

# Embracing Massive Medical Data

Yu-Cheng Chou, Zongwei Zhou\*, and Alan Yuille

Johns Hopkins University

**Abstract.** As massive medical data become available with an increasing number of scans, expanding classes, and varying sources, prevalent training paradigms—*where AI is trained with multiple passes over fixed, finite datasets*—face significant challenges. First, training AI all at once on such massive data is impractical as new scans/sources/classes continuously arrive. Second, training AI continuously on new scans/sources/classes can lead to catastrophic forgetting, where AI forgets old data as it learns new data, and vice versa. To address these two challenges, we propose an online learning method that enables training AI from massive medical data. Instead of repeatedly training AI on randomly selected data samples, our method identifies the most significant samples for the current AI model based on their data **uniqueness** and prediction **uncertainty**, then trains the AI on these selective data samples. Compared with prevalent training paradigms, our method not only improves data efficiency by enabling training on continual data streams, but also mitigates catastrophic forgetting by selectively training AI on significant data samples that might otherwise be forgotten, outperforming by 15% in Dice score for multi-organ and tumor segmentation. The code is available at <https://github.com/MrGiovanni/OnlineLearning>

**Keywords:** Online Learning · Catastrophic Forgetting.

## 1 Introduction

Massive medical data are becoming publicly available [29,3], but are we ready for training AI models on increasing number of scans, varying sources, and expanding classes? Prevalent training paradigms face significant challenges, as training AI all at once on such massive data is impractical [28]. First, these paradigms are largely designed for fixed, finite datasets. Second, they exhibit signs of catastrophic forgetting when adapted to new scans, sources, or classes [30].

To address these challenges, an ideal training paradigm must meet four requirements: **Req. 1:** handling incomplete annotations, as a combination of public medical datasets is often only partially labeled due to the high cost of annotating medical images at the voxel level [15]. Existing methods [40,22] require fully labeled inputs. **Req. 2:** learning new scans/sources without forgetting old ones, where prevalent training paradigms that rely on ‘epoch’—a complete, shuffled pass through the entire dataset—are impractical [42]. Existing methods [36,40]

---

\* Correspondence to: Zongwei Zhou ([zzhou82@jh.edu](mailto:zzhou82@jh.edu))

require AI to see the dataset multiple times. **Req. 3:** handling new classes without repeated retraining, as most medical images are not fully labeled yet but will be in the future. Existing methods [27,9] discard the possibility of new classes. **Req. 4:** training AI with manageable resources, as clinical environments usually lack large storage and computational capacities. Existing methods [37,14] require extensive memory for data selection and class extension.

To meet these requirements, we explore online learning [22,7]. **First**, we use text encoding [20] to segment classes based on their text descriptions, allowing us to handle new classes and tasks within partially labeled datasets by simply modifying the text descriptions (Reqs. 1,3). **Second**, we store recent samples as new data becomes available, enabling training continuously from new data without revisiting old data and avoiding the need for complete passes through the entire dataset (Req. 2, §2.1). **Third**, to store the best possible training data, we propose to select samples based on their uniqueness, which prevents catastrophic forgetting by preserving samples from different time periods and sources (Reqs. 2,4, §2.2). **Fourth**, inspired by human learning that focuses on challenging tasks [1], we propose a simple yet effective method that deliberately learns the samples with high uncertainty to improve performance without the need for extra computational resources (Reqs. 4, §2.3).

We have examined the proposed method on a large-scale single-site private dataset and a sequential-site dataset composed of 16 public datasets [20]. On the single-site dataset, extensive experiments demonstrate that our method achieves comparable performance to the prevalent training paradigm that trains the model over multiple passes (82.2% vs. 82.6%). More importantly, the results in Table 1 show that our method enhances data efficiency by enabling training continuously from new data without revisiting old data (§3.2). On the sequential-site dataset with varying sources of data (Table 2), we observe a 6% performance increase by effectively selecting unique data compared to directly storing recent samples. Notably, the results in Figure 2 and Figure 3 demonstrate that our method successfully mitigates catastrophic forgetting by enhancing data diversity and preserving previous knowledge (§3.3). Furthermore, by deliberately learning samples with high uncertainty, our method boosts performance by 9% that even outperforms the prevalent training paradigm (§3.4).

