Pathological Prior-Guided Multiple Instance Learning For Mitigating Catastrophic Forgetting in Breast Cancer Whole Slide Image Classification

Weixi Zheng
School of Computer Science
Wuhan University
Wuhan, China
acnicotine@whu.edu.cn

Aoling Huang
Department of Pathology
Renmin Hospital of Wuhan University
Wuhan, China
huangaoling@whu.edu.cn

Jingping Yuan
Department of Pathology
Renmin Hospital of Wuhan University
Wuhan, China
yuanjingping@whu.edu.cn

Haoyu Zhao
School of Computer Science
Wuhan University
Wuhan, China
haoyu.zhao@whu.edu.cn

Zhou Zhao
School of Computer Science
Central China Normal University
Wuhan, China
zhaozhou@ccnu.edu.cn

Yongchao Xu*
School of Computer Science
Wuhan University
Wuhan, China
yongchao.xu@whu.edu.cn

Thierry Géraud
EPITA Research Laboratory
Le Kremlin-Bicêtre, France
thierry.geraud@epita.fr

***Abstract*—In histopathology, intelligent diagnosis of Whole Slide Images (WSIs) is essential for automating and objectifying diagnoses, reducing the workload of pathologists. However, diagnostic models often face the challenge of forgetting previously learned data during incremental training on datasets from different sources. To address this issue, we propose a new framework PaGMIL to mitigate catastrophic forgetting in breast cancer WSI classification. Our framework introduces two key components into the common MIL model architecture. First, it leverages microscopic pathological prior to select more accurate and diverse representative patches for MIL. Secondly, it trains separate classification heads for each task and uses macroscopic pathological prior knowledge, treating the thumbnail as a prompt guide (PG) to select the appropriate classification head. We evaluate the continual learning performance of PaGMIL across several public breast cancer datasets. PaGMIL achieves a better balance between the performance of the current task and the retention of previous tasks, outperforming other continual learning methods. Our code will be open-sourced upon acceptance.**

***Index Terms*—Whole Slide Image Classification, Multiple Instance Learning, Continual Learning, Breast Cancer**

I. Introduction

WSIs provide crucial histopathological information for breast cancer diagnosis. Over 30 million breast cancer slides are analyzed annually, underscoring the need for intelligent WSI diagnosis. Deep learning advancements offer potential for automated diagnosis, prognosis, and treatment, reducing pathologists' workload [1]–[3]. However, the large size of WSIs and costly pixel-level annotations present challenges for deep learning models [4]–[6]. To address this, CLAM [7] has been proposed as an effective multiple instance learning (MIL) approach for WSI analysis. It utilizes slide-level labels to divide WSIs into patches, analyzes them, and aggregates results for slide-level predictions.

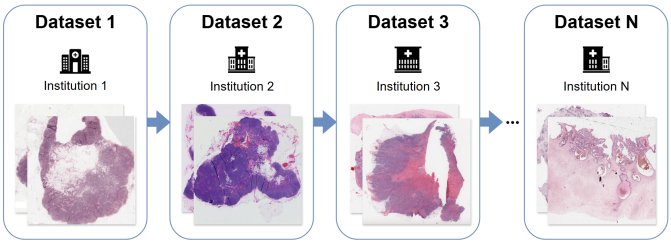


Fig. 1. Illustration of dynamic input of WSI data. The actual model update involves different medical institutions sending in some data for training. Significant differences in data from different batches arise due to variations in staining methods or equipment used across different medical institutions.

Despite encouraging results, existing methods [2], [8]–[12] typically use static models trained and tested on fixed datasets. However, WSI imaging is dynamic, influenced by variations in equipment and staining [13], which limits model performance across different environments. In practical applications, models encounter pathology images from various medical institutions (see Fig. 1), requiring adaptability to maintain robustness [14]–[18]. Fine-tuning pre-trained models on new datasets is one solution, but it often causes catastrophic forgetting, where models overfit to new data and lose knowledge from previous datasets [19]–[21].

Continual learning (CL) [20], [22] aims to mitigate catastrophic forgetting by allowing models to adapt to new tasks while retaining knowledge from previous ones, enhancing robustness and accommodating data growth. CL has shown

*: Corresponding Authors.

