massachusetts Dartmouth, funded by the Office of Naval Research (ONR) under Grant No. N00014-20-1-2170, National Science Foundation under Grant No. 2144772, and UMass Dartmouth Cybersecurity Center.

the non-overlapped dataset setting. WA and FOSTER were less affected by the overlapped dataset. This indicates vulnerabilities of the target model can be better revealed by a non-overlapping dataset.

Similar to Table 1, Table 2 presents the forgetting and transferability rates for each attack setting using the overlapped dataset. We observe that there are more negative values in forgetting when using overlapped dataset compared to non-overlapped results, indicating adversarial samples generated by non-overlapped dataset are more effective, allowing us to reveal the model's vulnerabilities better.

Similar to Figure 11, Figure 12 visualizes the endpoint accuracy of clean and poisoned target models, but on the overlapped dataset.

References

[1] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum, 2018.

[2] Yinpeng Dong, Tianyu Pang, Hang Su, and Jun Zhu. Evading defenses to transferable adversarial examples by translation-invariant attacks, 2019.

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.

[4] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks, 2018.

[5] Qian Huang, Isay Katsman, Horace He, Zeqi Gu, Serge Belongie, and Ser-Nam Lim. Enhancing adversarial example transferability with an intermediate level attack, 2020.

[6] Alex Krizhevsky. Learning multiple layers of features from tiny images. 2009.

[7] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H. Lampert. icarl: Incremental classifier and representation learning, 2017.

[8] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015.

[9] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition, 2015.

[10] Mingxing Tan and Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks, 2020.

[11] Fu-Yun Wang, Da-Wei Zhou, Han-Jia Ye, and De-Chuan Zhan. Foster: Feature boosting and compression for class-incremental learning, 2022.

[12] Zhibo Wang, Hengchang Guo, Zhifei Zhang, Wenxin Liu, Zhan Qin, and Kui Ren. Feature importance-aware transferable adversarial attacks, 2022.

[13] Shipeng Yan, Jiangwei Xie, and Xuming He. Der: Dynamically expandable representation for class incremental learning, 2021.

[14] Zebin Yun, Achi-Or Weingarten, Eyal Ronen, and Mahmood Sharif. The ultimate combo: Boosting adversarial example transferability by composing data augmentations, 2023.

[15] Bowen Zhao, Xi Xiao, Guojun Gan, Bin Zhang, and Shutao Xia. Maintaining discrimination and fairness in class incremental learning, 2019.

[16] Da-Wei Zhou, Qi-Wei Wang, Zhi-Hong Qi, Han-Jia Ye, De-Chuan Zhan, and Ziwei Liu. Deep class-incremental learning: A survey, 2023.

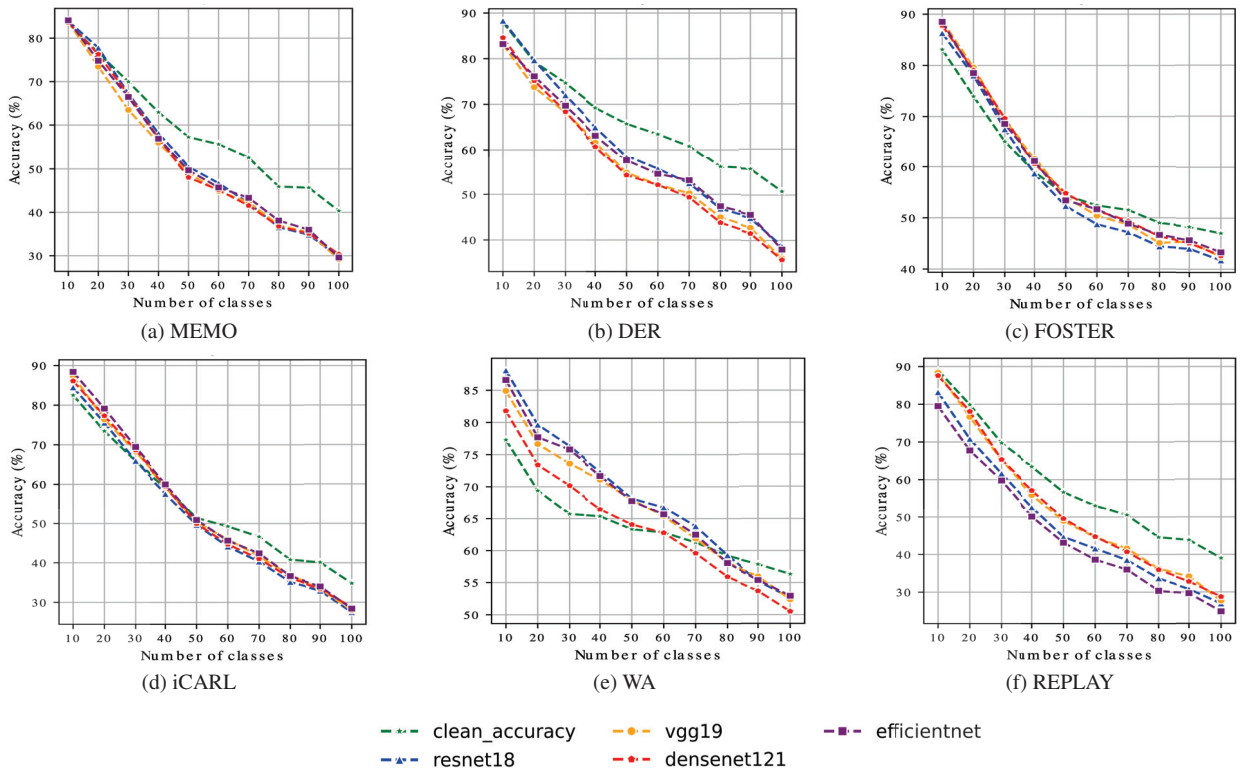
[17] Da-Wei Zhou, Qi-Wei Wang, Han-Jia Ye, and De-Chuan Zhan. A model or 603 exemplars: Towards memory-efficient class-incremental learning, 2023.