With advancements in automatic annotation [13,38,17] and the generation of vast synthetic data [11,8,16], the future of AI seems destined for continuous streams of massive data. Embracing massive medical data, we are among the first to take proactive action that can potentially lead to transformative shifts in AI development. In summary, the main contributions of this work are as follows:

1. We show that storing recent samples in the data stream can improve data efficiency, allowing us to train on continual data streams to achieve comparable performance to the prevalent training paradigm (Table 1).
2. We demonstrate the effectiveness of the data deduplication for adapting varying sources, which increases the performance by 6% on Dice score for multi-organ and tumor segmentation (Table 2).

3. We follow the human learning pattern to deliberately learn significant samples and further boost performance by 9%, even outperforming the prevalent training paradigm (Table 2).
4. We prove the capability of mitigating catastrophic forgetting with both data deduplication and deliberate learning by retaining the previous knowledge (Figure 2 and Figure 3).

**Related Work.** To mitigate catastrophic forgetting, significant research has been conducted in the domain of natural images. Regularization-based methods constrain model flexibility, typically through techniques applied to weights [6] and gradients [7], or via knowledge distillation focused on output logits [5,34] and intermediate features [43]. Despite their utility, these methods often fail to ensure optimal performance on challenging tasks, particularly without stored exemplars. Expandable models address this by either incrementally growing with new tasks [18,41] or maintaining a common backbone model adaptable through small task-specific adaptation blocks [23]. However, this often leads to inflated model sizes and bloated network architectures. In contrast, rehearsal techniques involve storing and replaying a small subset of training samples [4,21], embedded features [12], or generators [25] from prior tasks. Nonetheless, these methods violate the streaming data setting that our study adheres to.

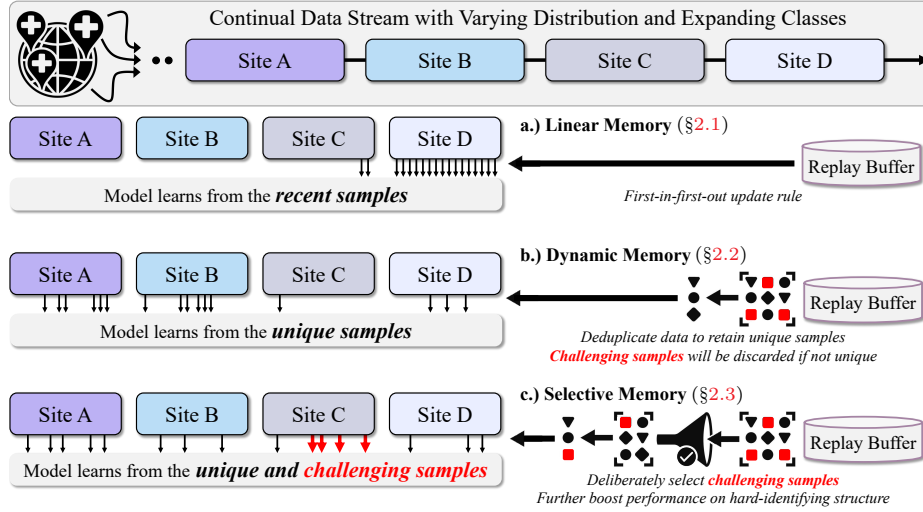
## 2 Method

To handle continual data streams and dynamically adapt without revisiting old data and forgetting acquired knowledge, we propose utilizing a linear memory to store recent samples in the continual data stream (§2.1), dynamic memory to deduplicate the stored samples (§2.2), and selective memory to learn from the significant samples to further improve the performance (§2.3).

### 2.1 Linear Memory

Prevalent training paradigms typically rely on fixed-size datasets, characterized by their finiteness, immutability, and ready availability. Leveraging these attributes, samples within such datasets can be conveniently indexed, shuffled, and accessed throughout the training process. The prevalent training paradigm often needs iterating over the datasets multiple times. In sharp contrast, online learning necessitates a departure from this paradigm, embracing a data stream potentially of infinite length. Within this context, data retrieval from a streaming source at any given moment  $t$  yields the current sample  $x_t$ , with future samples inaccessible and past samples retrievable only if stored upon acquisition.

