Journal of Machine Learning Research 30 (2025) 1-12 Submitted 30/25; Revised x/xx; Published x/xx

Context-guided Prompt-learning for Continual WSI Classification

Giulia Corso* GIULIA.CORSO@UNIMORE.IT
Francesca Miccolis* FRANCESCA.MICCOLIS@UNIMORE.IT
Angelo Porrello ANGELO.PORRELLO@UNIMORE.IT
Federico Bolelli ✉ FEDERICO.BOLELLI@UNIMORE.IT
Simone Calderara SIMONE.CALDERARA@UNIMORE.IT
Elisa Ficarra ELISA.FICARRA@UNIMORE.IT
University of Modena and Reggio Emilia

Editor: TBD

Abstract

Whole Slide Images (WSIs) are crucial in histological diagnostics, providing high-resolution insights for analyzing cellular structures. In addition to challenges like the gigapixel scale of WSIs and the lack of pixel-level annotations, privacy restrictions further complicate their analysis. For instance, in a hospital network, different facilities need to collaborate on WSI analysis without the possibility of sharing sensitive patient data. A more practical and secure approach involves sharing models capable of continual adaptation to new data. However, without proper measures, catastrophic forgetting can occur. Traditional continual learning techniques rely on storing previous data, which violates privacy restrictions. To address this issue, this paper introduces Context Optimization Multiple Instance Learning (CooMIL), a rehearsal-free continual learning framework designed explicitly for WSI analysis. It employs a WSI-specific prompt learning procedure to adapt classification models across tasks, efficiently preventing catastrophic forgetting. Evaluated on four public WSI datasets from TCGA projects, our model significantly outperforms state-of-the-art methods within the WSI-based continual learning framework. The source code is available at `https://github.com/FrancescaMiccolis/CooMIL`.

Keywords: WSI, Multi-instance Learning, Prompt Learning, Continual Learning

1 Introduction

Whole Slide Images (WSIs) are valuable tools in digital pathology and clinical diagnostics. In addition to the vast dimensions and lack of precise pixel-level annotations (Lu et al., 2020; Huang et al., 2022; Javed et al., 2020; Van der Laak et al., 2021)—WSIs are typically subject to privacy restrictions, which hinder data sharing and collaboration. Continual Learning (CL) offers a solution by allowing models to learn incrementally from data distributed across multiple healthcare institutions, while respecting patient privacy and data governance policies. Indeed, fine-tuning even large-scale models on new datasets often leads to *catastrophic forgetting*, where the model forgets previously learned information (Robins, 1995; Boschini et al., 2022b). This issue is critical in WSI analysis, where tissue subtypes

*Equal contribution. Authors are allowed to list their name first on their CVs.

and treatment protocols evolve rapidly. CL aims to mitigate this problem and enhance models' adaptability to new datasets and tasks. Among various continual learning methodologies such as regularization (Aljundi et al., 2018; Zenke et al., 2017) and architectural strategies (Rusu et al., 2016), only rehearsal-based models seem to be effective against catastrophic forgetting in WSI analysis (Huang et al., 2023a). However, such approaches rely on memory buffers, which are impractical in privacy-sensitive medical contexts. To address these challenges, we propose a rehearsal-free CL strategy for WSI analysis, employing a multimodal multi-resolution MIL classifier with an attribute word bank that pairs unique identifiers (keys) with prompts enriched with context-derived information.

Paper Contributions. In addressing the outlined challenges and introducing a novel architecture for the analysis of WSIs within a continual learning framework, our work makes several significant contributions to the field of medical image analysis: *(i)* the proposed approach overcomes the limitations of traditional continual learning strategies on WSIs, which rely on rehearsing previous data to prevent catastrophic forgetting. Our method offers a more efficient and privacy-compliant solution for continual learning in WSI analysis. *(ii)* We develop a novel prompt-learning-based MIL in WSI analysis. Different from other strategies, it exploits a MIL approach to contextualize prompts. *(iii)* Additionally, we introduce a novel solution of prompt learning tailored to the multi-resolution characteristics of WSIs, enabling our model to focus effectively on relevant features across different scales. To the best of our knowledge, this is the **first rehearsal-free CL strategy that employs MIL in WSI image classification.**