Attack	Target Model	Forgetting ↑				Transferability ↑			
		ResNet18	VGG19	DenseNet121	EfficientNet	ResNet18	VGG19	DenseNet121	EfficientNet
MI-FGSM	MEMO	6.29	7.59	6.95	6.56	46.21	46.06	48.94	49.37
	DER	6.20	9.63	9.83	7.56	45.99	51.64	50.03	48.46
	FOSTER	1.50	-0.16	-0.29	-0.24	51.64	55.33	53.15	48.71
	ICARL	3.08	1.64	1.85	0.9	51.97	54.35	52.99	50.79
	WA	-4.36	-2.93	0.03	-3.54	25.01	23.56	23.21	24.29
	REPLAY	10.56	7.05	6.92	13.01	52.76	54.26	51.59	55.96
DST-TI-FGSM	MEMO	9.05	8.75	9.29	8.45	45.65	49.41	51.46	48.86
	DER	7.70	6.0	7.50	11.69	48.70	49.75	48.93	49.73
	FOSTER	3.35	-0.06	0.13	2.79	52.28	53.0	51.04	52.12
	ICARL	2.93	6.83	1.59	6.82	52.08	61.66	52.10	56.49
	WA	-0.20	-3.77	-3.56	-4.20	22.42	23.96	23.47	24.46
	REPLAY	10.62	15.18	7.98	8.26	52.88	56.43	56.43	52.30
TI-MI-FGSM	MEMO	3.41	6.89	8.98	6.79	45.45	49.44	51.35	49.90
	DER	4.92	8.36	9.24	7.49	44.26	47.05	49.12	45.71
	FOSTER	-0.75	-1.15	1.0	2.47	49.13	48.50	51.87	54.85
	ICARL	0.94	0.50	2.71	3.01	56.41	55.23	56.66	55.35
	WA	-0.66	-2.98	-2.17	-5.27	24.04	23.32	23.61	21.68
	REPLAY	7.88	5.51	6.70	11.53	57.31	55.38	57.14	57.99
ILA	MEMO	7.06	6.09	6.41	5.19	47.92	50.02	49.92	30.32
	DER	6.69	8.36	7.46	6.75	46.34	44.97	50.94	45.10
	FOSTER	-0.73	1.72	0.91	1.09	46.87	46.68	50.29	46.67
	ICARL	2.01	2.97	3.45	4.44	48.25	52.64	52.15	50.78
	WA	-1.64	-1.64	-5.25	-1.06	23.18	22.74	23.73	24.32
	REPLAY	10.82	8.59	8.92	9.02	55.65	54.79	53.56	49.82
FIA	MEMO	4.51	7.47	7.30	7.24	46.72	52.15	46.81	49.44
	DER	5.91	7.26	7.46	7.67	49.50	49.51	48.29	47.37
	FOSTER	3.55	0.99	-0.21	-0.11	50.04	52.65	46.86	49.12
	ICARL	1.86	2.46	1.50	1.78	51.72	54.31	47.83	50.08
	WA	-3.47	-4.77	-4.52	-4.79	22.80	23.68	21.95	22.21
	REPLAY	11.67	6.40	7.10	6.92	53.56	49.58	49.49	49.00

Table 1. Forgetting & Transferability of our attack model on the **non-overlapped** dataset.

Attack	Target Model	Forgetting ↑				Transferability ↑			
		ResNet18	VGG19	DenseNet121	EfficientNet	ResNet18	VGG19	DenseNet121	EfficientNet
MI-FGSM	MEMO	3.31	4.44	7.24	5.20	48.00	48.68	51.51	47.49
	DER	9.60	7.95	11.03	6.49	47.52	43.20	47.29	39.82
	FOSTER	1.30	1.28	2.30	-0.65	48.12	48.73	52.33	44.45
	ICARL	-0.52	4.07	6.15	2.10	50.74	51.45	56.67	49.63
	WA	-3.06	-5.25	-3.24	-2.38	25.90	26.24	26.19	24.56
	REPLAY	9.92	7.60	5.71	8.07	56.59	53.61	57.61	53.66
DST-TI-FGSM	MEMO	7.68	4.23	10.18	8.65	51.62	849.03	53.79	52.13
	DER	6.99	10.20	4.94	10.54	48.19	50.15	45.59	55.19
	FOSTER	-1.54	0.04	-0.35	-2.19	46.31	50.07	47.48	43.24
	ICARL	2.60	3.23	4.98	4.3	57.75	56.41	56.55	55.25
	WA	-1.88	-6.03	-2.76	-5.6	27.24	26.23	27.56	24.75
	REPLAY	9.00	13.52	5.69	8.50	57.95	59.97	58.17	52.70
TI-MI-FGSM	MEMO	4.96	6.71	4.25	5.13	49.22	54.33	48.83	50.46
	DER	7.95	9.79	6.00	6.44	42.67	46.04	37.29	36.57
	FOSTER	-2.35	-1.53	2.76	-0.83	44.04	49.30	53.25	47.66
	ICARL	3.81	3.99	-1.93	0.72	58.71	60.78	53.19	55.60
	WA	0.26	-2.10	-4.86	-3.48	26.31	24.65	24.24	23.67
	REPLAY	5.46	5.82	5.28	6.36	55.22	61.21	58.75	59.72
ILA	MEMO	6.88	2.75	2.76	3.37	49.39	51.58	51.44	47.18
	DER	5.37	7.58	8.90	5.91	46.02	45.37	46.01	45.02
	FOSTER	3.04	0.57	0.56	2.73	48.44	50.23	46.81	47.42
	ICARL	6.682	1.35	3.11	1.78	59.15	50.70	53.47	49.97
	WA	0.41	-5.09	-1.30	-2.98	25.44	27.16	25.17	23.75
	REPLAY	11.84	8.15	8.25	8.60	54.87	51.91	54.59	51.32
FIA	MEMO	6.57	7.52	6.76	6.28	52.12	50.23	48.49	47.72
	DER	7.19	10.11	6.30	8.22	48.90	48.25	45.60	47.24
	FOSTER	0.91	0.97	0.73	-0.12	47.78	49.39	49.13	43.27
	ICARL	0.10	0.38	1.31	-0.14	51.50	51.97	48.58	49.37
	WA	-2.77	-5.90	-6.5	-4.97	27.33	27.58	23.74	24.01
	REPLAY	9.03	5.72	6.12	6.86	52.74	52.63	50.53	51.07

