APPENDIX

Anonymous authors
Paper under double-blind review

A LLM USAGE

We have used LLMs to refine the grammar, sentences and tables.

B REPRODUCIBILITY

All our codes and implementations are available at https://github.com/anony34/FoSSIL.

C FOSSIL BENCHMARK DATASET DETAILS

Med FoSSIL-Disjoint: The models with U-Net backbone are evaluated for their robustness and ability to handle a large number of incremental sessions while facing diverse domain shifts and scarce few-shot data. This represents a highly realistic setting in clinical domains, where annotations are costly and data is collected over time from multiple medical institutions. We evaluated the baselines from this benchmark on only two incremental sessions, as nearly all baselines collapsed after two sessions. We evaluated FoSSIL with a U-Net backbone across all five incremental sessions. The results demonstrate that the framework effectively handles the constraints and maintains strong performance throughout all sessions. This underscores the inability of existing baselines to effectively handle multiple constraints in continual segmentation across a large number of sessions. NC-FSCIL, which performs well for classification in few-shot class-incremental learning, fails on this benchmark. This is consistent with, which shows that a semantic segmentation model with a classifier fixed as a simplex equiangular tight frame (ETF) performs significantly worse than a model with a learnable classifier. MDIL , despite handling different domains with a shared encoder, struggles under strict constraints and suffers a drop in performance. Moreover, having multiple domains is not a practical solution when the number of domains increases. Approaches like generative replay (Gen-Replay) struggle to generate 3D medical volumes even with a diffusion model, which leads to poor performance. Most methods that perform well in incremental or few-shot incremental learning fail when confronted with domain shifts, including segmentation approaches such as GAPS, MiB, PIFS, and CLIP-CT. Representation and meta-learning-based approaches, such as MAML, SupCL, UnSupCL, and MTL, also fail to retain performance over the sessions.

Med FoSSIL-Mixed: A more realistic clinical setting involves the same classes appearing across different domains. The goal is to evaluate the model's ability to learn previously seen classes in new domains while retaining performance on the original domains. Additionally, multiple domains may appear within the same session, making learning and adaptation more challenging, as the model must simultaneously learn from diverse domains while managing the usual constraints. In this setting, we evaluated different transformer backbones, including MedFormer, SwinUNetr, and a large pre-trained (CLIP-driven U-Net model, which had already been exposed to most of the 35 classes. Results show that even robust transformer backbones are unable to maintain model performance across sessions. An interesting observation is that the CLIP-driven model, already pre-trained on 21 of the 35 classes, experiences a drop in performance across sessions, highlighting the severity of the constraints in the benchmark.

Med Semi-Supervised-FoSSIL: This benchmark is a variant of Med FoSSIL-Disjoint, where few-shot classes have access to unlabeled data. The inclusion of unlabeled data significantly boosts the performance of FoSSIL, an improvement not observed with other methods. Benchmark also highlights how existing semi-supervised approaches fail to handle the constraints and are unable to effectively leverage unlabeled data. **FoSSIL clearly demonstrates that using readily available**

unlabeled data can significantly improve multi-constraint continual learning for semantic segmentation.

Natural-FoSSIL: In autonomous driving, scenes evolve over time with distributional shifts and limited labeled data, making this a highly realistic benchmark. Even large-scale pre-trained models like SAM (trained on over a billion natural scene masks) exhibit forgetting in this realistic and challenging setting. It can be observed that FoSSIL with a SAM backbone achieves further improvements, highlighting that the proposed framework enhances not only simple backbones (e.g., U-Net) and transformer-based backbones (e.g., MedFormer, SwinUNetr) but also heavily pre-trained models like SAM.

Semi-Supervised Natural-FoSSIL: This benchmark provides access to abundant unlabeled data, and all baselines are evaluated with a DeepLab backbone. In this setting, we employ FoSSIL as a plug-and-play module on top of the GAPS baseline, further improving its results, by leveraging unlabeled data. This demonstrates that unlabeled data can significantly enhance performance and that FoSSIL can effectively boost existing baselines. Similar to the medical benchmarks, existing semi-supervised methods perform poorly in this setting and are unable to effectively leverage unlabeled data. We also evaluated active learning approaches, such as HALO, and coreset selection methods, like RETRIEVE, to select the most informative data; however, these methods fail to achieve significant improvements.

D Detection results

To check the effectiveness of FoSSIL on detection tasks with distributional shifts, we tested FoSSIL on the detection dataset COCO-O[1] having different domains with RCNN[2] backbone as shown in Table 6. We created base session with 77 classes from 'Painting' domain and incremental session 1 from 'Weather' domain with three novel disjoint few-shot classes, 'person', 'car', and 'bench'. The performance of base model is 0.2840 mAP. We took 41, 224, 174 unlabeled samples from 'bench', 'car', and 'person' classes, respectively with 30 few-shot labeled samples from each class for session 1. Table 6 illustrates that FoSSIL is able to improve the performance over the vanilla backbones like RCNN, even on detection tasks with large domain shifts.

E Implementation details

Baselines: We have re-implemented and adapted the following baselines - *Class-incremental Learning*: MiB, CLIP-CT; *Domain Incremental learning*: MDIL; *Few-shot Class-incremental Learning*: PIFS, Subspace, C-FSCIL, FACT, NC-FSCIL, Gen-Replay, GAPS, SoftNet, FSCIL-SS, FeCAM. *Semi-Supervised Learning based approaches*: RETRIEVE, NNCSL, UaD-CE. *Other methods*: SupCL, UnSupCL, UnSupCL-HNM, MTL, MAML, CLIP-driven, HALO.