2 Related Work

Continual Learning (CL) for Histology. Continual learning—the ability to incrementally acquire new knowledge while retaining previously learned information—is vital in medical image analysis. It is generally categorized into three main approaches: *(i)* regularization-based methods, which penalize changes using various regularization terms (Kirkpatrick et al., 2017; Li and Hoiem, 2017; Zenke et al., 2017); *(ii)* rehearsal-based strategies, which retain and replay past data during training (Li and Hoiem, 2017; Buzzega et al., 2020; Boschini et al., 2022a; Chaudhry et al., 2019; Caccia et al., 2022); and *(iii)* architectural solutions, which expand model's parameters to accommodate new tasks (Rusu et al., 2016). Notably, some works have investigated continual learning in pathology (Derakhshani et al., 2022; Veena et al., 2022; Thandiackal et al., 2024). However, these efforts are primarily limited to patch-level analysis. Addressing these challenges at the slide level, ConSlide (Huang et al., 2023a) introduces a continual learning framework specifically designed for WSI classification. ConSlide employs a rehearsal-based strategy by maintaining a memory buffer of representative slide-level features from previous tasks. During training, it replays this stored data to mitigate catastrophic forgetting, effectively balancing the learning of new and old tasks. While ConSlide advances continual learning in digital pathology, its reliance on rehearsal methods introduces concerns related to data storage and potential privacy issues, critical factors in medical applications where data security is crucial. In contrast, our work is the first to propose a rehearsal-free continual learning framework for WSI classification.

Continual Prompt Learning. Vision-Language Models (Jia et al., 2021; Radford et al., 2021) showcased remarkable capabilities in learning versatile visual representations on stan-

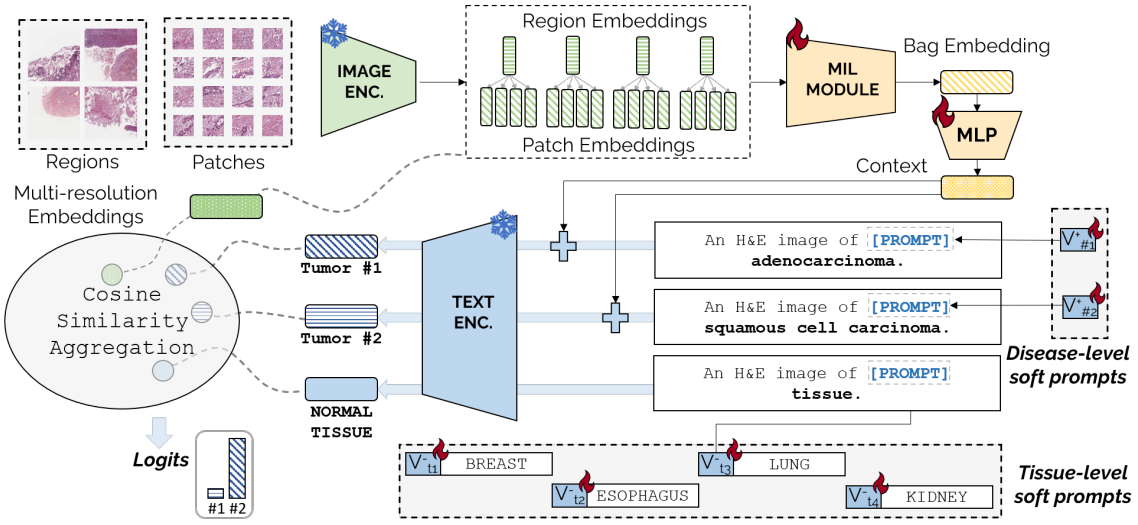

Figure 1: WSIs are decomposed into regions and patches, from which multi-resolution embeddings are extracted using a frozen image encoder (top left). These features are aggregated using a MIL module (top right) to produce a context-aware bag embedding. The context is injected into learnable soft prompts which are processed by a frozen text encoder (center). Classification logits at the slide level are obtained by computing the cosine similarity between the visual and textual embeddings (bottom left).

dard benchmarks (Radford et al., 2021) and histological data (Huang et al., 2023b). Building on these advancements, recent research has focused on optimizing training methodologies for specific classification tasks, moving beyond conventional model fine-tuning to prevent a degradation of the representation space (Gao et al., 2023; Wortsman et al., 2022; Yao et al., 2021; Zhang et al., 2021; Dong et al., 2019; He et al., 2016). In particular, recent efforts in continual learning have introduced visual prompt tuning (Wang et al., 2022b,a), integrating a minimal set of adaptable parameters directly into the input, thereby furnishing the pre-trained models with additional guidance for enhanced performance on downstream tasks (Li et al., 2021b). L2P (Wang et al., 2022b) bridges visual prompting with continual learning, utilizing a shared prompt pool for task sequence adaptation. DualPrompt (Wang et al., 2022a), on the other hand, enriches the pre-trained model with dual visual prompts, delineating both generic and task-specific directives, taking inspiration from complementary dual systems (Arani et al., 2022). However, none of the aforementioned prompting strategies can be directly applied to the gigapixel nature of WSIs.