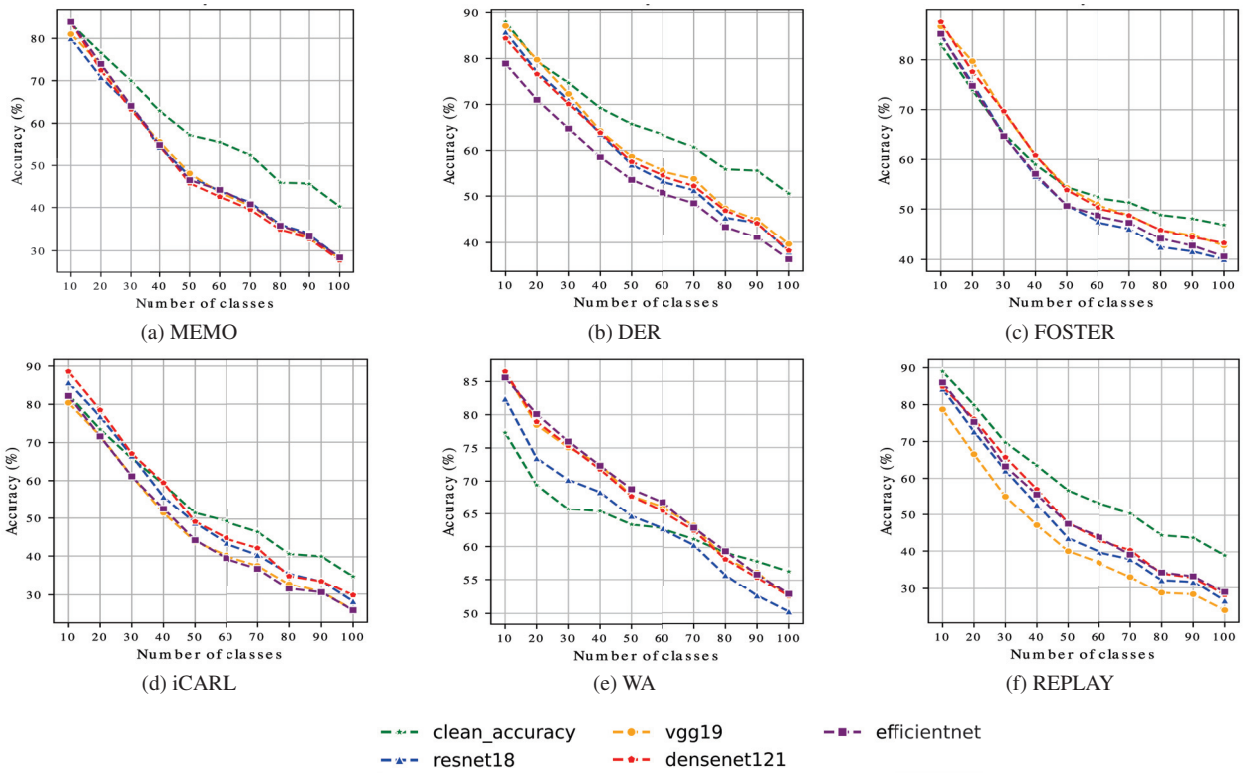
Table 2. Forgetting & Transferability of our attack model on the **overlapped** dataset.



(a) MEMO (b) DER (c) FOSTER

(d) iCARL (e) WA (f) REPLAY

Figure 1. CIL performance for each task under MIFGSM attack on the non-overlapped dataset.



(a) MEMO (b) DER (c) FOSTER

(d) iCARL (e) WA (f) REPLAY

Figure 2. CIL performance for each task under DST-TI-FGSM attack on the non-overlapped dataset.