All baselines were adapted with their original settings and hyperparameters. Medical experiments were run for 200 epochs for both base and incremental sessions, while natural experiments were run for 100 epochs per session. All SAM experiments were run for 10 epochs. We used one A100 40GB GPU for all experiments.

F Additional results

We report session wise results for FoSSIL and some baselines. We showcase how class performances vary across the sessions. We report results of FoSSIL with guided noise injection in Table 14 and without pseudo-label refinement in Table 15. Table 24 and Table 12 highlight that the performance of pre-trained model deteriorates in the proposed multi-constraint benchmark.

[1] https://openaccess.thecvf.com/content/ICCV2023/papers/Mao_COCO-O_A_Benchmark_for_Object_Detectors_under_Natural_Distribution_Shifts_ICCV_2023_paper.pdf

[2] https://github.com/microsoft/SoftTeacher

Table 1: Dataset Splits for Med FoSSIL-Disjoint benchmark. Session 0 (base) is from domain TotalSegmentator (TS), Session 1 is from Amos, Session 2 is from BCV, Session 3 is from MOTS, Session 4 is from Brats, Session 5 is from Verse. For incremental sessions, Train is 5-shots.

Class	Source	Train	Val	Test
Sacrum - 1	TS	136	27	43
Stomach - 2	TS	124	29	53
Lung upper lobe left - 3	TS	188	41	92
Lung lower lobe left - 4	TS	179	39	83
Brain - 5	TS	125	35	79
Atrium (left) - 6	TS	127	25	52
Ventricle (left) - 7	TS	123	24	50
Pulmonary artery - 8	TS	116	17	43
Aorta - 9	TS	200	39	85
Gallbladder - 10	TS	120	27	47
Trachea - 11	TS	157	30	68
Rib left1 - 12	TS	156	28	66
Rib right1 - 13	TS	156	29	67
Rib left2 - 14	TS	156	29	67
Rib right2 - 15	TS	155	29	66
Pancreas - 16	Amos	120	30	60
Duodenum - 17	Amos	120	30	60
Bladder - 18	Amos	99	26	47
Prostate/uterus - 19	Amos	96	26	47
Postcava - 20	Amos	120	30	60
Spleen - 21	BCV	18	4	8
Liver - 22	BCV	18	4	8
Left kidney - 23	BCV	18	4	8
Right kidney - 24	BCV	18	4	8
Left adrenal - 25	BCV	18	4	8
Right adrenal - 26	BCV	18	4	8
Hepatic Vessel - 27,28	MOTS	7	4	4
Hepatic Vessel Tumor	MOTS	7	4	4
Colon Tumor - 29	MOTS	7	4	4
Lung Tumor - 30	MOTS	7	4	4
NCR — label 1	Brats	40	4	8
ED — label 2	Brats	40	4	8
ET — label 4	Brats	37	4	8
C3(3)	Verse	12	5	6
C4(4)	Verse	12	5	6
T1(8)	Verse	31	5	8
T2(9)	Verse	26	5	7

Table 2: Dataset Splits for Med FoSSIL-Mixed benchmark with classes and domains repeating. Session 0 (base) is from the domain Amos, Session 1 is from BCV and MOTS, Session 2 is from TS and Amos with three classes repeating (Right kidney, Left kidney and Stomach) from TS domain. Session 3 is from MOTS and TS, Session 4 is from Brats and Verse. For incremental sessions, Train is 5-shots.

Class	Source	Train	Val	Test
Spleen - 1	Amos	203	66	68
Right kidney - 2	Amos	234	77	77
Left kidney - 3	Amos	233	78	78
Gall bladder - 4	Amos	190	61	67
Esophagus - 5	Amos	204	67	68
Prostate/uterus - 6	Amos	167	55	52
Stomach - 7	Amos	233	77	77
Arota / aorta - 8	Amos	204	68	68
Duodenum - 9	Amos	203	68	68
Postcava - 10	Amos	204	68	68
Pancreas - 11	BCV	18	6	6
Inferior vena cava - 12	BCV	18	6	6
Portal vein and splenic vein - 13	BCV	18	6	6
Left adrenal - 14	BCV	18	6	6
Right adrenal - 15	BCV	18	6	6
Liver - 16	BCV	18	6	6
Hepatic Vessel - 17	MOTS	7	4	4
Hepatic Vessel Tumor - 18	MOTS	7	4	4
Small bowel - 19	TS	60	20	20
Hip_left - 20	TS	60	20	20
Hip_right - 21	TS	60	20	20
Lung upper lobe left - 22	TS	60	20	20
Lung lower lobe left - 23	TS	60	20	20
Bladder - 24	Amos	9	4	4
Colon Tumor - 25	MOTS	7	4	4
Lung Tumor - 26	MOTS	7	4	4
Femur_left - 27	TS	18	6	6
Femur_right - 28	TS	18	6	6
T1	Brats	40	4	8
T2	Brats	40	4	8
Flair	Brats	37	4	8
C3(3)	Verse	12	5	6
C4(4)	Verse	12	5	6
T1(8)	Verse	31	5	8
T2(9)	Verse	26	5	7

Table 3: Unlabeled sample counts per session in Med Semi-Supervised-FoSSIL benchmark.

Session	Count
Session 1 - Amos	25
Session 2 - BCV	30
Session 3 - MOTS	8
Session 4 - Brats	15
Session 5 - Verse	20

Table 4: Dataset Splits for Natural-FoSSIL benchmark. For incremental sessions, Train is 10-shots and Val is 3 for each class.