3 Method

Overview. To address the challenges of continual WSI classification preserving privacy, we propose a solution that integrates rehearsal-free continual learning techniques with multi-instance learning. The proposed model comprises several components designed to enable effective classification in a continual learning setting. The pipeline (Fig. 1) starts with an image encoder that fuses features extracted from patches at multiple resolutions (Sec. 3.1). Subsequently, a context-aware MIL module provides a bag-level representation, which is

injected into learnable soft prompts that are then processed by a text encoder (Sec. 3.2). Classification logits are obtained by computing cosine similarities between the visual and textual embeddings (Sec. 3.3). In addition, a continual word bank (Fig. 2), facilitates the dynamic retrieval of the most relevant prompts over time, as detailed in Sec. 3.4.

3.1 Multi-scale Slide Representation

We compute a multi-scale representation for each WSI employing the image encoder $f(\cdot)$ of a foundation model pre-trained on histological images (Lu et al., 2024). Each slide is represented as a multi-resolution embedding B (green box in Fig. 1) of instances x_i:

$$B = \{x_1, ..., x_n\}; \;\; x_i = \bigcup_{\forall \; p_j \in r_i;} \{\frac{f(p_j) + f(r_i)}{2}\}; \tag{1}$$

where r_i denotes the i-th region at a coarser resolution, and p_j are the patches at a finer resolution within r_i. The notation $p_j \in r_i$ indicates that patch p_j is contained within region r_i. Averaging features from finer patches p_j and their corresponding coarser regions r_i captures both macro and micro details, yielding robust representations.

3.2 Context-aware Prompt Learning

Textual prompts often generalize poorly (Zhou et al., 2022). To address this, we draw inspiration from DSMIL (Li et al., 2021a) and CoCoOp (Zhou et al., 2022), and propose a mechanism to inject image-derived contextual information directly into the prompts, enhancing their relevance and adaptability.

Specifically, in this MIL module, the instance-level representation x_i is transformed into two vectors, corresponding to query q_i and value v_i, computed as: $q_i = W_q x_i, \; v_i = W_v x_i$ with $i = 0, \ldots, N-1$, where W_q and W_v are learnable weight matrices. We use a distance measurement U to quantify the similarity between an arbitrary instance and the critical instance x_{crit}, selected using max pooling, as in Eq. (2):

$$U(x_i, x_{crit}) = \frac{\exp(\langle q_i, q_{crit} \rangle)}{\sum_{k=1}^{n} \exp(\langle q_k, q_{crit} \rangle)} \tag{2}$$

$$b = \sum_{i=1}^{n} U(x_i, x_{crit})\, v_i \tag{3}$$

For constructing the bag embedding b in Eq. (3), we perform an element-wise weighted sum of the value vector v_i across all instances using $U(x_i, x_{crit})$ as weights. To facilitate the information flow from the MIL module to the learnable prompts, we employ a MLP denoted as M. The context to be injected in the prompt, denoted as π (yellow box in Fig. 1) is obtained as $\pi = M(b)$ (4).

In the proposed model, we introduce two distinct prompts with complementary functions, i.e., localization and classification. Each prompt is composed of a static template ("An H&E image of"), a learnable word embedding vector V, and a given class name `CLS`:

$$P^{tumor,normal} = \left[TEMPLATE, V^{+,-}, \texttt{CLS}^{tumor,normal}\right] \tag{5}$$

The P^{normal} is designed to distinguish normal tissues from pathological anomalies, thus localizing areas of interest. The P^{tumor}, on the other hand, is used to classify the identified instances based on the type of tumor. This two-step process ensures that the system not