Addressing the challenges, we propose a straightforward solution by leveraging the concept of Linear Memory, tasked with storing a limited number of recent samples. This notion draws inspiration from experience replay [19], a technique widely employed in reinforcement learning [2,24,33], supervised continual learning [10,31], and self-supervised learning [28]. As shown in Figure 1-(a), the linear



**Fig. 1. Different Training Method.** Linear memory stores only a few recent samples, causing significant forgetting. Dynamic memory adapts to varying data distributions by retaining unique samples, while selective memory further identifies and selects challenging samples, including those that might be duplicated, ensuring they are not missed by dynamic memory (§3.4).

memory serves to store the recent samples from streaming data sources for the training pipeline. Specifically, incoming streaming data is seamlessly integrated into the replay buffer  $\mathcal{B}$  with memory size  $N$ , replacing the oldest samples according to a first-in-first-out update rule. At the same time, mini-batches  $x$  of training data can be generated at any time by random sampling from the buffer. This setup eliminates the need for the training pipeline to revisit previous samples, enabling it to leverage streaming data sources potentially of infinite length. Simultaneously, linear memory facilitates the reuse of samples through multiple sampling instances, thereby reducing the overall data cost. As a result, data usage is determined by sampling rate  $S$ , defined as the ratio between the number of mini-batches and those acquired from the streaming source.

## 2.2 Dynamic Memory

Medical data obtained from hospitals often exhibit duplicated patterns due to similar sampling strategies or data distribution. Initially, distributions within a period of time adhere to the same scanning protocol and annotation policy, resulting in a high degree of duplication. However, these distributions may shift over time. This varying nature stands in sharp contrast to the static data used in the prevalent training paradigm, where models are trained on single, meticulously curated datasets. Take AbdomenAtlas dataset [17] for instance, which

allows for random sampling images from a collection of 25 classes. However, the continual data stream in online learning setups often challenges this assumption.

To address this challenge, we propose Dynamic Memory, a method designed to optimize the linear memory by retaining only unique samples and actively reducing duplication, as illustrated in Figure 1-(b). Consider a replay buffer  $\mathcal{B}$  containing  $N$  samples, each with a corresponding embedding  $z_i$ . When adding a new sample  $x_t$  to  $\mathcal{B}$ , we calculate the cosine similarity between all samples in  $\mathcal{B}$ . This helps us identify and discard the sample  $x_{i^*}$ , where  $i^*$  is the index in  $\mathcal{B}$  with the highest cosine similarity between embeddings  $z_i$  and  $z_j$  for  $i, j \in \mathcal{B}$ . We then employ a moving average to track embedding over iterations. Note that because each sample represents a small region after data augmentation, rather than the entire input, its embedding shows significant differences. By removing duplicated samples from the memory, the correlation between samples decreases, thereby increasing the data diversity in the memory.

### 2.3 Selective Memory

In the realm of varying data streams, we encounter numerous challenges that, intriguingly, parallel the concept of deliberate practice, reminiscent of human learning patterns. Different from common practice that simply repetition of certain actions, deliberate practice involves first identifying the problem, and then systematically targeting through specific exercises to correct mistakes [1]. Here, machine learning moves beyond its traditional passive role of processing fixed data, aligning more closely with the active learning mechanisms observed in humans. Inspired by human learning patterns, we prioritize samples based on their significance, constructing a memory that emphasizes the most challenging samples while discarding the easier ones. As a result, our goal is to focus on samples with greater uncertainty and structures carrying heavier penalties.

To achieve this, we introduce Selective Memory (SM). As shown in Figure 1-(c), given a prediction  $y$  and the corresponding ground truth  $g$ , we first calculate the penalty  $\alpha_c$  of each structure:

$$\alpha_c = \begin{cases} 1, & \text{for } 0 < \hat{\alpha}_c < 1 \\ \hat{\alpha}_c, & \text{else} \end{cases}, \text{ where } \hat{\alpha}_c = \text{Norm}\left(\frac{\sum_c^m S_c}{S_c}\right) * m, \quad (1)$$

and  $S_c$  is the size of the corresponding structure,  $m$  is the number of structure in  $g$ . We then apply the obtained penalty  $\alpha$  on the entropy calculation  $L_{BCE}$  to give the hard-identifying structure more penalty.

To calculate uncertainty  $U_x$ , we resort to the entropy. In information theory, entropy is a measure of uncertainty or unpredictability associated with a random variable [35]. It quantifies the amount of information contained in a message or a set of outcomes, which inherently matches the idea of deliberate practice to identify problems (knowledge with more information) and strategically target them. As a result, we apply  $\alpha$  weighted  $L_{BCE}$  to calculate uncertainty  $U_x$  and only discard samples from the ones under top  $K\%$  entropy when calculating the

**Table 1. Data Efficiency.** The results demonstrate that the linear memory trained on continual data streams achieves comparable performance to the prevalent training paradigm that trains models repeatedly with 100 epochs. Linear memory enables training without the need to revisit old data, thereby enhancing data efficiency. Please see [supp. material](#) for full results.