Class	Source	Train	Val	Test
road - 0	BDD - Session 0	4691	1184	847
sidewalk - 1	BDD - Session 0	3173	796	612
building - 2	BDD - Session 0	4270	1070	758
wall - 3	BDD - Session 0	772	189	111
fence - 4	BDD - Session 0	1494	360	206
traffic light - 5	BDD - Session 0	2187	556	443
traffic sign - 6	BDD - Session 0	3625	919	660
person - 7	BDD - Session 0	1562	393	305
car - 8	BDD - Session 0	4738	1180	833
truck - 9	BDD - Session 0	1451	389	283
Bridge/tunnel - 10	IDD - Session 1	360	3	54
Parking - 11	IDD - Session 1	1262	3	173
rail track - 12	IDD - Session 1	896	3	155
Autorickshaw - 13	IDD - Session 1	933	3	139
pole - 14	IDD - Session 1	1432	3	220
Bus - 15 (from BDD)	BDD - Session 2	397	3	58
Sky - 16 (from BDD)	BDD - Session 2	886	3	139
bicycle - 17 (from BDD)	BDD - Session 2	336	3	45
person - 7 (repeat from IDD) - 18	IDD - Session 2	528	3	82
car - 8 (repeat from IDD) - 19	IDD - Session 2	590	3	91

Table 5: Dataset Splits for Semi-Supervised Natural-FoSSIL benchmark with 400 samples per class for unlabeled data. For incremental sessions, Train is 10-shots and Val is 3 for each class.

Class	Source	Train	Val	Test
road - 0	BDD : Session 0	4691	1184	847
sidewalk - 1	BDD : Session 0	3173	796	612
building - 2	BDD : Session 0	4270	1070	758
wall - 3	BDD : Session 0	772	189	111
fence - 4	BDD : Session 0	1494	360	206
traffic light - 5	BDD : Session 0	2187	556	443
traffic sign - 6	BDD : Session 0	3625	919	660
person - 7	BDD : Session 0	1562	393	305
car - 8	BDD : Session 0	4738	1180	833
truck - 9	BDD : Session 0	1451	389	283
parking - 10	Cityscapes : Session 1	720	3	111
vegetation - 11	Cityscapes : Session 1	2891	3	486
rail track - 12	IDD : Session 2	1258	3	214
sky - 13	IDD : Session 2	1945	3	300
motorcycle - 14	IDD : Session 3	4411	3	609
autorickshaw - 15	IDD : Session 3	2825	3	361
bridge/tunnel - 16	IDD : Session 3	445	3	65

Table 6: Performance of FoSSIL on COCO-O detection benchmark with one incremental session on three few-shot novel classes. We compared FoSSIL without Guided Noise Injection (GNI) and Pseudo-Label Refinement (PLR).

Method	Session 1 (mAP)
RCNN (Vanilla with only unlabeled data)	0
FoSSIL	0.063
FoSSIL w/o GNI	0.004
FoSSIL w/o PLR	0.0510

Table 7: Configuration of FoSSIL in different settings.

Method	Base Frozen (Yes/No)	Incremental Sessions Frozen (which layers)
Med FoSSIL-Disjoint (U-Net)	No	Yes (All but last)
Med FoSSIL-Mixed (MedFormer)	No	Yes (All but last)
Med FoSSIL-Mixed (SwinUNetr)	No	Yes (All but last)
Natural-FoSSIL (DeepLab)	No	No
Semi-Supervised Natural-FoSSIL (Deeplab)	No	No
Med Semi-Supervised-FoSSIL (U-Net)	No	Yes (All but last)
Med Semi-Supervised-FoSSIL (MedFormer)	No	Yes (All but last)
Natural-FoSSIL (SAM)	No	Yes (All but last)

Table 8: Session-wise performance of FoSSIL on Med FoSSIL-Disjoint with U-Net.

Class	Session 0	Session 1	Session 2	Session 3	Session 4	Session 5
class_0	0.9788	0.9619	0.8480	0.7644	0.9440	0.6723
class_1	0.7916	0.6139	0.6106	0.5528	0.0023	0.7911
class_2	0.6125	0.2090	0.3410	0.4537	0.0023	0.5406
class_3	0.8557	0.7920	0.7192	0.7174	0.0032	0.8342
class_4	0.7555	0.6356	0.6228	0.6463	0.0032	0.7355
class_5	0.9101	0.7691	0.7295	0.6009	0.4577	0.8773
class_6	0.6534	0.6410	0.5457	0.5572	0.1274	0.6222
class_7	0.7035	0.5512	0.5585	0.5677	0.0290	0.6560
class_8	0.8745	0.8354	0.8022	0.7330	0.1821	0.8104
class_9	0.7980	0.7185	0.7110	0.6871	0.0480	0.7631
class_10	0.6101	0.2715	0.4802	0.5417	0.0019	0.4439
class_11	0.8326	0.7311	0.6869	0.6825	0.0298	0.8250
class_12	0.8239	0.6738	0.6813	0.6497	0.0822	0.7820
class_13	0.7290	0.6382	0.5920	0.5789	0.0693	0.7016
class_14	0.7618	0.6346	0.5950	0.3970	0.0458	0.6886
class_15	0.7046	0.4307	0.5000	0.5387	0.0189	0.6537
class_16	–	0.1285	0.1417	0.1337	0.0401	0.1224
class_17	–	0.0434	0.0584	0.0622	0.0173	0.0839
class_18	–	0.0503	0.0675	0.0709	0.0208	0.0718
class_19	–	0.0534	0.0165	0.0121	0.0155	0.0152
class_20	–	0.0956	0.0726	0.0701	0.0172	0.0621
class_21	–	–	0.0324	0.4322	0.0307	0.1649
class_22	–	–	0.2867	0.1252	0.0237	0.1106
class_23	–	–	0.0530	0.0932	0.0026	0.0176
class_24	–	–	0.0371	0.0307	0.0013	0.0421
class_25	–	–	0.0038	0.0079	0.0151	0.0097
class_26	–	–	0.0002	0.0032	0.0016	0.0013
class_27	–	–	–	0.0745	0.0004	0.0870
class_28	–	–	–	0.0305	0.0001	0.0319
class_29	–	–	–	0.3377	0.0004	0.2917
class_30	–	–	–	0.2773	0.0138	0.0432
class_31	–	–	–	–	0.1094	0.0604
class_32	–	–	–	–	0.1847	0.0084
class_33	–	–	–	–	0.2989	0.1579
class_34	–	–	–	–	–	0.0555
class_35	–	–	–	–	–	0.1332
class_36	–	–	–	–	–	0.2215
class_37	–	–	–	–	–	0.0661