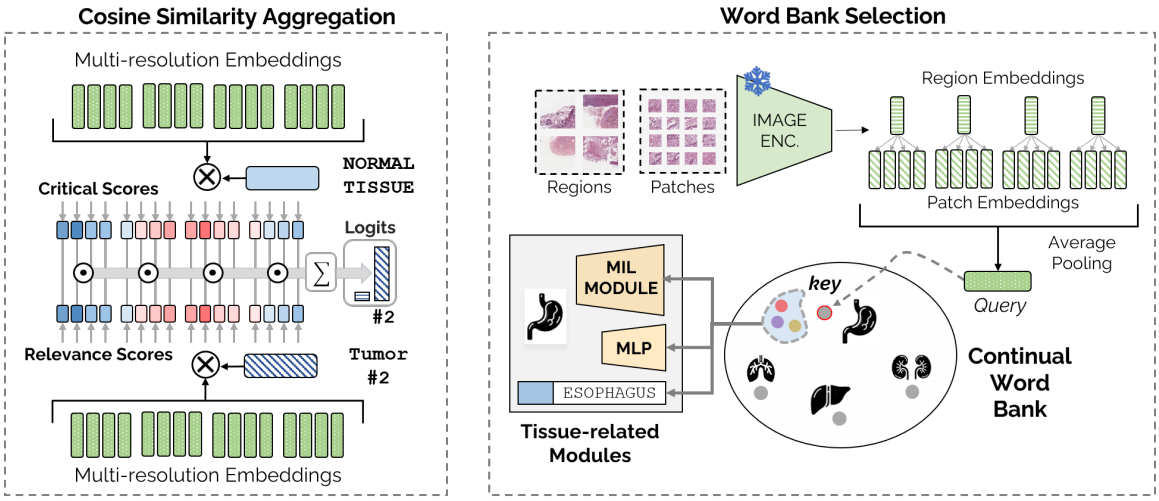

Figure 2: Cosine Similarity Aggregation (left) and Continual Word Bank (right). On the left, multi-resolution embeddings are matched against specific tumor and normal textual prompts. The comparison with normal tissue prompts allows for modulating tumor scores for the final prediction. On the right, a query representing the average instance embedding is compared with keys to select the corresponding set of parameters.

only detects areas of concern but also provides a precise classification. Having defined both the context π, Eq. (4), and the base prompt $P^{tumor,normal}$, Eq. (5), the final prompt after context injection is obtained as $P_\pi^{tumor,normal} = P^{tumor,normal} + \pi$ and processed by the text encoder $e(\cdot)$.

3.3 Cosine Similarity Aggregation

For each multi-scale instance feature x_i, Eq. (1), class c and tissue t, we compute the cosine similarity between the visual and textual features as $S_{x_i,c}^{tumor} = \langle x_i, e(P_c^{tumor}) \rangle$ and $S_{x_i,t}^{normal} = \langle x_i, e(P_t^{normal}) \rangle$. To make a single prediction S_c for the whole image, we aggregate the *relevance scores* $S_{x_i,c}^{tumor}$ modulated by the corresponding magnitude of the *critical scores* $S_{x_i,t}^{normal}$ (Fig. 2, left):

$$S_c = \sum_{x_i} \mathbb{1}_{c \in t} \left(exp(-S_{x_i,t}^{normal}) \cdot S_{x_i,c}^{tumor} \right) \tag{6}$$

In a nutshell, we compute the aggregated slide-level representation by summing the similarity score of the tumor regions weighted by the inverse of their degree of "normality".

3.4 Continual Word Bank

To work effectively within a continual learning framework, we introduce the Continual Word Bank (Fig. 2, right). This dynamic repository accumulates and refines task-specific prompts over time, each tailored to a tissue type and its associated classes. By selectively retrieving these prompts, the model can process a sequence of tasks without catastrophic forgetting. At each task t_i, we compute the average patch representation for each bag B_h as in Eq. (7); then we apply the K-Means clustering algorithm to partition them into k clusters. Each

cluster G_j is associated with a centroid g_j, Eq. (8), representing the mean feature vector of the points belonging to that cluster:

$$\bar{b}_h = \frac{1}{|B_h|} \sum_{x_i \in B_h} x_i, \qquad (7) \qquad\qquad g_j = \frac{1}{|G_j|} \sum_{\bar{b}_h \in G_j} \bar{b}_h. \qquad (8)$$

These centroids are updated with each new task and serve as keys to select the most relevant prompts. In this key-value system, the values store both the prompts and the MIL modules, which are lightweight components that ensure scalability and efficiency across tasks. During inference, the average instance representation for each new slide is computed similarly to how it was during training, and it serves as a query to retrieve the most pertinent entries from the continual word bank via a nearest-neighbor search. During training, we minimize the Cross-Entropy loss:

$$L(B, y, t_i) = \mathop{\mathbb{E}}_{\forall (B,y) \in t_i} (y \cdot log(softmax(y_c))) \qquad (9)$$

where y_c represents the predicted class scores. For the text prompts, $y_c = S_c$, while for the MIL module, $y_c = W_{cls} \cdot b$, where W_{cls} is a learnable weight and b is defined in Eq. (3). Finally, the loss is computed as $L = L_{MIL} + L_{text}$. In a continual context, a set of parameters (including V^+, V^-, and the MIL module) is instantiated for each task t_i. Only parameters corresponding to the current task are optimized during training.