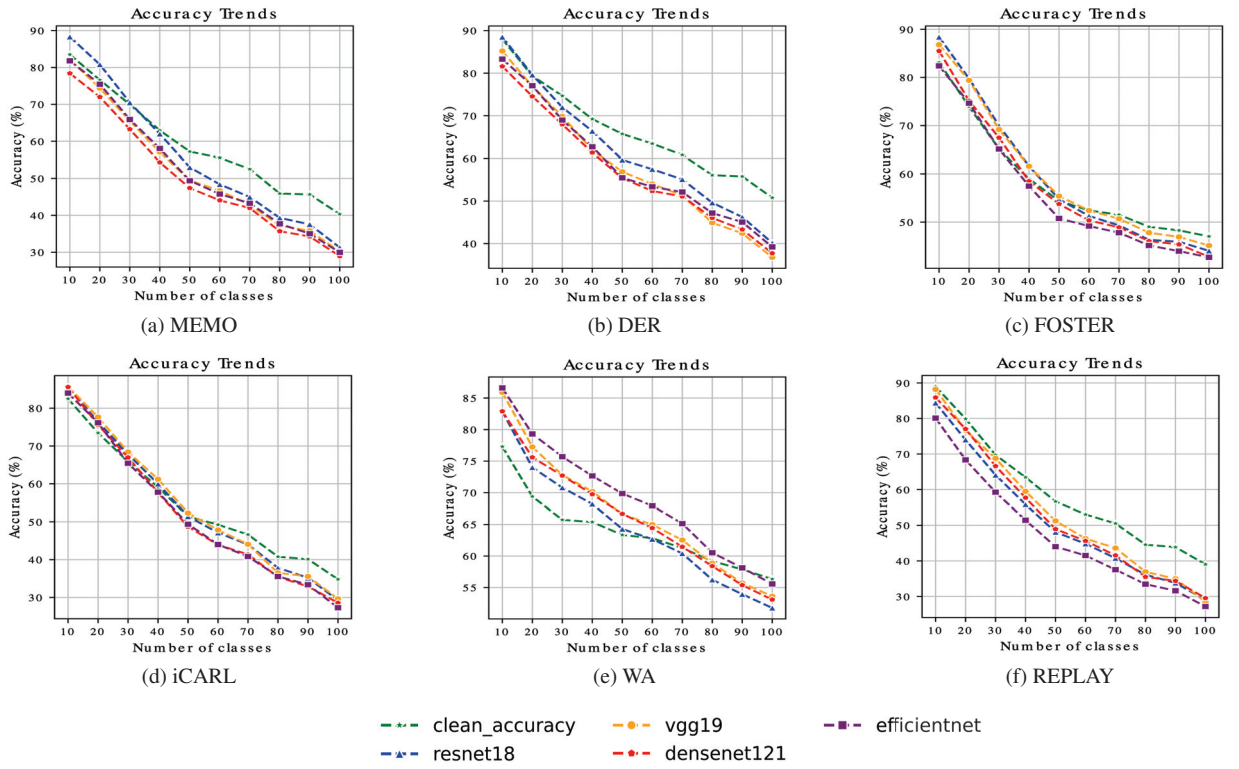


Figure 3. CIL performance for each task under TI-MI-FGSM attack on the non-overlapped dataset.

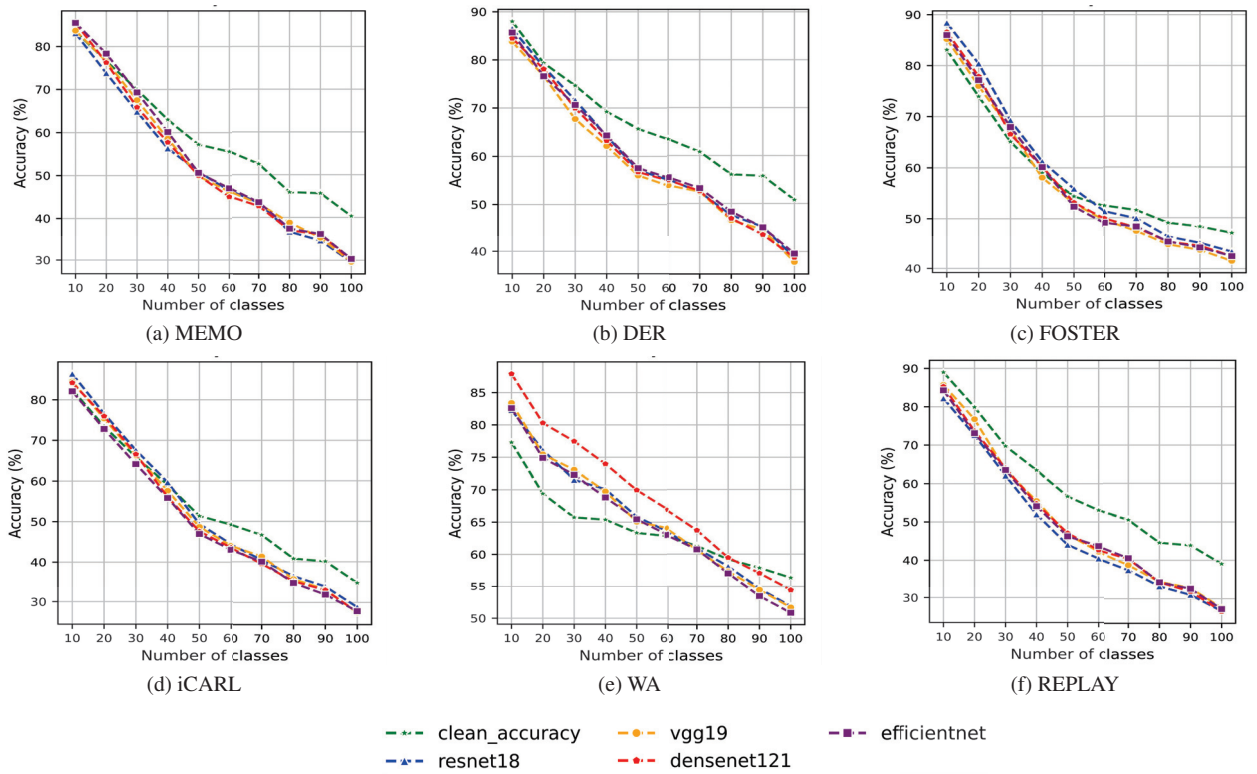
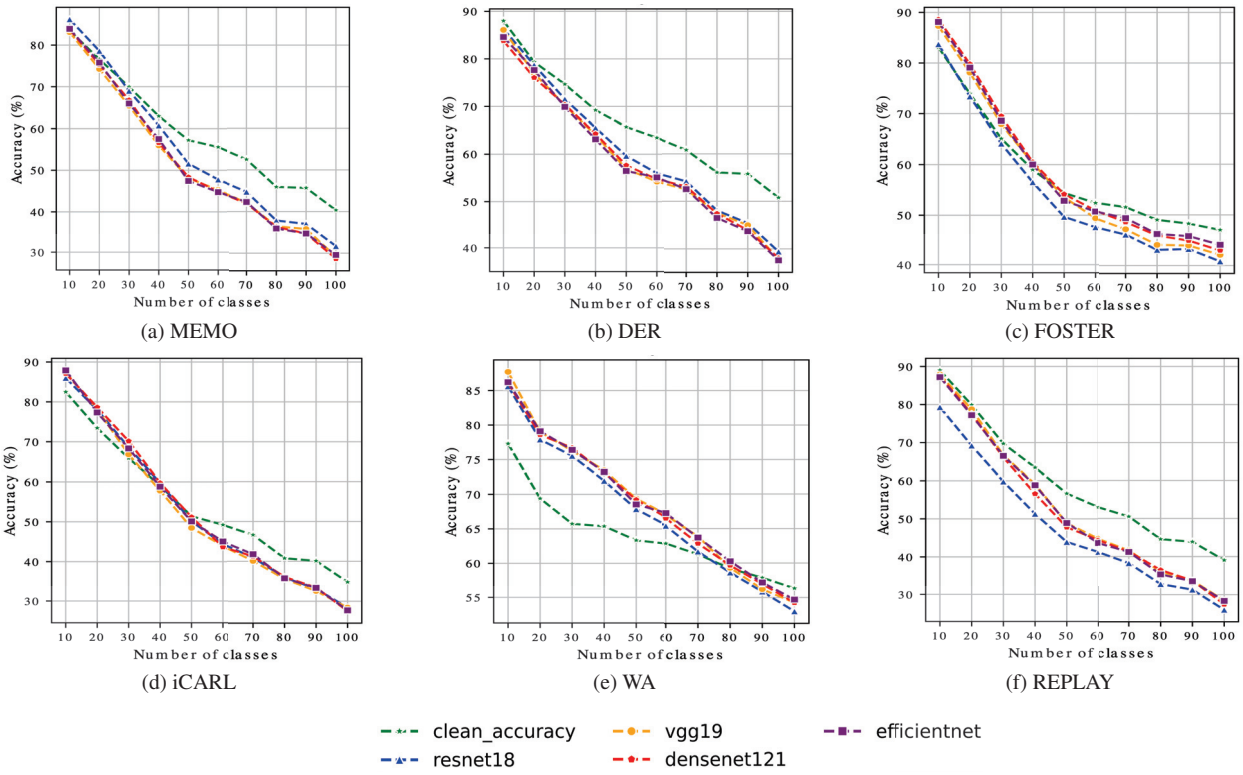