Single-Site Dataset				
Strategy	Linear Memory	Linear Memory	Linear Memory	Repeatedly
Epochs	-	-	-	100
Sampling Rate $S$	100	100	100	-
Memory Size $N$	64	128	256	-
Average Dice	0.8217	0.8222	0.8225	0.8260

cosine similarity between all pairs of samples:

$$i, j \in \mathcal{B}_k, \text{ where } \mathcal{B}_k = \mathcal{B} \setminus \text{Top}K(\{U_x\}). \quad (2)$$

### 3 Experiment & Result

#### 3.1 Experimental Setting

**Dataset.** We adopt two large-scale CT datasets in our experiments, including a single-site private dataset [26,39] and a sequential-site dataset [20]. (1) Our **single-site** dataset comprises abdominal CT scans with per-pixel annotation, encompassing 15 classes of abdominal organs, all sourced from a single hospital with a consistent distribution. We split the dataset randomly into 2,101 training cases and 516 testing cases. (2) Our **sequential-site** dataset consists of 16 partially labeled sub-datasets ( $D_d, d \in [1, 16]$ ), collectively offering 32 classes. This combination of sub-datasets incorporates various annotation policies and scanning protocols, mirroring the varying distributions of a data stream. We adhered to the data splitting protocol in [20], resulting in 2,100 training cases and 583 testing cases. It is worth noting that we did not shuffle the data in any of our experiments for the data stream setting.

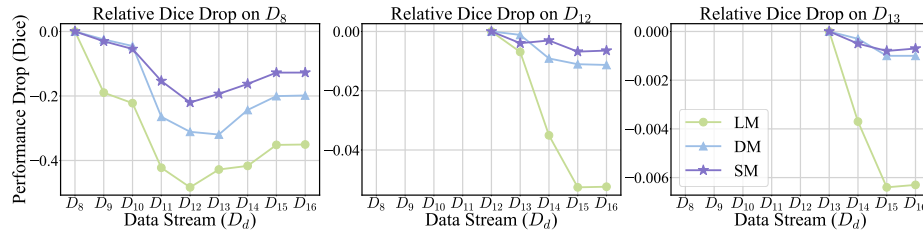
**Implementation and baseline.** For adapting partially labeled sub-datasets, we adopt the CLIP-Driven Universal Model [20] for multi-organ and tumor segmentation. U-Net [32] is chosen as the backbone, and we follow the default training setting in [20] with the standard binary cross-entropy loss and the Dice loss. We adopt the Dice score as the metric to evaluate the results in our experiments. Note that the prevalent training paradigm trains models repeatedly with 100 epochs, which violates the streaming setting in our experiments.

#### 3.2 Linear Memory Enables Training Without Revisiting Old Data

We investigate the effectiveness of linear memory by training on the continual data stream. As shown in Table 1, we trained the model with and without linear

**Table 2. Dynamic Adaptation.** Under the varying distributions in the streaming source, Dynamic Memory (DM) and Selective Memory (SM) enable the identification of the significant samples and thereby enhance the segmentation performance. Please see [supp. material](#) for full results.

Sequential-Site Dataset [20]						
Strategy	LM	DM	SM	SM	SM	Repeatedly
Epochs	-	-	-	-	-	100
Sampling Rate $S$	100	100	100	100	100	-
Memory Size $N$	128	128	128	128	128	-
Top K%	-	-	12.5%	25%	50%	-
Tumor Average Dice	0.2559	0.3125	0.3436	<b>0.3878</b>	0.3426	0.3520
Organ Average Dice	0.3471	0.4066	0.4805	<b>0.4958</b>	0.4834	0.4783
Average Dice	0.3272	0.3860	0.4505	<b>0.4722</b>	0.4525	0.4506

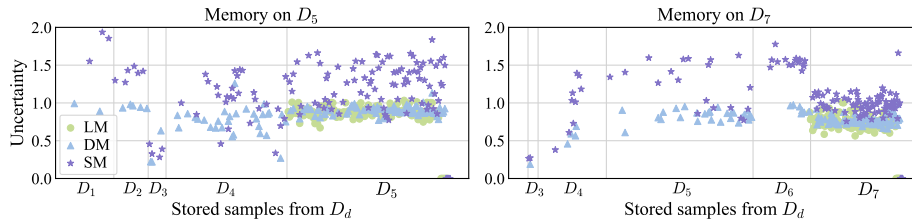


**Fig. 2. Catastrophic Forgetting.** To evaluate forgetting, we calculate the relative Dice drop after training on the incoming sub-datasets. Both DM and SM store samples from previous sub-datasets, thereby alleviating forgetting observed with LM.

memory on our single-site dataset with a consistent distribution. Linear memory notably enhances data efficiency in the continual data stream setting without the need to revisit old data. Additionally, under the same amount of updates, the linear memory with a sampling rate  $S = 100$  achieved similar performance to the prevalent training paradigm. Table 1 also shows that by increasing memory size  $N$ , the performance raised slightly due to increasing samples in the memory. However, this enhancement comes with a notable memory overhead. Consequently, we set  $N = 128$  for a cost-accuracy trade-off.