4 Experiments

4.1 Datasets

Continual WSI benchmark. To validate our proposed architecture in a class-incremental learning setting, we conducted experiments using an improved version of the benchmark introduced by ConSlide (Huang et al., 2023a). Class-incremental learning requires models to recognize new classes without forgetting previously learned ones. In this context, the order of the datasets plays a crucial role, particularly with datasets of varying sizes and complexities. Unlike the original ConSlide benchmark, our version explicitly accounts for dimensionality differences, class imbalance, and presentation order. Tab. 1 reports the mean and standard deviation of a 10-fold cross-validation performed on two task orders: one from the most to the least numerous (Tab. 1a), and its reverse (Tab. 1b). The data include four tumor types: **NSCLC**, **BRCA**, **RCC**, **ESCA**.

Preprocessing. Each slide is processed with CLAM (Lu et al., 2021) to extract non-overlapping regions r of dimensions $4\,096 \times 4\,096$ at up to $20\times$ magnification. These tiles are partitioned into 64 non-overlapping patches p of size 512×512. All the instances are resized to 224×224 and encoded with CONCH's vision encoder (Lu et al., 2024).

Experimental Setting. Our model employs CooMIL as the backbone, while all continual learning baselines adopt HIT, the architecture from ConSlide (Huang et al., 2023a). To ensure fairness, all methods rely on the same CONCH-based embeddings. Models are trained for 50 epochs using the Adam optimizer with a learning rate of 0.0003, employing a 10-fold cross-validation approach.

Evaluation Metrics. Besides Accuracy (**ACC**), we report CL metrics (De Lange et al., 2021) such as Task Accuracy (**Task ACC**), i.e., performance on the current task only, and Forgetting (**Fgt.**), the decline in accuracy on earlier tasks (Boschini et al., 2022b). Metrics reported in Tab. 1 were saved at the end of the final task in a ten-fold validation fashion.

Table 1: Comparison of continual learning methods across different dataset orders.

		(a) NSCLC → BRCA → RCC → ESCA			*(b) ESCA → RCC → BRCA → NSCLC*		
CL Type	**Method**	**ACC (↑)**	**Task ACC (↑)**	**Fgt.(↓)**	**ACC (↑)**	**Task ACC (↑)**	**Fgt.(↓)**
	Joint (Upper)	91.6 ± 2.4	91.5 ± 3.2		91.6 ± 2.4	91.5 ± 3.2	
	Naïve (Lower)	21.7 ± 3.0	38.4 ± 9.2	51.0 ± 12.9	32.6 ± 1.3	38.1 ± 5.6	75.7 ± 7.4
Regularization Based	LwF	19.7 ± 4.3	28.9 ± 0.9	44.2 ± 5.3	32.9 ± 1.6	39.8 ± 11.7	81.7 ± 15.6
	EWC	28.1 ± 2.6	56.0 ± 1.4	64.5 ± 4.1	46.6 ± 5.0	55.1 ± 1.8	77.9 ± 6.1
Rehearsal Based	GDumb	48.4 ± 12.2	18.1 ± 3.6	1.1 ± 2.3	42.8 ± 15.1	17.6 ± 7.7	6.0 ± 4.9
	ER-ACE	86.8 ± 2.9	87.8 ± 1.8	2.8 ± 1.4	88.8 ± 2.1	90.6 ± 2.7	7.3 ± 5.0
	DER++	88.4 ± 1.2	90.3 ± 1.2	3.7 ± 0.7	89.6 ± 1.1	91.2 ± 3.1	5.6 ± 4.3
	DER++ w/o buf.	29.9 ± 3.8	57.9 ± 2.1	62.9 ± 5.2	48.6 ± 4.3	58.0 ± 2.1	79.2 ± 4.1
	ConSlide	66.3 ± 3.7	80.3 ± 1.5	25.8 ± 3.4	69.0 ± 3.8	81.8 ± 2.7	49.2 ± 5.4
	ConSlide w/o buf.	26.5 ± 4.2	54.8 ± 1.9	66.5 ± 5.5	37.6 ± 4.5	53.1 ± 2.6	89.2 ± 3.7
Prompt Based	**CooMIL (Ours)**	**88.6 ± 2.7**	**90.7 ± 2.5**	**3.6 ± 1.4**	**89.9 ± 3.4**	**91.3 ± 2.4**	**5.1 ± 3.8**

Table 2: (a) Task accuracy with varying numbers of centroids. (b) Impact of context and multi-scale on classification performance. (c) Impact of context type, here **V** denotes injecting context only into the learnable part of the prompt, and **P** refers to appending the context to the entire prompt. **Tumor** and **Normal** indicate where injection is performed.