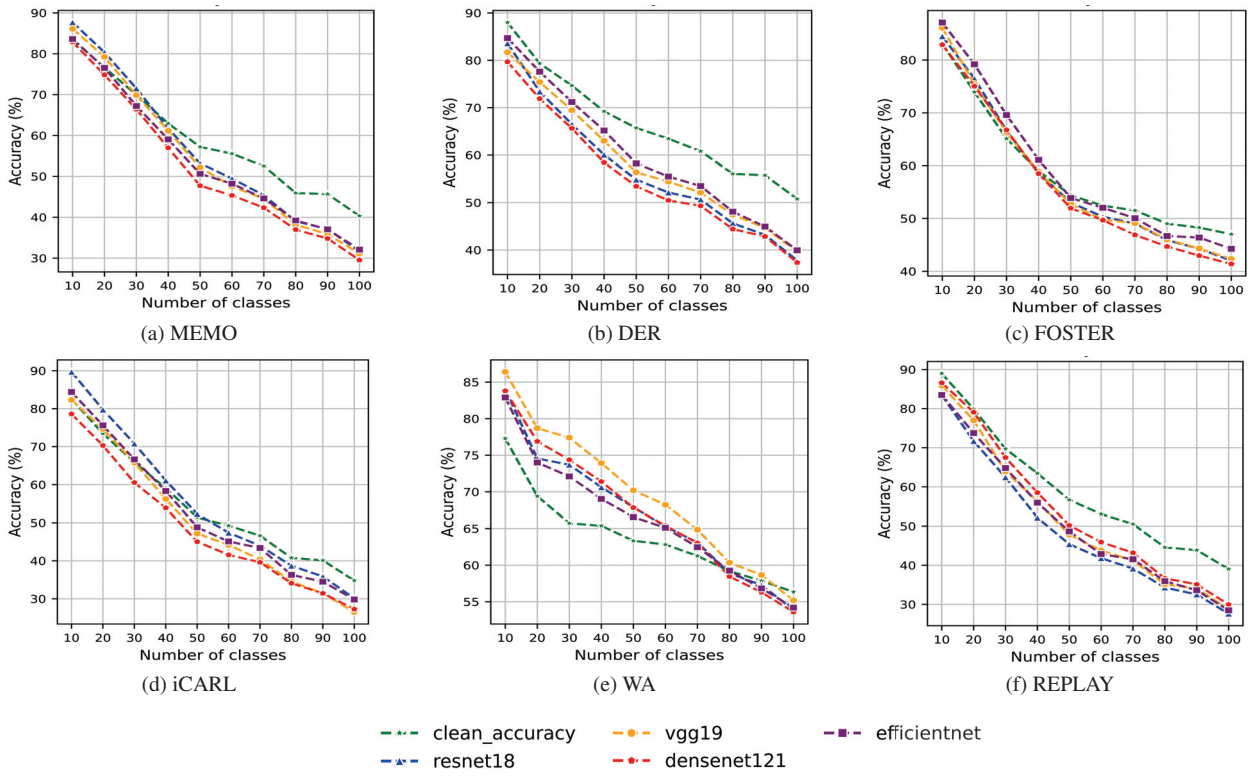
Figure 4. CIL performance for each task under ILA attack on the non-overlapped dataset.