Table 9: Session-wise performance of FoSSIL on Med FoSSIL-Mixed with MedFormer.

Class	Session 0	Session 1	Session 2	Session 3	Session 4
class_0	0.9708	0.9197	0.9159	0.8546	0.8374
class_1	0.7268	0.7113	0.6460	0.6556	0.6865
class_2	0.7383	0.5017	0.6265	0.6863	0.6322
class_3	0.6131	0.4992	0.5688	0.5894	0.4674
class_4	0.3726	0.2752	0.3008	0.3137	0.2946
class_5	0.6654	0.5540	0.5692	0.6117	0.6156
class_6	0.5733	0.3193	0.4376	0.4662	0.4364
class_7	0.5801	0.5310	0.5484	0.5241	0.5149
class_8	0.8693	0.8658	0.7777	0.8670	0.7484
class_9	0.5762	0.4770	0.4581	0.4691	0.3021
class_10	0.7013	0.6055	0.5097	0.5402	0.5088
class_11	–	0.1599	0.2282	0.1403	0.1801
class_12	–	0.4752	0.7193	0.3971	0.5676
class_13	–	0.0146	0.1974	0.1999	0.1843
class_14	–	0.0033	0.0331	0.0473	0.0491
class_15	–	0.0025	0.0778	0.0733	0.0810
class_16	–	0.5768	0.2846	0.2346	0.1912
class_17	–	0.0101	0.0969	0.0995	0.0897
class_18	–	0.0291	0.0856	0.0788	0.0694
class_19	–	–	0.1619	0.2260	0.2233
class_20	–	–	0.0879	0.1620	0.1251
class_21	–	–	0.0248	0.1559	0.1500
class_22	–	–	0.0446	0.0755	0.0773
class_23	–	–	0.1205	0.1521	0.1847
class_24	–	–	0.0182	0.1185	0.1355
class_25	–	–	–	0.0968	0.0781
class_26	–	–	–	0.0656	0.1789
class_27	–	–	–	0.0127	0.0324
class_28	–	–	–	0.0085	0.3007
class_29	–	–	–	–	0.0153
class_30	–	–	–	–	0.0967
class_31	–	–	–	–	0.0839
class_32	–	–	–	–	0.0108
class_33	–	–	–	–	0.0134
class_34	–	–	–	–	0.0235
class_35	–	–	–	–	0.0190

Table 10: Session-wise performance of FoSSIL on Med FoSSIL-Mixed with SwinUNetr.

Class	Session 1	Session 2	Session 3	Session 4
class_0	0.9177	0.8392	0.7489	0.6863
class_1	0.6182	0.6542	0.7171	0.7281
class_2	0.4002	0.3129	0.3832	0.3683
class_3	0.5285	0.5071	0.5261	0.5420
class_4	0.1691	0.1811	0.1975	0.1695
class_5	0.5035	0.5071	0.5149	0.5352
class_6	0.3991	0.4497	0.4729	0.5168
class_7	0.4998	0.5500	0.5399	0.4988
class_8	0.8356	0.8382	0.8403	0.7952
class_9	0.3164	0.4256	0.4493	0.4320
class_10	0.5107	0.4342	0.3817	0.2558
class_11	0.2323	0.1852	0.2837	0.2882
class_12	0.4545	0.6057	0.5522	0.6818
class_13	0.0313	0.1421	0.1570	0.1548
class_14	0.0031	0.0183	0.0164	0.0185
class_15	0.0241	0.0324	0.0474	0.0506
class_16	0.5017	0.4349	0.2722	0.4420
class_17	0.0325	0.1932	0.2231	0.2093
class_18	0.0126	0.0663	0.0636	0.0616
class_19		0.1410	0.1670	0.1570
class_20		0.0395	0.1898	0.1369
class_21		0.0932	0.1067	0.0940
class_22		0.0536	0.0728	0.0698
class_23		0.1064	0.0754	0.0698
class_24		0.0708	0.1549	0.1368
class_25			0.2741	0.3101
class_26			0.0948	0.1861
class_27			0.0031	0.2009
class_28			0.0213	0.1307
class_29				0.0256
class_30				0.1113
class_31				0.0726
class_32				0.0081
class_33				0.0038
class_34				0.0406
class_35				0.0135

Table 11: Per-class performance of FoSSIL across incremental Sessions on Med Semi-Supervised-FoSSIL benchmark.