(a) Clusters per Task

# Centroids	Task ACC
1	85.7 ± 2.3
4	86.5 ± 1.0
8	87.5 ± 1.9
12	87.2 ± 2.0
14	**90.7 ± 2.5**
16	88.5 ± 2.4

(b) Context and Multi-Scale

Context	Multi Scale	ACC
✗	✗	84.6 ± 2.8
✗	✓	86.6 ± 2.0
✓	✗	86.9 ± 3.0
✓	✓	**88.6 ± 2.7**

(c) Context Type

V/P	Tumor	Normal	ACC
V	✓	✓	87.2 ± 3.2
V	✓	✗	85.2 ± 3.1
V	✗	✓	84.2 ± 3.6
P	✓	✓	**88.6 ± 2.7**
P	✓	✗	86.2 ± 2.2
P	✗	✓	85.0 ± 2.9

4.2 Experimental Results

Continual Comparison We evaluated our proposed model against several leading continual learning baselines, covering regularization and rehearsal approaches in Tab. 1. Regularization-based methods, such as LwF (Li and Hoiem, 2017) and EWC (Kirkpatrick et al., 2017), aim to mitigate forgetting by constraining updates to important parameters for previously learned tasks. On the other hand, rehearsal-based methods, including GDumb (Prabhu et al., 2020), ER-ACE (Caccia et al., 2022), DER++ (Buzzega et al., 2020), and ConSlide (Huang et al., 2023a), rely on storing and replaying a subset of past data to help retain knowledge. These models were evaluated with a fixed buffer size of 5 WSIs.

Specifically, our model boosts overall accuracy by over 20% and task-specific accuracy by 10% compared to ConSlide, in both normal and reverse task orders. Although ConSlide relies on a memory buffer, it still suffers a high forgetting rate (25.8% in normal order, nearly doubling in reverse). By contrast, our model—without any buffer—achieves substantially lower forgetting rates of just 3.6% in normal order and 5.1% in reverse order. The only methods with metrics comparable to ours in both the order settings are ER-ACE and DER++. However, they rely on a memory buffer and require significantly higher computational resources than our model. These results underscore the superior performance of our approach, which avoids storing past data while still achieving better overall metrics.

4.3 Further Analysis

What is the optimal number of centroids? In Tab. 2(a), we explore how the number of centroids per task influences the performance of the task predictor. A large number of clusters can better represent the entire task variability. If too high, the centroids overfit the training representation. A good balance is obtained considering 14 centroids per task, which performs 90.7% in task identification accuracy.

Are multi-scale and context effective? In Tab. 2(b), multi-scale representations consistently enhance performance. Without context, incorporating multi-scale features leads to a 2% increase in accuracy. When context is included, the addition of multi-scale representations yields a further 1.7% improvement. Similarly, adding context improves accuracy by 2.3% when multi-scale features are not used, and by 2% when they are. These results confirm the effectiveness of both multi-scale and contextual features, independently and in combination. Tab. 2(c) investigates the impact of different context injection strategies. Results show that appending context to the full prompt (P) generally outperforms partial injection (V). The best accuracy (88.6% $\pm$ 2.7) is achieved when both tumor and normal context are used with full-prompt injection. Removing either context source results in a consistent performance drop, confirming the complementary value of tumor and normal contextual signals.

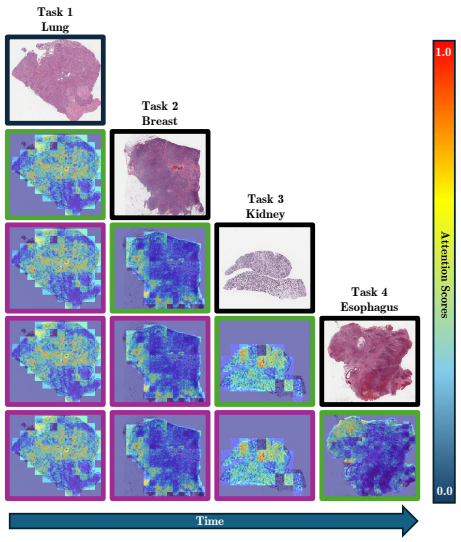

Figure 3: Patch-level attention maps over four consecutive tasks.