(a) MEMO (b) DER (c) FOSTER

(d) iCARL (e) WA (f) REPLAY

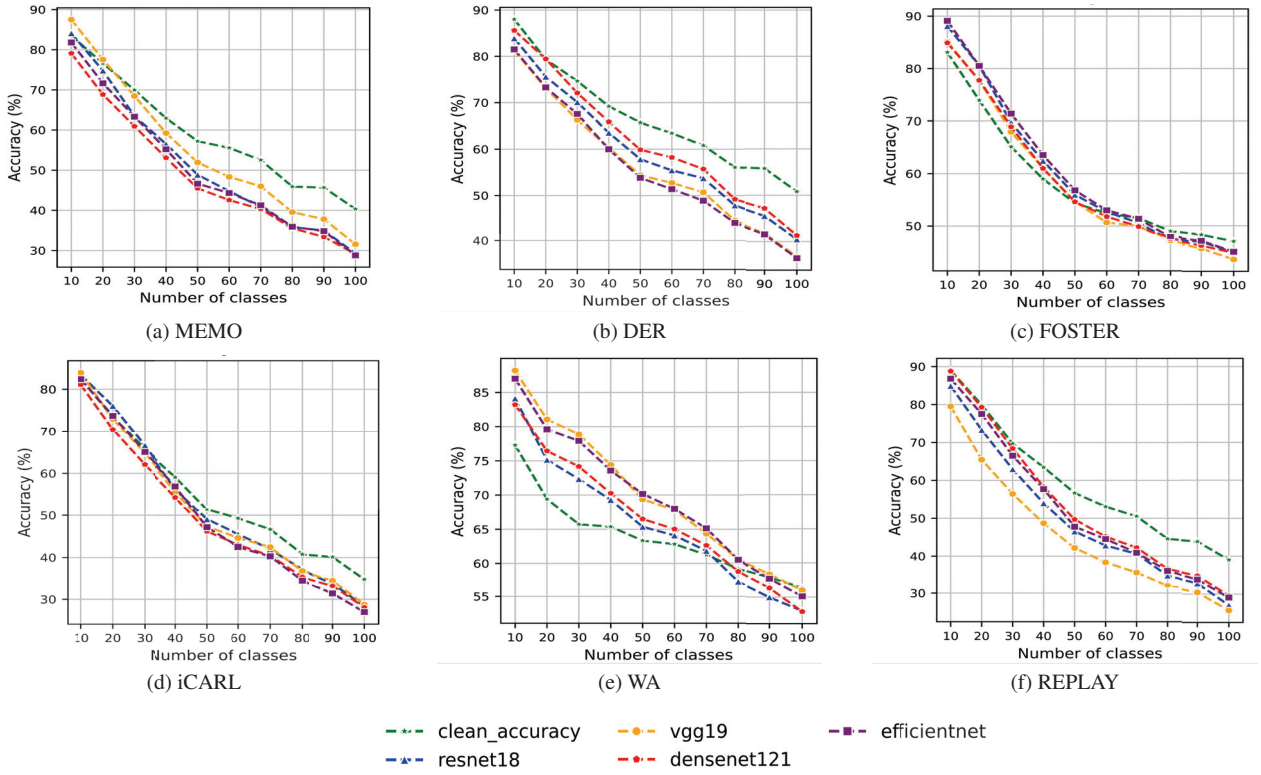
Figure 5. CIL performance for each task under FIA attack on the non-overlapped dataset.



(a) MEMO (b) DER (c) FOSTER

(d) iCARL (e) WA (f) REPLAY

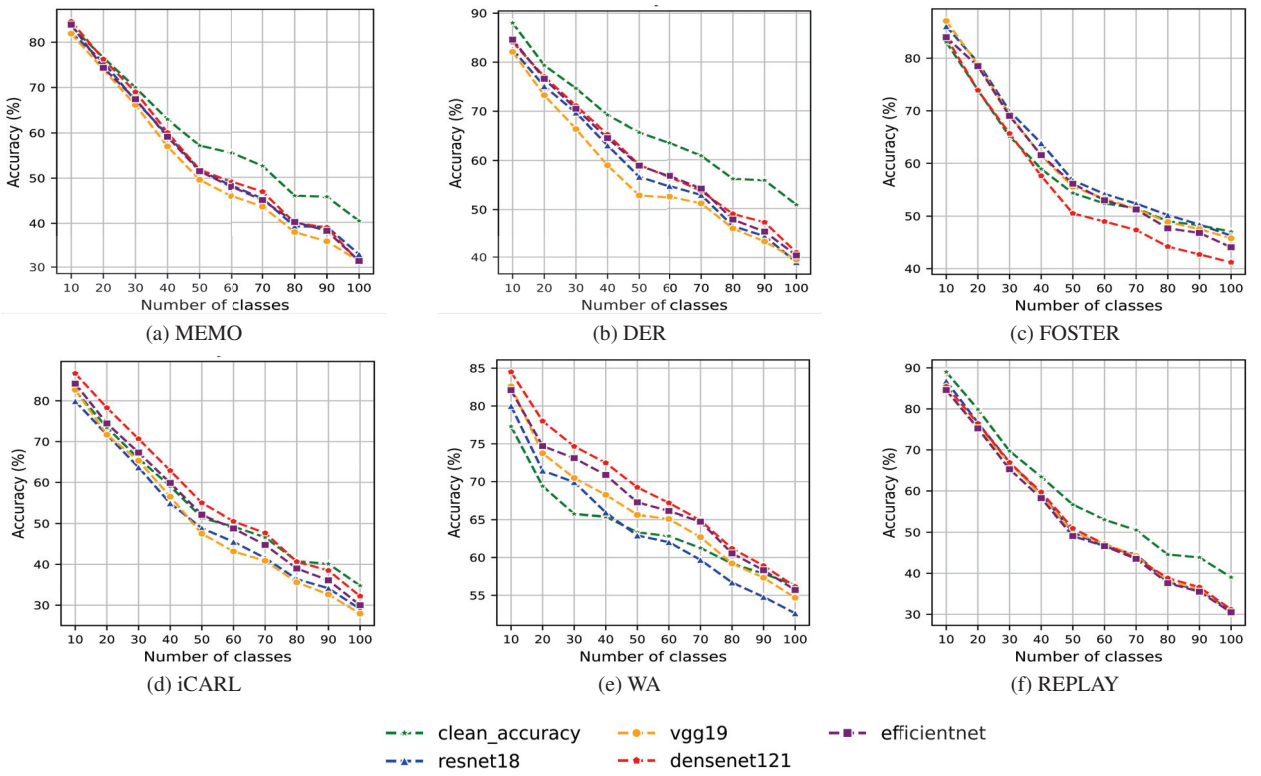
Figure 6. CIL performance for each task under MI-FGSM attack on the overlapped dataset.