This work was supported in part by the National Key Research and Development Program of China (2023YFC2705700), NSFC 62222112, and 62176186, the Innovative Research Group Project of Hubei Province under Grants (2024AFA017), the Cross-disciplinary Innovation Talent Project of Renmin Hospital of Wuhan University (JCRCZN-2022-015).

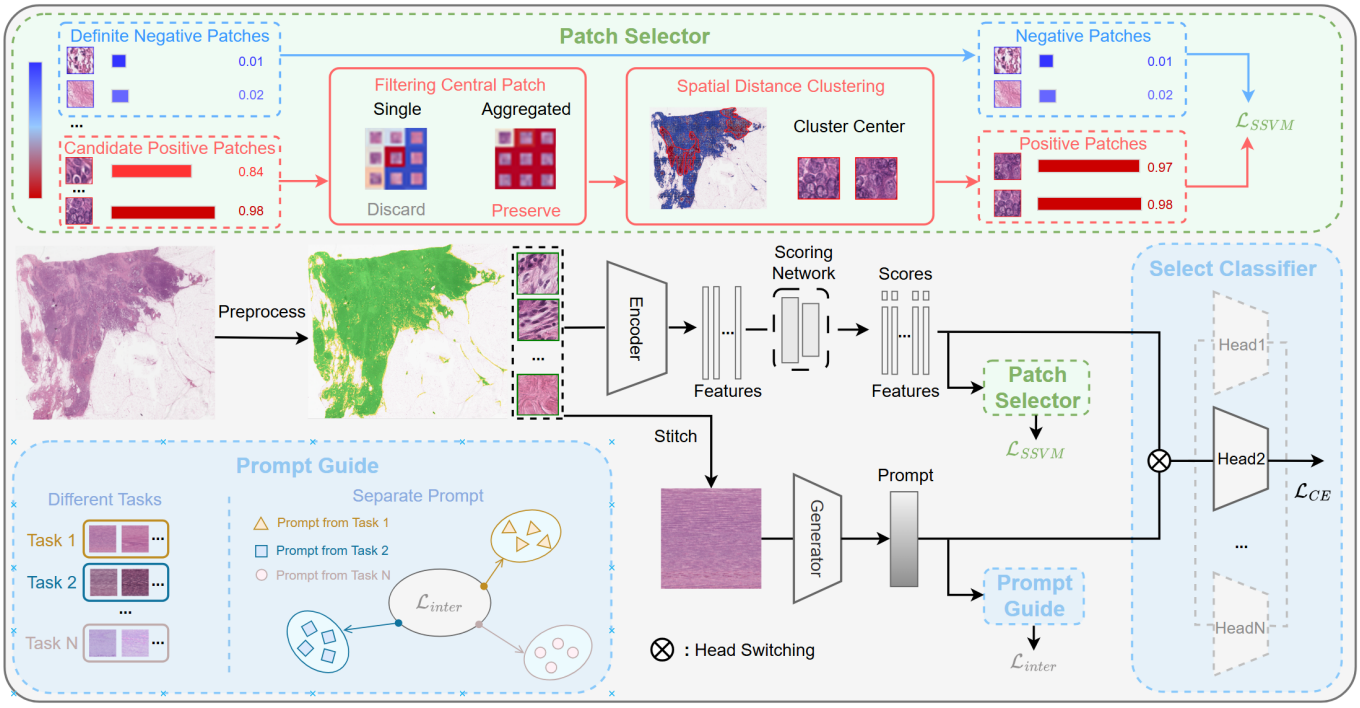


Fig. 2. Overview of our proposed PaGMIL method. Based on the MIL pipeline, we add two new modules. The Patch Selector module uses microscopic pathology prior to select more accurate representative patches, and the PG module uses macroscopic pathology priors to select the correct classification head.

promise on natural images and includes approaches like parameter regularization [23]–[25], knowledge distillation [26]–[28], dynamic network structure [29]–[31], and data replay [32]–[34]. However, applying these CL methods to WSI images is challenging due to their large size and the presence of background noise, making data replay, regularization, and distillation less effective. In recent years, a data replay method [35] is designed specifically for the characteristics of WSIs, achieving promising results.