Class	Session 0	Session 1	Session 2	Session 3	Session 4	Session 5
class_0	0.9851	0.9693	0.8416	0.7628	0.9355	0.7675
class_1	0.8313	0.6586	0.6283	0.7085	0.0191	0.7044
class_2	0.5754	0.2502	0.4779	0.5395	0.0007	0.4648
class_3	0.8016	0.8832	0.8863	0.8738	0.0249	0.9047
class_4	0.7202	0.7198	0.7110	0.7121	0.0221	0.7107
class_5	0.7864	0.8488	0.8562	0.8303	0.1329	0.7821
class_6	0.7334	0.7154	0.7129	0.7151	0.0813	0.7048
class_7	0.6955	0.7601	0.7130	0.7307	0.0141	0.8014
class_8	0.6646	0.7098	0.7619	0.7817	0.1507	0.7955
class_9	0.7993	0.8283	0.7965	0.8121	0.1180	0.8420
class_10	0.2296	0.5717	0.5166	0.5487	0.0024	0.5056
class_11	0.8144	0.8090	0.8265	0.8340	0.2347	0.8635
class_12	0.6966	0.7603	0.7483	0.7286	0.1190	0.8134
class_13	0.7259	0.6408	0.6532	0.6952	0.0535	0.7216
class_14	0.7112	0.7201	0.7487	0.7391	0.0857	0.7902
class_15	0.7190	0.7123	0.6706	0.5109	0.0264	0.7454
class_16		0.1781	0.1720	0.1706	0.0514	0.1725
class_17		0.0598	0.0871	0.1120	0.0076	0.0931
class_18		0.1094	0.0892	0.0766	0.0044	0.1066
class_19		0.0743	0.0478	0.0679	0.0004	0.0520
class_20		0.0738	0.1123	0.1037	0.0071	0.1008
class_21			0.0488	0.0717	0.0037	0.0943
class_22			0.1646	0.0792	0.0257	0.1129
class_23			0.0906	0.0601	0.0012	0.0932
class_24			0.0387	0.0017	0.0003	0.0077
class_25			0.0319	0.0185	0.0113	0.0205
class_26			0.0141	0.0226	0.0003	0.0242
class_27				0.1818	0.0010	0.2306
class_28				0.0257	0.0003	0.0522
class_29				0.3655	0.0008	0.1170
class_30				0.2980	0.0522	0.1439
class_31					0.2517	0.1313
class_32					0.1708	0.0517
class_33					0.2266	0.1370
class_34						0.1114
class_35						0.1777
class_36						0.2738
class_37						0.1645

Table 12: Performance of SAM and FoSSIL on Natural-FoSSIL benchmark.

Class	Base	SAM Session 1	FoSSIL Session 1	SAM Session 2	FoSSIL Session 2
1	95.29	86.54	88.42	88.18	88.72
2	64.01	35.13	34.07	40.03	39.89
3	92.77	89.24	89.12	82.23	84.21
4	46.02	21.70	24.45	35.56	39.80
5	57.90	16.48	17.70	42.10	40.99
6	48.19	16.31	18.14	17.24	15.33
7	53.94	33.65	37.09	39.76	37.75
8	60.06	28.26	39.18	56.20	57.56
9	90.56	87.93	87.03	84.44	80.50
10	51.58	45.65	38.13	35.77	37.38
11		16.96	16.23	2.17	2.13
12		1.67	1.23	0.03	0.02
13		3.01	3.48	0.19	0.21
14		1.98	0.53	0.00	0.01
15		4.42	3.72	1.01	0.98
16				0.65	1.13
17				28.94	35.42
18				0.00	0.01

Table 13: Per-class performance of FoSSIL on Semi-Supervised Natural-FoSSIL benchmark.

Class	Session 0	Session 1	Session 2	Session 3
road - 0	91.18	82.36	81.35	81.48
sidewalk - 1	48.82	31.20	29.84	27.94
building - 2	86.42	71.03	72.53	67.41
wall - 3	20.84	1.95	1.39	1.14
fence - 4	25.28	2.81	1.41	2.49
traffic light - 5	38.34	10.82	0.53	1.23
traffic sign - 6	31.79	4.12	0.18	0.06
person - 7	28.33	0.59	0.81	0.08
car - 8	81.19	56.30	47.89	51.92
truck - 9	25.40	12.12	9.94	8.96
parking - 10		5.20	7.69	6.23
vegetation - 11		54.76	58.41	53.49
rail track - 12			15.03	18.23
sky - 13			84.62	83.67
motorcycle - 14				17.53
autorickshaw - 15				30.59
bridge/tunnel - 16				24.58

Table 14: FoSSIL performance on Med FoSSIL-Mixed benchmark without Guided noise injection. The results highlight the importance of the proposed mechanism.