Is localization stable over time? In Fig. 3 we present a qualitative analysis showing the model's ability to maintain consistent attention to relevant image regions across sequential tasks. Visualizations reveal stable localization for both current and past tasks (green and violet, respectively), highlighting the model's capacity for knowledge retention. This is especially important in medical contexts, where consistent and interpretable localization across resolutions and tasks enhances clinical trust and decision-making.

5 Conclusion

This work addresses key challenges in continual learning for WSI classification, including catastrophic forgetting, large-scale image analysis, multi-resolution processing, and privacy concerns. By integrating critical information into learnable prompts, CooMIL enhances classification performance and context awareness. Evaluations on four WSI datasets show improved accuracy and reduced forgetting. Despite these advantages, limitations remain. Although aligned with the existing literature, the benchmark tasks are relatively simple, and future work should incorporate more diverse imaging conditions and classification scenarios. While our approach involves incremental parameter growth, the overhead is minimal. Exploring prompt learning within the vision encoder also offers promising future directions.

References

Rahaf Aljundi, Francesca Babiloni, Mohamed Elhoseiny, Marcus Rohrbach, and Tinne Tuytelaars. Memory Aware Synapses: Learning what (not) to forget . In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

Elahe Arani, Fahad Sarfraz, and Bahram Zonooz. Learning Fast Learning Slow: A General Continual Learning Method based on Complementary Learning System. In *International Conference on Learning Representations*, 2022.

Matteo Boschini, Lorenzo Bonicelli, Pietro Buzzega, Angelo Porrello, and Simone Calderara. Class-incremental continual learning into the extended der-verse. *IEEE transactions on pattern analysis and machine intelligence*, 2022a.

Matteo Boschini, Lorenzo Bonicelli, Angelo Porrello, Giovanni Bellitto, Matteo Pennisi, Simone Palazzo, Concetto Spampinato, and Simone Calderara. Transfer without Forgetting. In *European Conference on Computer Vision*, 2022b.

Pietro Buzzega, Matteo Boschini, Angelo Porrello, Davide Abati, and Simone Calderara. Dark experience for general continual learning: a strong simple baseline. *Advances in neural information processing systems*, 33:15920–15930, 2020.

Lucas Caccia, Rahaf Aljundi, Nader Asadi, Tinne Tuytelaars, Joelle Pineau, and Eugene Belilovsky. New Insights on Reducing Abrupt Representation Change in Online Continual Learning. In *International Conference on Learning Representations*, 2022.

Arslan Chaudhry, Marc'Aurelio Ranzato, Marcus Rohrbach, and Mohamed Elhoseiny. Efficient lifelong learning with a-gem. In *International Conference on Learning Representations*, 2019.

Matthias De Lange, Rahaf Aljundi, Marc Masana, Sarah Parisot, Xu Jia, Aleš Leonardis, Gregory Slabaugh, and Tinne Tuytelaars. A continual learning survey: Defying forgetting in classification tasks. *PAMI*, 2021.

Mohammad Mahdi Derakhshani, Ivona Najdenkoska, Tom van Sonsbeek, Xiantong Zhen, Dwarikanath Mahapatra, Marcel Worring, and Cees G. M. Snoek. LifeLonger: A Benchmark for Continual Disease Classification. In *Medical Image Computing and Computer Assisted Intervention*, volume 13432, pages 314–324. Springer, 2022.

Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. Unified language model pre-training for natural language understanding and generation. *Advances in neural information processing systems*, 32, 2019.

Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *International Journal of Computer Vision*, pages 1–15, 2023.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

Shih-Chiang Huang, Chi-Chung Chen, Jui Lan, Tsan-Yu Hsieh, Huei-Chieh Chuang, Meng-Yao Chien, Tao-Sheng Ou, Kuang-Hua Chen, Ren-Chin Wu, Yu-Jen Liu, et al. Deep neural network trained on gigapixel images improves lymph node metastasis detection in clinical settings. *Nature Communications*, 13(1):1–14, 2022.