(a) MEMO (b) DER (c) FOSTER

(d) iCARL (e) WA (f) REPLAY

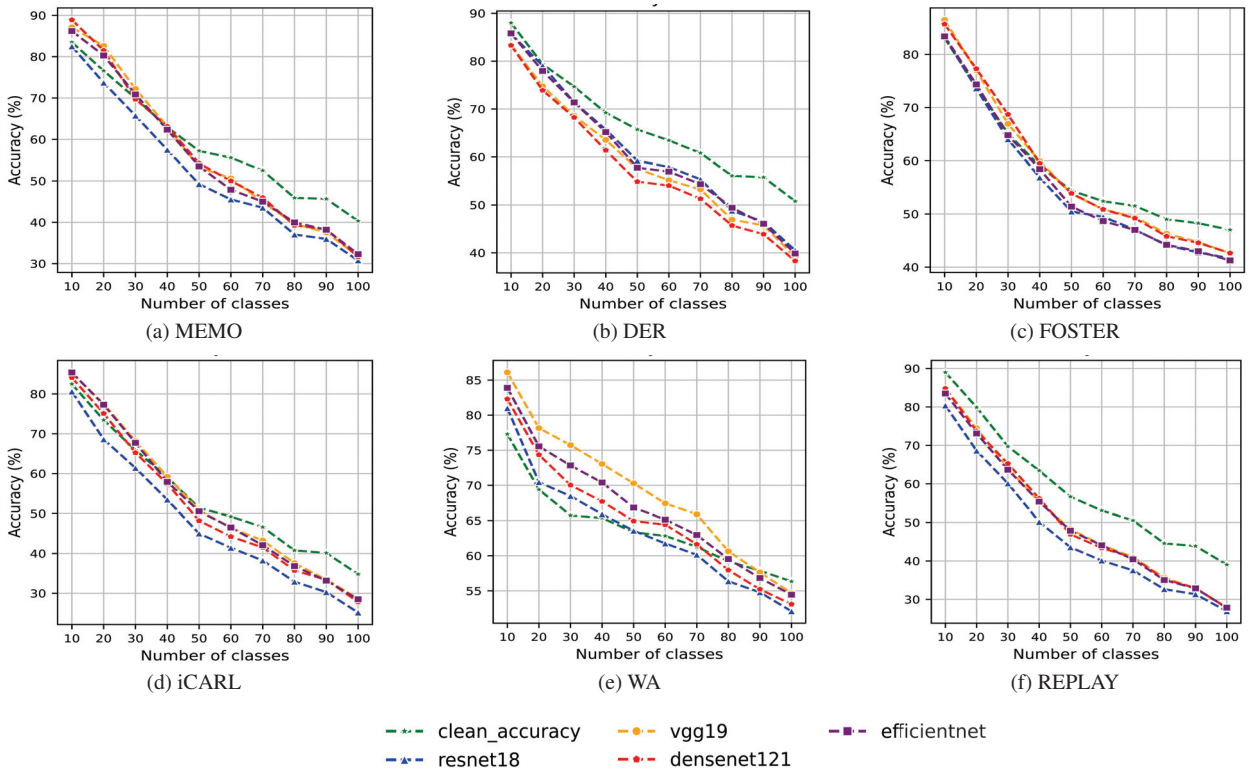
Figure 7. CIL performance for each task under DST-TI-FGSM attack on the overlapped dataset.



(a) MEMO (b) DER (c) FOSTER

(d) iCARL (e) WA (f) REPLAY

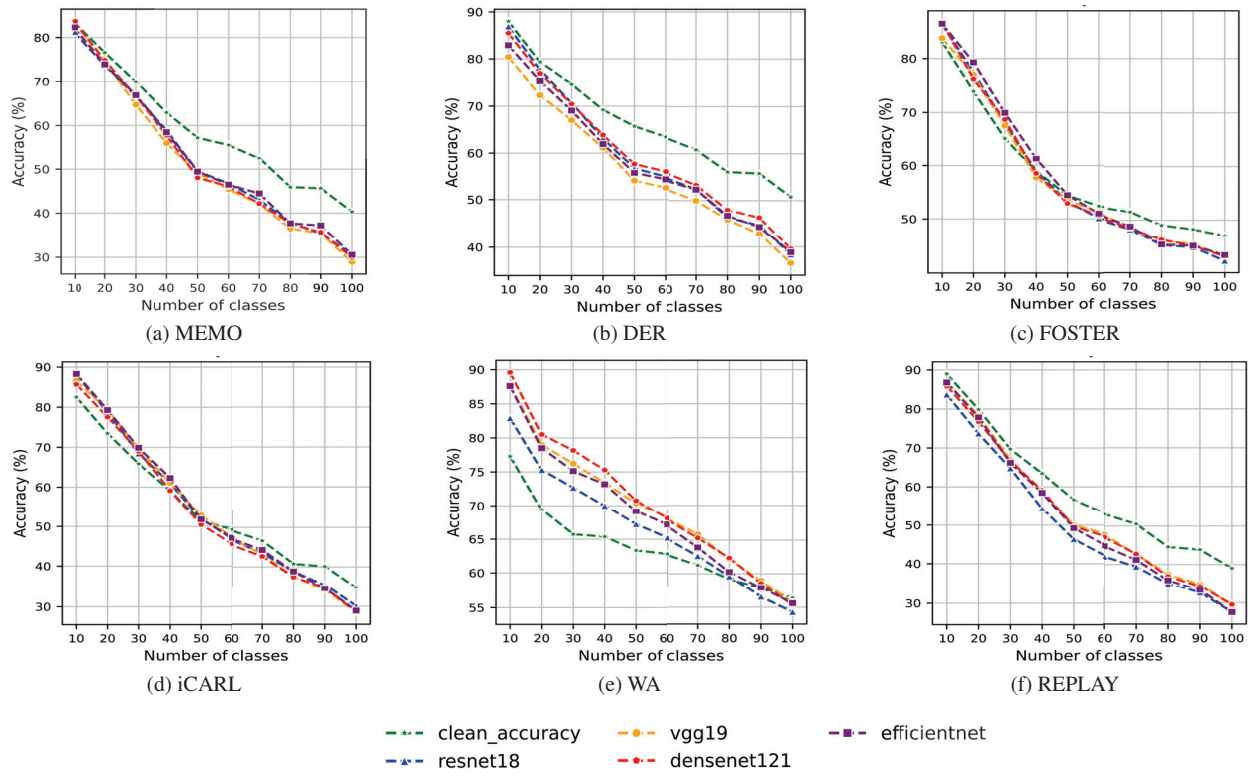
Figure 8. CIL performance for each task under TI-MI-FGSM attack on the overlapped dataset.