In this paper, we propose PaGMIL, a novel MIL continual learning network that leverages both micro and macro pathological priors to guide the learning of new data and mitigate catastrophic forgetting. At the micro level, we introduce positional information by recording the coordinates of each patch. By incorporating prior knowledge about cancer clustering, it is possible to rule out misjudgments caused by staining bias or noise. At the same time, we select patches with large spatial distances to capture diverse tissue environments. This encourages the attention network to learn generalizable knowledge to mitigate forgetting. At the macro level, we introduce global information by converting the WSI into thumbnail and feeding it into a Generator to generate prompt. This prompt serves as global information to indicate the current task during classification and helps select the corresponding classification head.

In summary, the main contributions are threefold: 1) We introduce positional information to select more representative patches for continual learning. 2) We use global information to select the correct classification head. 3) We evaluate PaGMIL on four public breast cancer WSI datasets (*TCGA*, *CAMELYON16*, *BRACS*, and *BACH*) and demonstrate superior balance between the performance of the current task and the retention of previous tasks.

II. Method

A. Overview

Our proposed PaGMIL method, shown in Fig. 2, follows the general MIL approach of CLAM [7], with the addition of two modules: Patch Selector (PS) and Prompt Guide (PG). The PS module utilizes microscopic pathological priors to process ranked patches. It defines low-scoring patches as negative patches and identifies high-scoring patches as candidate positive patches for further screening. The PS module filters out isolated high-score patches, then clusters the remaining high-score patches into B categories based on their positions, and selects B cluster centers as positive patches. The PG module leverages macroscopic pathological priors to convert the WSI into a thumbnail, which is then processed to generate prompts. Prompts for the same task are positioned closer, while those for different tasks are spaced further apart. PaGMIL trains a distinct classification head per task, which is linked to its prompts. For a new WSI, the appropriate classification head is selected based on its prompt for prediction.

B. Patch Selector

According to CLAM [7], we score and rank each patch to select positive and negative patches. Negative patches represent patches that are not important for the WSI and can

be selected as B patches with the lowest scores, denoted as $\mathcal{N}$. B is a parameter that needs to be set, representing the number of patches. Positive patches, as representative patches of the WSI, need to be carefully selected. We first identify the top $k\%$ of patches with the highest scores. These patches are then stored as candidate positive patches, denoted as $\mathcal{P}_c$.

In cancer tissue slides, representative patches must contain cancerous cells, and cancer tissues always cluster together [36]. Utilizing this knowledge, we can filter out some isolated high-scoring patches because these patches are usually staining bias or noise [37]. We record each patch's position to calculate distances. For each patch in $\mathcal{P}_c$, if there is any other patch in $\mathcal{P}_c$ adjacent to it, both the current patch and its neighbors are added to $\mathcal{P}_s$.

Within the same WSI, cancer cells and their surrounding environments may vary across regions, which brings patch diversity. This diversity helps the model focus on different features, allowing the model to acquire generalizable knowledge that mitigates catastrophic forgetting. In order to find high-score patches that are far away as positive patches, we perform K-means clustering on the patches in $\mathcal{P}_s$ based on location, resulting in B clusters. The center patches of these clusters serve as the B positive patches.

We then feed the positive and negative patches into the instance classifier like in CLAM [7] and supervise it using the smooth SVM Loss $\mathcal{L}_{SSVM}$ [38].

C. Prompt Guide

For each WSI, we stitch all tissue-containing patches into a large square and resize it to fit the Generator's input. We use a ResNet [39] as the generator to convert the thumbnails into 768-dimensional features, which serve as the prompt. This approach can retain macroscopic information such as color to infer which dataset the WSI belongs to.

During the training process, we generate prompts for the different data with each task. Once the training for task i is completed, we calculate the average value of the prompts and designate it as the prompt m_i for that task. It is crucial to minimize the differences within the prompts of the same task and maximize the differences between the prompts of different tasks. Therefore, we need to design two different types of losses to control the generation of the prompts.