### 3.3 Dynamic Memory Mitigates Catastrophic Forgetting

Compared to our single-site dataset, the sequential-site dataset [20] features varying distributions in scanning protocol, annotation policy, and slicing scope. This variability leads to significant performance degradation, as illustrated in Table 2. While the naive linear memory can somewhat mitigate this issue, it fails to match the performance of the prevalent training paradigm (32.7% vs. 45.1%) and still experiences substantial forgetting. Specifically, linear memory shows a marked performance decline on  $D_d$  when learning from the incoming sub-datasets, as shown in Figure 2. Conversely, the dynamic memory, which retains unique samples from each sub-dataset, demonstrates considerable improvements



**Fig. 3. Diverse Memory.** We visualize the memory to demonstrate the diversity of stored samples from previous  $D_d$ . Both DM and SM can retain the samples from previous sub-datasets. SM can further identify samples with higher uncertainty.

(Figure 3). This method boosts performance by 6% (Table 2) and effectively mitigates catastrophic forgetting, as evidenced in Figure 2.

### 3.4 Selective Memory Outperforms Prevalent Training Paradigms

Although dynamic memory can retain unique samples, the performance remains unsatisfactory. By deliberately targeting the samples with higher uncertainty, selective memory significantly increases the performance that even outperforms the prevalent training paradigm (47.2% vs. 45.1%, Table 2) and is able to identify small structures such as Adrenal Gland (detailed in [supp. material](#)). Especially, as shown in Figure 3 (right), compared to dynamic memory, selective memory stores the high uncertainty samples of  $D_6$  when training on  $D_7$ . This illustrates that selective memory can identify challenging but potentially duplicated samples that are ignored by dynamic memory. Last, we ablate this method with different  $K\%$  settings and set  $K = 25\%$  to ensure a balance of diversity and uncertainty that maximizes the model’s performance.

## 4 Discussion & Conclusion

As the volume of public medical data continues to expand, prevalent training paradigms struggle to handle the continual streams of massive medical data seamlessly. Our study, as the pioneer in anticipating this trend and proactively taking action in response, introduced linear memory and dynamic memory to address data inefficiencies while retaining significant samples to effectively accommodate varying distributions. Additionally, inspired by human learning patterns, we introduced selective memory to further actively focus on challenging samples. We experimentally show that our method not only surpasses the prevalent training paradigm but also significantly mitigates catastrophic forgetting. In conclusion, we present potential solutions tailored to evolving clinical scenarios, and we believe that further exploration in this direction is required to continually learn by more fine-grained and, even, no annotation.



**Acknowledgments.** This work was supported by the Lustgarten Foundation for Pancreatic Cancer Research and the Patrick J. McGovern Foundation Award.

**Disclosure of Interests.** The authors have no competing interests to declare that are relevant to the content of this article.