(a) MEMO (b) DER (c) FOSTER

(d) iCARL (e) WA (f) REPLAY

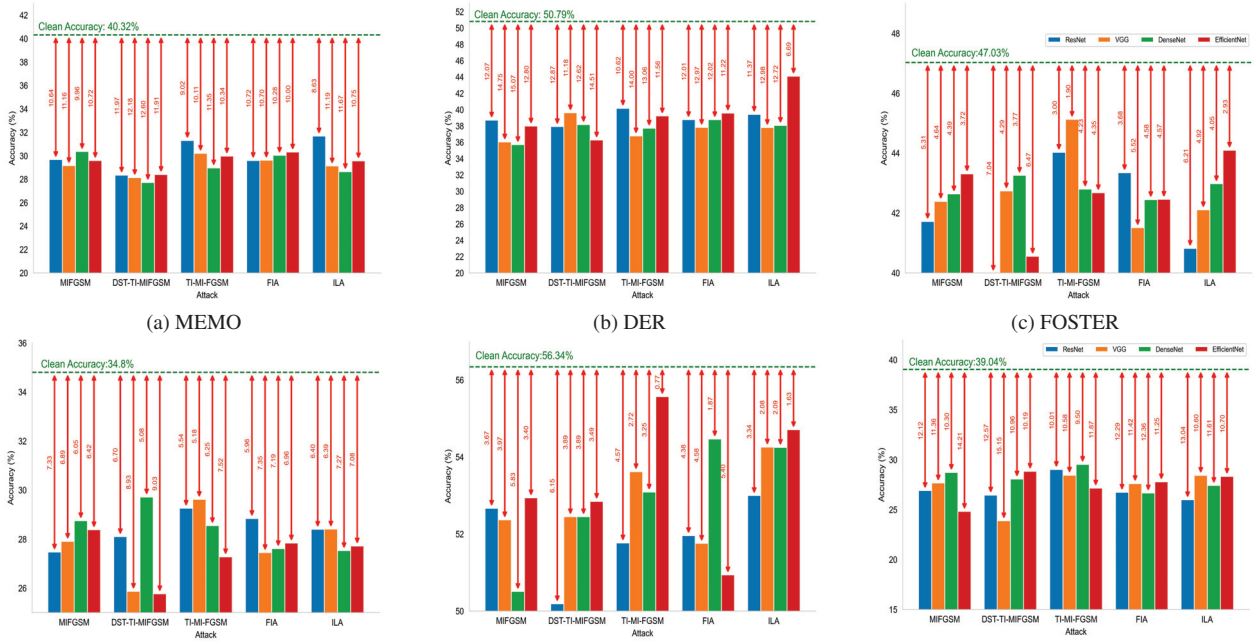
Figure 9. CIL performance for each task under ILA attack on the overlapped dataset.



(a) MEMO (b) DER (c) FOSTER

(d) iCARL (e) WA (f) REPLAY

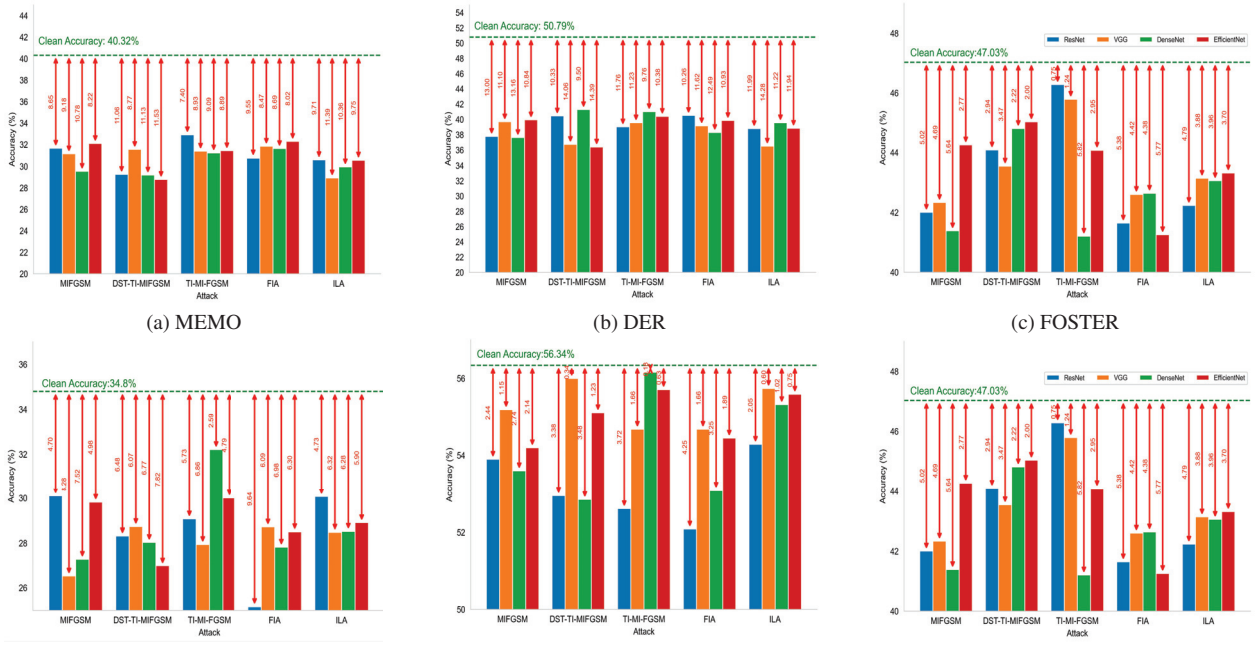
Figure 10. CIL performance for each task under FIA attack on the overlapped dataset.



(a) MEMO (b) DER (c) FOSTER

(d) iCARL (e) WA (f) REPLAY

Figure 11. Endpoint CIL accuracy on non-overlapped dataset.



(a) MEMO (b) DER (c) FOSTER

(d) iCARL (e) WA (f) REPLAY

Figure 12. Endpoint CIL accuracy on overlapped dataset.