Intra-class loss, to bring prompts of the same task closer:

$$\mathcal{L}_{intra} = \frac{1}{2} * \sum_{i=1}^{N} ||m_i - \bar{m}||^2, \tag{1}$$

where m_i represents the prompt generated for each WSI during the training process, and $\bar{m}$ represents the average prompt generated in the current epoch. At the start of a new epoch, $\bar{m}$ is recalculated. $||\dots||$ denotes the L2 (Euclidean) distance.

Inter-class loss, to keep prompts of different tasks apart:

$$\mathcal{L}_{inter} = -\frac{1}{2NT} \sum_{t=1}^{T} \sum_{i=1}^{N} \left[{d_{it}}^2 + \max(0, min - d_{it})^2 \right], \tag{2}$$

$$d_{it} = ||m_i - \bar{m}_t||, \tag{3}$$

where T represents the number of tasks trained previously, N represents the number of data points in the current task, d_{it} represents the Euclidean distance between the current prompt and the previous average prompt, and min is the minimum distance set, giving a greater penalty when the prompts of different tasks are closer to each other.

During testing, the thumbnail of the WSI is sent to the Prompt Generator, and the generated prompt is used to calculate the cosine similarity with the stored prompts. The classification head corresponding to the most similar prompt is selected. This typically implies that the style of the test data closely aligns with the style of the training set associated with a certain stored classification head.

III. Experiment

A. Datasets

We select breast cancer datasets from four distinct countries and regions. These datasets exhibit significant variations in staining and distribution. The datasets include: 1) *CAMELYON16*, Netherlands challenge dataset with official train/test sets; 2) *BRACS*, an Italian breast cancer dataset , from which we select 4 classes: normal, benign, in situ carcinoma, and invasive carcinoma; 3) *TCGA-BRCA*, a public WSI database in the United States, from which we chose a subset of images with relatively consistent distribution; and 4) *BACH*, a Portuguese breast cancer dataset, from which we chose the portion labeled as "Photos".

B. Experimental settings

We define a sequential order for the four datasets: *CAMELYON16*, *BRACS*, *TCGA-BRCA*, and *BACH*. This order places the most difficult dataset, CAMELYON16, at the beginning, puts the simple data in the middle, and places the relatively difficult data at the end. This arrangement poses the greatest challenge to the model's robustness. We evaluate the final performance of the model by measuring the Area Under the Curve (AUC) and accuracy (ACC).

We compare our method with the following approaches: 1) CLAM (Baseline) [7], sequentially training the model for newly arrived task; 2) EWC [23], a classic regularization-based method; 3) LWF [26], a classic knowledge distillation-based method; 4) DER++ [34], a classic data replay-based method; 5) MIND [27] and PEC [28], the state-of-the-art methods for regularization and distillation; 6) ConSlide [35], a method for continual learning on WSI.

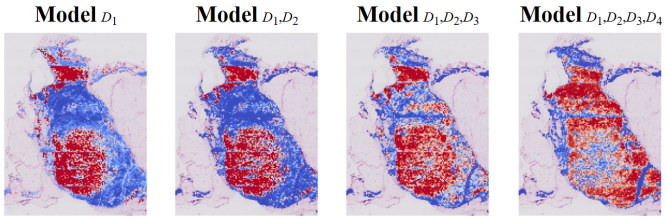
C. Implementation details

We crop regions of 224×224 pixels from the segmented foreground tissue at a 20x magnification to obtain patch-level images. Then we utilize Ctranspath [40] to extract features from the corresponding images. The previously mentioned parameter B is set to 8, meaning that PaGMIL will select 8 positive patches and 8 negative patches. Each task throughout experiment is conducted for 30 epochs. The experiments are run on an NVIDIA GeForce RTX 4090 GPU.

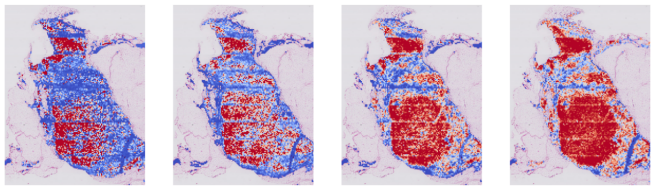
TABLE I
QUANTITATIVE COMPARISON OF DIFFERENT METHODS TRAINED SEQUENTIALLY.