## References

1. Anders Ericsson, K.: Deliberate practice and acquisition of expert performance: a general overview. *Academic emergency medicine* **15**(11), 988–994 (2008)
2. Andrychowicz, M., Wolski, F., Ray, A., Schneider, J., Fong, R., Welinder, P., McGrew, B., Tobin, J., Pieter Abbeel, O., Zaremba, W.: Hindsight experience replay. *NeurIPS* **30** (2017)
3. Blankemeier, L., Cohen, J.P., Kumar, A., Van Veen, D., Gardezi, S.J.S., Paschali, M., Chen, Z., Delbrouck, J.B., Reis, E., Truys, C., et al.: Merlin: A vision language foundation model for 3d computed tomography. *arXiv preprint arXiv:2406.06512* (2024)
4. Buzzega, P., Boschini, M., Porrello, A., Abati, D., Calderara, S.: Dark experience for general continual learning: a strong, simple baseline. *NeurIPS* **33**, 15920–15930 (2020)
5. Castro, F.M., Marín-Jiménez, M.J., Guil, N., Schmid, C., Alahari, K.: End-to-end incremental learning. In: *ECCV*. pp. 233–248 (2018)
6. Chaudhry, A., Dokania, P.K., Ajanthan, T., Torr, P.H.: Riemannian walk for incremental learning: Understanding forgetting and intransigence. In: *ECCV*. pp. 532–547 (2018)
7. Chaudhry, A., Ranzato, M., Rohrbach, M., Elhoseiny, M.: Efficient lifelong learning with a-gem. *arXiv preprint arXiv:1812.00420* (2018)
8. Chen, Q., Chen, X., Song, H., Xiong, Z., Yuille, A., Wei, C., Zhou, Z.: Towards generalizable tumor synthesis. In: *CVPR* (2024)
9. González, C., Ranem, A., Pinto dos Santos, D., Othman, A., Mukhopadhyay, A.: Lifelong nnu-net: a framework for standardized medical continual learning. *Scientific Reports* **13**(1), 9381 (2023)
10. Hsu, Y.C., Liu, Y.C., Ramasamy, A., Kira, Z.: Re-evaluating continual learning scenarios: A categorization and case for strong baselines. *arXiv preprint arXiv:1810.12488* (2018)
11. Hu, Q., Chen, Y., Xiao, J., Sun, S., Chen, J., Yuille, A.L., Zhou, Z.: Label-free liver tumor segmentation. In: *CVPR*. pp. 7422–7432 (2023)
12. Iscen, A., Zhang, J., Lazebnik, S., Schmid, C.: Memory-efficient incremental learning through feature adaptation. In: *ECCV*. pp. 699–715. Springer (2020)
13. Jaus, A., Seibold, C., Hermann, K., Walter, A., Giske, K., Haubold, J., Kleesiek, J., Stiefelhagen, R.: Towards unifying anatomy segmentation: Automated generation of a full-body ct dataset via knowledge aggregation and anatomical guidelines. *arXiv preprint arXiv:2307.13375* (2023)
14. Ji, Z., Guo, D., Wang, P., Yan, K., Lu, L., Xu, M., Wang, Q., Ge, J., Gao, M., Ye, X., et al.: Continual segment: Towards a single, unified and non-forgetting continual segmentation model of 143 whole-body organs in ct scans. In: *CVPR*. pp. 21140–21151 (2023)
15. Kang, M., Li, B., Zhu, Z., Lu, Y., Fishman, E.K., Yuille, A., Zhou, Z.: Label-assemble: Leveraging multiple datasets with partial labels. In: *ISBI*. pp. 1–5. IEEE (2023)

16. Lai, Y., Chen, X., Wang, A., Yuille, A., Zhou, Z.: From pixel to cancer: Cellular automata in computed tomography. arXiv preprint arXiv:2403.06459 (2024)
17. Li, W., Yuille, A., Zhou, Z.: How well do supervised models transfer to 3d image segmentation? In: ICLR. vol. 1 (2024)
18. Li, X., Zhou, Y., Wu, T., Socher, R., Xiong, C.: Learn to grow: A continual structure learning framework for overcoming catastrophic forgetting. In: ICML. pp. 3925–3934. PMLR (2019)
19. Lin, L.J.: Reinforcement learning for robots using neural networks. Carnegie Mellon University (1992)
20. Liu, J., Zhang, Y., Chen, J.N., Xiao, J., Lu, Y., A Landman, B., Yuan, Y., Yuille, A., Tang, Y., Zhou, Z.: Clip-driven universal model for organ segmentation and tumor detection. In: CVPR. pp. 21152–21164 (2023)
21. Liu, Y., Schiele, B., Sun, Q.: Rmm: Reinforced memory management for class-incremental learning. NeurIPS **34**, 3478–3490 (2021)
22. Lopez-Paz, D., Ranzato, M.: Gradient episodic memory for continual learning. NeurIPS **30** (2017)
23. Mallya, A., Davis, D., Lazebnik, S.: Piggyback: Adapting a single network to multiple tasks by learning to mask weights. In: ECCV. pp. 67–82 (2018)
24. Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A.A., Veness, J., Bellemare, M.G., Graves, A., Riedmiller, M., Fidjeland, A.K., Ostrovski, G., et al.: Human-level control through deep reinforcement learning. nature **518**(7540), 529–533 (2015)
25. Ostapenko, O., Puscas, M., Klein, T., Jahnichen, P., Nabi, M.: Learning to remember: A synaptic plasticity driven framework for continual learning. In: CVPR. pp. 11321–11329 (2019)
26. Park, S., Chu, L., Fishman, E., Yuille, A., Vogelstein, B., Kinzler, K., Horton, K., Hruban, R., Zinreich, E., Fouladi, D.F., et al.: Annotated normal ct data of the abdomen for deep learning: Challenges and strategies for implementation. Diagnostic and interventional imaging **101**(1), 35–44 (2020)
27. Perkonigg, M., Hofmanninger, J., Herold, C.J., Brink, J.A., Pianykh, O., Prosch, H., Langs, G.: Dynamic memory to alleviate catastrophic forgetting in continual learning with medical imaging. Nature communications **12**(1), 5678 (2021)
28. Purushwalkam, S., Morgado, P., Gupta, A.: The challenges of continuous self-supervised learning. In: ECCV. pp. 702–721. Springer (2022)
29. Qu, C., Zhang, T., Qiao, H., Liu, J., Tang, Y., Yuille, A., Zhou, Z.: Abdomenatlas-8k: Annotating 8,000 abdominal ct volumes for multi-organ segmentation in three weeks. In: NeurIPS. vol. 21 (2023)
30. Robins, A.: Catastrophic forgetting, rehearsal and pseudorehearsal. Connection Science **7**(2), 123–146 (1995)
31. Rolnick, D., Ahuja, A., Schwarz, J., Lillicrap, T., Wayne, G.: Experience replay for continual learning. NeurIPS **32** (2019)
32. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: MICCAI. pp. 234–241. Springer (2015)
33. Schaul, T., Quan, J., Antonoglou, I., Silver, D.: Prioritized experience replay. arXiv preprint arXiv:1511.05952 (2015)
34. Schwarz, J., Czarnecki, W., Luketina, J., Grabska-Barwinska, A., Teh, Y.W., Pascanu, R., Hadsell, R.: Progress & compress: A scalable framework for continual learning. In: ICML. pp. 4528–4537. PMLR (2018)
35. Seidenfeld, T.: Entropy and uncertainty. Philosophy of Science **53**(4), 467–491 (1986)
36. Smith, J.S., Tian, J., Halbe, S., Hsu, Y.C., Kira, Z.: A closer look at rehearsal-free continual learning. In: CVPR. pp. 2409–2419 (2023)