Yanyan Huang, Weiqin Zhao, Shujun Wang, Yu Fu, Yuming Jiang, and Lequan Yu. Conslide: Asynchronous hierarchical interaction transformer with breakup-reorganize rehearsal for continual whole slide image analysis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21349–21360, 2023a.

Zhi Huang, Federico Bianchi, Mert Yuksekgonul, Thomas J Montine, and James Zou. A visual–language foundation model for pathology image analysis using medical twitter. *Nature Medicine*, pages 1–10, 2023b.

Sajid Javed, Arif Mahmood, Naoufel Werghi, Ksenija Benes, and Nasir Rajpoot. Multiplex cellular communities in multi-gigapixel colorectal cancer histology images for tissue phenotyping. *IEEE Transactions on Image Processing*, 29:9204–9219, 2020.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017.

Bin Li, Yin Li, and Kevin W Eliceiri. Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14318–14328, 2021a.

Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. In *NeurIPS*, 2021b.

Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE transactions on pattern analysis and machine intelligence*, 40(12):2935–2947, 2017.

Ming Y Lu, Drew FK Williamson, Tiffany Y Chen, Richard J Chen, Matteo Barbieri, and Faisal Mahmood. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature biomedical engineering*, 5(6):555–570, 2021.

Ming Y Lu, Bowen Chen, Drew FK Williamson, Richard J Chen, Ivy Liang, Tong Ding, e Guillaume Jaum, Igor Odintsov, Long Phi Le, Georg Gerber, et al. A visual-language foundation model for computational pathology. *Nature Medicine*, 30:863–874, 2024.

Wenqi Lu, Simon Graham, Mohsin Bilal, Nasir Rajpoot, and Fayyaz Minha. Capturing cellular topology in multi-gigapixel pathology images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 260–261, 2020.

Ameya Prabhu, Philip HS Torr, and Puneet K Dokania. Gdumb: A simple approach that questions our progress in continual learning. In *European conference on computer vision*, pages 524–540, 2020.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

Anthony Robins. Catastrophic forgetting rehearsal and pseudorehearsal. *Connection Science*, 1995.

Andrei A Rusu, Neil C Rabinowitz, Guillaume Desjardins, Hubert Soyer, James Kirkpatrick, Koray Kavukcuoglu, Razvan Pascanu, and Raia Hadsell. Progressive neural networks. *arXiv preprint arXiv:1606.04671*, 2016.

Kevin Thandiackal, Luigi Piccinelli, Rajarsi Gupta, Pushpak Pati, and Orcun Goksel. Multi-scale feature alignment for continual learning of unlabeled domains. *IEEE Transactions on Medical Imaging*, 2024.

Jeroen Van der Laak, Geert Litjens, and Francesco Ciompi. Deep learning in histopathology: the path to the clinic. *Nature medicine*, 27(5):775–784, 2021.

Kaustaban Veena, Ba Qinle, Bhattacharya Ipshita, Sobh Nahil, Mukherjee Satarupa, Martin Jim, Miri Mohammad Saleh, Guetter Christoph, and Chaturvedi Amal. Characterizing continual learning scenarios for tumor classification in histopathology images. In *International Workshop on Medical Optical Imaging and Virtual Microscopy Image Analysis*, pages 177–187. Springer, 2022.

Zifeng Wang, Zizhao Zhang, Sayna Ebrahimi, Ruoxi Sun, Han Zhang, Chen-Yu Lee, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, et al. Dualprompt: Complementary prompting for rehearsal-free continual learning. In *European Conference on Computer Vision*, pages 631–648, 2022a.

Zifeng Wang, Zizhao Zhang, Chen-Yu Lee, Han Zhang, Ruoxi Sun, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, and Tomas Pfister. Learning to prompt for continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 139–149, 2022b.

Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo Lopes, Hannaneh Hajishirzi, Ali Farhadi, Namkoong Hongseok, et al. Robust fine-tuning of zero-shot models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7959–7971, 2022.

Yuan Yao, Ao Zhang, Zhengyan Zhang, Zhiyuan Liu, Tat-Seng Chua, and Maosong Sun. Cpt: Colorful prompt tuning for pre-trained vision-language models. *arXiv preprint arXiv:2109.11797*, 2021.

Friedemann Zenke, Ben Poole, and Surya Ganguli. Continual learning through synaptic intelligence. In *International Conference on Machine Learning*, pages 3987–3995, 2017.

Renrui Zhang, Rongyao Fang, Wei Zhang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free clip-adapter for better vision-language modeling. *arXiv preprint arXiv:2111.03930*, 2021.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16816–16825, 2022.