Class	Session 0	Session 1	Session 2	Session 3	Session 4
class_0	0.9708	0.7916	0.8477	0.7670	0.8451
class_1	0.7268	0.1127	0.1044	0.2663	0.2423
class_2	0.7383	0.0956	0.1345	0.1047	0.2635
class_3	0.6131	0.1410	0.1116	0.0868	0.2519
class_4	0.3726	0.0598	0.0816	0.0671	0.1605
class_5	0.6654	0.0695	0.0977	0.1460	0.2086
class_6	0.5733	0.1145	0.0471	0.0566	0.1832
class_7	0.5801	0.1303	0.0600	0.0870	0.1193
class_8	0.8693	0.1127	0.1652	0.2193	0.2181
class_9	0.5762	0.1091	0.1763	0.1526	0.0967
class_10	0.7013	0.1082	0.1637	0.0942	0.1573
class_11	–	0.0165	0.0157	0.0182	0.1150
class_12	–	0.0051	0.0312	0.0030	0.2561
class_13	–	0.0042	0.0076	0.0245	0.0463
class_14	–	0.0011	0.0014	0.0201	0.0325
class_15	–	0.0108	0.0061	0.0041	0.0056
class_16	–	0.0193	0.0045	0.0032	0.0367
class_17	–	0.0127	0.0214	0.0067	0.0192
class_18	–	0.0075	0.0129	0.0204	0.0179
class_19	–	–	0.0370	0.0290	0.0247
class_20	–	–	0.0138	0.0096	0.0416
class_21	–	–	0.0075	0.0120	0.0114
class_22	–	–	0.0072	0.0194	0.0107
class_23	–	–	0.0173	0.0103	0.0239
class_24	–	–	0.0083	0.0076	0.0115
class_25	–	–	–	0.0500	0.0793
class_26	–	–	–	0.0719	0.1064
class_27	–	–	–	0.0069	0.0135
class_28	–	–	–	0.0057	0.0124
class_29	–	–	–	–	0.0033
class_30	–	–	–	–	0.0196
class_31	–	–	–	–	0.0047
class_32	–	–	–	–	0.0021
class_33	–	–	–	–	0.0143
class_34	–	–	–	–	0.0024
class_35	–	–	–	–	0.0067

Table 15: FoSSIL performance on Semi-Supervised Natural-FoSSIL benchmark without Pseudo-label refinement.

Class	Session 0	Session 1	Session 2	Session 3
road - 0	91.18	81.79	77.50	78.68
sidewalk - 1	48.82	27.03	26.18	24.36
building - 2	86.42	70.01	66.08	62.76
wall - 3	20.84	3.13	0.75	0.78
fence - 4	25.28	4.06	0.81	0.88
traffic light - 5	38.34	6.35	1.31	4.36
traffic sign - 6	31.79	2.64	0.45	0.45
person - 7	28.33	0.63	0.02	0.03
car - 8	81.19	49.48	43.93	42.89
truck - 9	25.40	11.19	6.89	9.09
parking - 10	–	4.16	4.29	8.90
vegetation - 11	–	51.60	53.31	60.64
rail track - 12	–	–	9.24	12.86
sky - 13	–	–	73.77	78.93
motorcycle - 14	–	–	–	14.33
autorickshaw - 15	–	–	–	17.82
bridge/tunnel - 16	–	–	–	15.08

Table 16: Session-wise performance of MedFormer Vanilla on Med FoSSIL-Mixed benchmark.

Class	Session 0	Session 1	Session 2	Session 3	Session 4
class_0	0.9690	0.9548	0.9648	0.9565	0.9555
class_1	0.6385	3.73e-07	3.73e-07	3.73e-07	7.66e-08
class_2	0.6595	3.34e-07	0.1170	3.34e-07	1.75e-07
class_3	0.7042	7.53e-08	0.0526	7.53e-08	1.68e-08
class_4	0.3711	2.80e-08	1.75e-08	1.75e-08	6.62e-09
class_5	0.5961	1.19e-07	7.77e-08	7.77e-08	1.81e-09
class_6	0.5888	1.99e-08	1.99e-08	1.99e-08	2.48e-09
class_7	0.5652	1.35e-09	0.2320	9.48e-09	0
class_8	0.8350	9.71e-08	9.54e-08	9.54e-08	0
class_9	0.4536	3.97e-08	0	0	0
class_10	0.7170	0	0	0	1.54e-08
class_11	–	0.4038	0	0	0
class_12	–	0.6981	0	0	0
class_13	–	0.4677	0	1.99e-08	0
class_14	–	0.2619	4.47e-07	4.47e-07	0
class_15	–	0.2363	0	0	0
class_16	–	0.8702	0	0	0
class_17	–	0.5323	0	0	0
class_18	–	0.0964	0	0	0
class_19	–	–	0.53	0	0
class_20	–	–	0.58	0	1.79e-08
class_21	–	–	0.5871	0	0
class_22	–	–	0.3948	0	2.55e-08
class_23	–	–	0.4732	0	0
class_24	–	–	0.2474	0	0
class_25	–	–	–	0.72	0
class_26	–	–	–	0.70	0
class_27	–	–	–	0.03	0
class_28	–	–	–	8.94e-08	0
class_29	–	–	–	–	0.2766
class_30	–	–	–	–	0.4008
class_31	–	–	–	–	0.5864
class_32	–	–	–	–	0.2473
class_33	–	–	–	–	0.2792
class_34	–	–	–	–	0.3762
class_35	–	–	–	–	0.1527

Table 17: Performance of GAPS on Natural-FoSSIL benchmark.

Class	Session 0	Session 1	Session 2
road - 0	91.18	81.01	66.32
sidewalk - 1	48.82	29.91	5.24
building - 2	86.42	77.24	62.31
wall - 3	20.84	0.00	0.00
fence - 4	25.28	3.93	0.04
traffic light - 5	38.34	6.01	6.64
traffic sign - 6	31.79	2.64	1.95
person - 7	28.33	6.48	10.33
car - 8	81.19	68.94	45.45
truck - 9	25.40	1.97	0.01
Bridge/tunnel - 10	–	7.48	8.55
Parking - 11	–	22.11	20.01
rail track - 12	–	14.03	6.20
Autorickshaw - 13	–	16.80	8.02
pole - 14	–	12.67	9.69
Bus - 15 (from BDD)	–	–	0.02
Sky - 16 (from BDD)	–	–	48.76
bicycle - 17 (from BDD)	–	–	0.81
person - 7 (repeat from IDD) - 18	–	–	–
car - 8 (repeat from IDD) - 19	–	–	–

Table 18: Performance of C-FSCIL on Med FoSSIL-Disjoint benchmark.