37. Tiwari, R., Killamsetty, K., Iyer, R., Shenoy, P.: Gcr: Gradient coreset based replay buffer selection for continual learning. In: CVPR. pp. 99–108 (2022)
38. Wasserthal, J., Breit, H.C., Meyer, M.T., Pradella, M., Hinck, D., Sauter, A.W., Heye, T., Boll, D.T., Cyriac, J., Yang, S., et al.: Totalsegmentator: robust segmentation of 104 anatomic structures in ct images. *Radiology: Artificial Intelligence* **5**(5) (2023)
39. Xia, Y., Yu, Q., Chu, L., Kawamoto, S., Park, S., Liu, F., Chen, J., Zhu, Z., Li, B., Zhou, Z., et al.: The felix project: Deep networks to detect pancreatic neoplasms. *medRxiv* (2022)
40. Yan, S., Xie, J., He, X.: Der: Dynamically expandable representation for class incremental learning. In: CVPR. pp. 3014–3023 (2021)
41. Yoon, J., Yang, E., Lee, J., Hwang, S.J.: Lifelong learning with dynamically expandable networks. *arXiv preprint arXiv:1708.01547* (2017)
42. Zhang, Y., Li, X., Chen, H., Yuille, A.L., Liu, Y., Zhou, Z.: Continual learning for abdominal multi-organ and tumor segmentation. In: MICCAI. pp. 35–45. Springer (2023)
43. Zhu, K., Zhai, W., Cao, Y., Luo, J., Zha, Z.J.: Self-sustaining representation expansion for non-exemplar class-incremental learning. In: CVPR. pp. 9296–9305 (2022)

**Table 3. Dynamic Adaptation.** Under the varying distribution in the streaming source, the proposed Dynamic Memory (DM) and Selective Memory (SM) enable the identification of the critical samples and thereby enhance the segmentation performance. Specifically, SM significantly outperforms other methods on complex structures such as the Esophagus and small structures such as the Adrenal Gland.

Sequential-Site Dataset [20]						
Strategy	LM	DM	SM	SM	SM	Prevalent
Sampling Rate $S$	100	100	100	100	100	-
Memory Size $N$	128	128	128	128	128	-
Top K%	-	-	12.5%	25%	50%	-
Spleen	0.9506	0.9543	0.9553	0.9449	0.9542	0.9268
Right Kidney	0.8961	0.9236	0.9239	0.9202	0.9122	0.9189
Left Kidney	0.8972	0.9172	0.9145	0.9138	0.9052	0.9149
Gall Bladder	0.2205	0.6747	0.5367	0.5906	0.6325	0.3671
Esophagus	0.0004	0.1588	0.4376	0.4044	0.4537	0.0395
Liver	0.9615	0.9673	0.9668	0.9665	0.9632	0.9630
Stomach	0.5802	0.7705	0.7313	0.7890	0.7467	0.7785
Aorta	0.4895	0.6063	0.6751	0.5687	0.6633	0.7676
Postcava	0.2098	0.5373	0.6191	0.5444	0.5239	0.6433
Vein	0.0000	0.0370	0.0000	0.0000	0.0000	0.2310
Pancreas	0.6669	0.8028	0.8249	0.8112	0.7766	0.7376
Right Adrenal Gland	0.0000	0.0000	0.5830	0.5942	0.4307	0.0000
Left Adrenal Gland	0.0000	0.0000	0.5625	0.5126	0.4716	0.0000
Duodenum	0.1204	0.4293	0.4000	0.4987	0.5251	0.3599
Hepatic Vessel	0.4678	0.5571	0.5769	0.5524	0.5731	0.5063
Right Lung	0.7002	0.4919	0.5432	0.6112	0.5529	0.7617
Left Lung	0.8771	0.6187	0.7965	0.7162	0.7859	0.9102
Colon	0.0008	0.0003	0.0185	0.0352	0.0187	0.4986
Intestine	0.0000	0.0849	0.1081	0.2765	0.3648	0.4610
Rectum	0.0000	0.0000	0.0000	0.0000	0.0102	0.0000
Bladder	0.6383	0.5397	0.7316	0.7701	0.7722	0.7517
Prostate	0.0009	0.0923	0.1066	0.2276	0.0471	0.0000
Left Head of Femur	0.0000	0.0000	0.0000	0.0772	0.0000	0.0000
Right Head of Femur	0.0000	0.0000	0.0000	0.0690	0.0000	0.4190
Celiac Truck	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Kidney Tumor	0.2417	0.2936	0.2483	0.3258	0.3204	0.3781
Kidney Cyst	0.0268	0.0059	0.0729	0.2920	0.0056	0.1751
Liver Tumor	0.3579	0.5753	0.5755	0.6059	0.5817	0.6219
Pancreas Tumor	0.1643	0.2823	0.2957	0.3435	0.3560	0.2058
Hepatic Vessel Tumor	0.5717	0.6659	0.6516	0.6618	0.6713	0.5930
Lung Tumor	0.1910	0.1845	0.2862	0.1981	0.1516	0.3233
Colon Tumor	0.2377	0.1797	0.2751	0.2874	0.3106	0.1667
Tumor Average Dice	0.2559	0.3125	0.3436	<b>0.3878</b>	0.3426	0.3520
Organ Average Dice	0.3471	0.4066	0.4805	<b>0.4958</b>	0.4834	0.4783
Average Dice	0.3272	0.3860	0.4505	<b>0.4722</b>	0.4525	0.4506

**Table 4. Data Efficiency.** By integrating linear memory into the prevalent training paradigm, we enable training on continual data streams without the need to revisit old data, thereby enhancing data efficiency. The results demonstrate that the linear memory trained on continual data streams achieves comparable performance to the prevalent training paradigm.

Proprietary Dataset				
Strategy	Linear Memory	Linear Memory	Linear Memory	Prevalent
Sampling Rate $S$	100	100	100	-
Memory Size $N$	64	128	256	-
Aorta	0.8865	0.8893	0.8890	0.8848
Right Adrenal Gland	0.7530	0.7534	0.7501	0.7491
Left Adrenal Gland	0.7001	0.6980	0.6962	0.6964
Celiac Truck	0.5403	0.5428	0.5385	0.5610
Colon	0.6915	0.6946	0.6992	0.7196
Duodenum	0.7907	0.7903	0.7878	0.7896
Gall Bladder	0.8866	0.8888	0.8860	0.8886
Postcava	0.8129	0.8088	0.8161	0.8164
Right Kidney	0.9535	0.9536	0.9551	0.9527
Left Kidney	0.9473	0.9477	0.9475	0.9460
Liver	0.9715	0.9712	0.9717	0.9708
Pancreas	0.8699	0.8683	0.8703	0.8688
Intestine	0.6047	0.6088	0.6144	0.6375
Spleen	0.9664	0.9661	0.9663	0.9641
Stomach	0.9512	0.9509	0.9506	0.9453
Average Dice	0.8217	0.8222	0.8225	0.8260