Class	Session 0	Session 1	Session 2
class_0	0.982	0.715	0.606
class_1	0.779	0.438	0.552
class_2	0.735	0.300	0.444
class_3	0.877	0.725	0.563
class_4	0.794	0.669	0.615
class_5	0.802	0.810	0.821
class_6	0.742	0.438	0.471
class_7	0.866	0.519	0.439
class_8	0.844	0.364	0.350
class_9	0.869	0.476	0.521
class_10	0.639	0.271	0.275
class_11	0.854	0.492	0.300
class_12	0.780	0.095	0.145
class_13	0.780	0.288	0.326
class_14	0.752	0.103	0.282
class_15	0.688	0.139	0.247
class_16	–	0.112	0.053
class_17	–	0.067	0.068
class_18	–	0.188	0.231
class_19	–	0.088	0.067
class_20	–	0.092	0.079
class_21	–	–	0.410
class_22	–	–	0.286
class_23	–	–	0.054
class_24	–	–	0.078
class_25	–	–	0.010
class_26	–	–	0.037

Table 19: Performance of GAPS on Semi-Supervised Natural-FoSSIL benchmark.

Class	Session 0	Session 1	Session 2	Session 3
road - 0	91.180	79.102	69.082	67.173
sidewalk - 1	48.822	21.521	3.476	1.202
building - 2	86.420	60.155	46.163	42.604
wall - 3	20.840	1.312	0.489	0.329
fence - 4	25.275	0.000	0.000	0.085
traffic light - 5	38.336	0.000	0.006	0.000
traffic sign - 6	31.790	0.000	0.000	0.000
person - 7	28.332	0.000	0.000	0.000
car - 8	81.190	25.787	20.093	8.848
truck - 9	25.402	8.548	5.093	0.546
parking - 10	–	2.175	1.573	3.339
vegetation - 11	–	38.124	31.426	25.913
rail track - 12	–	–	5.042	5.680
sky - 13	–	–	80.241	69.647
motorcycle - 14	–	–	–	9.447
autorickshaw - 15	–	–	–	9.869
bridge/tunnel - 16	–	–	–	1.035

Table 20: Performance of NNCSL on Med Semi-Supervised-FoSSIL benchmark.

Class	Session 0	Session 1	Session 2	Session 3	Session 4	Session 5
class_0	0.985	0.814	0.925	0.957	0.959	0.962
class_1	0.831	0.068	0.024	0.015	0.011	0.009
class_2	0.575	0.039	0.029	0.011	0.009	0.008
class_3	0.802	0.038	0.022	0.011	0.010	0.006
class_4	0.720	0.017	0.008	0.005	0.001	0.002
class_5	0.786	0.034	0.012	0.017	0.014	0.010
class_6	0.733	0.050	0.029	0.018	0.019	0.015
class_7	0.695	0.005	0.008	0.005	0.004	0.005
class_8	0.665	0.030	0.022	0.009	0.008	0.004
class_9	0.799	0.023	0.008	0.005	0.003	0.008
class_10	0.230	0.042	0.019	0.016	0.013	0.011
class_11	0.814	0.016	0.013	0.004	0.004	0.002
class_12	0.697	0.017	0.025	0.023	0.018	0.017
class_13	0.726	0.013	0.007	0.008	0.003	0.003
class_14	0.711	0.013	0.015	0.014	0.011	0.009
class_15	0.719	0.020	0.018	0.015	0.012	0.010
class_16	–	0.121	0.012	0.007	0.012	0.017
class_17	–	0.071	0.022	0.012	0.016	0.024
class_18	–	0.093	0.015	0.011	0.012	0.013
class_19	–	0.189	0.012	0.008	0.007	0.005
class_20	–	0.044	0.023	0.014	0.015	0.026
class_21	–	–	0.023	0.003	0.002	0.002
class_22	–	–	0.518	0.032	0.018	0.029
class_23	–	–	0.276	0.021	0.019	0.018
class_24	–	–	0.013	0.005	0.003	0.003
class_25	–	–	0.019	0.008	0.008	0.012
class_26	–	–	0.049	0.009	0.004	0.007
class_27	–	–	–	0.166	0.011	0.005
class_28	–	–	–	0.052	0.006	0.005
class_29	–	–	–	0.276	0.016	0.014
class_30	–	–	–	0.106	0.006	0.007
class_31	–	–	–	–	0.023	0.016
class_32	–	–	–	–	0.048	0.013
class_33	–	–	–	–	0.002	0.003
class_34	–	–	–	–	–	0.216
class_35	–	–	–	–	–	0.413
class_36	–	–	–	–	–	0.237
class_37	–	–	–	–	–	0.289

Table 21: Performance of HALO on Semi-Supervised Natural-FoSSIL.

Class	Session 0	Session 1	Session 2	Session 3
road - 0	91.180	0	0	0
sidewalk - 1	48.822	0	0	0
building - 2	86.420	0	0	0
wall - 3	20.840	0	0	0
fence - 4	25.275	0	0	0
traffic light - 5	38.336	0	0	0
traffic sign - 6	31.790	0	0	0
person - 7	28.332	0	0	0
car - 8	81.190	0	0	0
truck - 9	25.402	0	0	0
parking - 10	–	0.545	0	0
vegetation - 11	–	20.873	0	0
rail track - 12	–	–	0.899	0
sky - 13	–	–	27.322	0
motorcycle - 14	–	–	–	1.150
autorickshaw - 15	–	–	–	17.824
bridge/tunnel - 16	–	–	–	2.604

Table 22: Performance of MDIL (domain-incremental) on Med FoSSIL-Disjoint benchmark.

Class	Session 0	Session 1	Session 2
class_0	0.9817	0.9722	0.9672
class_1	0.8514	1.11E-07	2.14E-04
class_2	0.7903	1.83E-06	1.71E-04
class_3	0.8636	9.67E-08	2.55E-03
class_4	0.7851	2.98E-08	2.46E-04
class_5	0.7668	2.15E-05	1.06E-03
class_6	0.7332	6.50E-08	3.15E-04
class_7	0.7982	0	4.80E-04
class_8	0.7869	4.58E-09	1.91E-03
class_9	0.7907	7.99E-08	1.99E-04
class_10	0.4745	9.91E-06	9.41E-04
class_11	0.8172	2.02E-07	5.83E-04
class_12	0.8254	4.26E-07	1.30E-04
class_13	0.8455	2.91E-08	3.45E-04
class_14	0.7842	1.46E-07	7.09E-04
class_15	0.7743	5.08E-07	9.64E-05
class_16	–	0.3737	1.18E-04
class_17	–	0.2560	4.67E-05
class_18	–	0.5987	6.75E-05
class_19	–	0.5410	3.65E-04
class_20	–	0.5216	5.41E-04
class_21	–	–	0.2879
class_22	–	–	0.6180
class_23	–	–	0.5137
class_24	–	–	0.4509
class_25	–	–	0.366
class_26	–	–	0.298

Table 23: Performance of Gen-Replay (generative replay with diffusion) on Med FoSSIL-Disjoint benchmark.

Class	Session 0	Session 1	Session 2
class_0	0.9851	0.9659	0.9604
class_1	0.8313	7.89E-03	0.0180
class_2	0.5754	1.33E-02	2.13E-03
class_3	0.8016	1.63E-03	2.89E-03
class_4	0.7202	1.80E-03	8.80E-03
class_5	0.7864	6.48E-03	0.0229
class_6	0.7334	6.39E-04	7.51E-05
class_7	0.6955	1.89E-04	6.75E-05
class_8	0.6646	6.46E-04	1.39E-04
class_9	0.7993	0.0166	9.66E-04
class_10	0.2296	2.81E-03	1.38E-04
class_11	0.8144	1.91E-04	2.91E-06
class_12	0.6966	7.45E-04	3.64E-05
class_13	0.7259	1.96E-04	9.67E-05
class_14	0.7112	2.76E-03	1.12E-04
class_15	0.7190	1.64E-03	8.19E-04
class_16	–	0.1923	2.72E-04
class_17	–	0.0700	3.44E-04
class_18	–	0.5455	0.00239
class_19	–	0.3336	0.0238
class_20	–	0.3287	6.24E-04
class_21	–	–	0.5066
class_22	–	–	0.6369
class_23	–	–	0.4109
class_24	–	–	0.4150
class_25	–	–	0.166
class_26	–	–	0.430

Table 24: Performance of CLIP-driven method on Med FoSSIL-Mixed benchmark. The model is pre-trained on a large number of classes in the benchmark.

Class	Session 0	Session 1	Session 2	Session 3	Session 4
class_1	0.8812	8.25E-01	7.02E-01	7.02E-01	0.0357
class_2	0.8443	4.60E-01	3.68E-01	3.68E-01	0
class_3	0.8435	8.29E-01	2.18E-01	2.18E-01	0
class_4	0.6985	5.98E-01	1.74E-01	1.74E-01	0
class_5	0.6275	8.51E-02	1.40E-03	1.40E-03	0
class_6	0.2730	8.80E-03	5.00E-04	5.00E-04	0
class_7	0.7567	6.96E-01	1.39E-01	1.39E-01	1.00E-04
class_8	0.8496	1.33E-02	0	0	0
class_9	0.6523	2.33E-01	2.30E-03	2.30E-03	4.00E-04
class_10	0.7396	7.26E-01	0	0	0
class_11	–	0.8083	4.75E-02	4.75E-02	0
class_12	–	0.8337	0	0	0.0142
class_13	–	0.0019	1.90E-03	1.90E-03	0.026
class_14	–	0.0127	0	0	0
class_15	–	0.0121	0	0	0
class_16	–	0.9461	9.41E-01	9.41E-01	0.2405
class_17	–	0.1056	0	0	0.0295
class_18	–	0.3143	2.61E-01	2.61E-01	0.0062
class_19	–	–	0.80	8.02E-01	0.1771
class_20	–	–	0.84	8.42E-01	0
class_21	–	–	0.8551	8.55E-01	0.1799
class_22	–	–	0	0	0
class_23	–	–	0	0	0
class_24	–	–	0.0977	9.77E-02	0
class_25	–	–	–	0.02	0.1365
class_26	–	–	–	0	2.00E-04
class_27	–	–	–	0.01	0
class_28	–	–	–	0.0091	6.00E-04
class_29	–	–	–	–	0.0245
class_30	–	–	–	–	0.1475
class_31	–	–	–	–	0.0498
class_32	–	–	–	–	0.4296
class_33	–	–	–	–	0.4167
class_34	–	–	–	–	0.7290
class_35	–	–	–	–	0.4713