Method	CAMELYON16 (D_1)		BRACS (D_2)		TCGA (D_3)		BACH (D_4)	
	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC
Separate	79.06	79.03	87.03	93.32	92.31	92.81	83.84	91.72
CLAM [7] (Baseline)	37.98	45.44	62.96	86.36	80.76	83.43	83.33	89.43
EWC [23] (PNAS2017)	37.98	58.39	68.51	90.19	84.23	85.12	85.25	90.49
LWF [26] (TPAMI2017)	37.98	47.73	76.62	85.36	86.54	92.50	84.44	92.35
DER++ [34] (NIPS2020)	38.75	47.60	81.48	89.48	84.61	90.62	76.56	86.70
MIND [27] (AAAI2024)	41.86	60.45	84.81	91.32	85.38	93.74	82.26	89.58
PEC [28] (ICLR2024)	39.53	59.38	78.42	82.75	89.21	90.69	84.04	90.68
ConSlide [35] (ICCV2023)	48.53	59.45	81.85	88.42	88.46	91.62	81.61	91.33
PaGMIL (Ours)	**62.40**	**68.64**	**85.18**	**92.13**	**90.38**	**93.74**	**85.45**	**92.85**

Note: Seperate denotes training a separate model for each dataset as a theoretical upper bound.



(a) Score heatmap of CLAM (Baseline) method



(b) Score heatmap of our method

Fig. 3. The image is from CAMELYON16, which is WSI with cancer. Red indicates high scores given by the scoring network, while blue indicates low scores. The models are trained sequentially from left to right in the order of datasets. As training progresses, the baseline method shows an opposite distribution of red and blue, changing the representative patches from cancerous to normal tissue. In contrast, our method maintains a similar distribution of red and blue.

D. Main results

Our method achieves good results in mitigating catastrophic forgetting according Table I. It is worth noting that, following sequential training, the baseline, EWC [23], and LWF [26] models all achieve a relatively low accuracy of 37.98% when tested on the *CAMELYON16* dataset. This is attributed to these models predicting all data as cancerous. We print out a score heatmap (see Fig.3) of a WSI from *CAMELYON16*, which might explain this phenomenon. The figure shows a heatmap based on the scoring, which is then weighted and used by the final classification head to determine whether the entire image is cancerous. In the Baseline method, the scoring network is always trained together with the classification head, leading them to make consistent decisions. After sequential training, the scoring network assigns high scores to normal tissue and low scores to cancerous tissue, misclassifying normal tissue as cancerous. Our method freezes the current classification head after training on one dataset, allowing the scoring network to focus solely on distribution of cancerous regions. Similar color distribution means that our PS module effectively captures generalizable knowledge and mitigates forgetting.

TABLE II
ABLATION STUDY ON THE IMPACT OF EACH MODULE.

PS	PG	CAMELYON16 (D_1)		BRACS (D_2)		TCGA (D_3)		BACH (D_4)	
		ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC
		37.98	45.44	62.96	86.36	80.76	83.43	83.33	89.43
✓		39.53	52.80	69.25	**92.75**	82.31	85.00	84.84	90.62
	✓	58.29	64.81	82.96	89.94	88.85	92.69	85.05	92.58
✓	✓	**62.40**	**68.64**	**85.18**	92.13	**90.38**	**93.74**	**85.45**	**92.85**

E. Ablation study

To validate the effectiveness of the two components in our method, we conduct ablation experiments, and results are shown in Table II. When using only the PS module, our results show improvement, indicating that the diversity of patches can help the model learn more generalizable knowledge to reduce catastrophic forgetting. Using only the PG module significantly improves results. This is because forgetting mainly occurs in the classification head. Using separate heads for each task and selecting the correct one during testing effectively reduces catastrophic forgetting.

IV. CONCLUSION

In this paper, we propose PaGMIL, a novel architecture for mitigating catastrophic forgetting in breast cancer diagnosis. The PS module selects more accurate and diverse representative patches based on microscopic pathological prior knowledge, while the PG module determines the appropriate classification head for each WSI based on macroscopic pathological knowledge. Extensive experiments on four breast cancer datasets demonstrate that PaGMIL achieves a superior balance between current task effectiveness and retention of previous knowledge. In the future, we will conduct continual learning research on different types of cancer.

REFERENCES

[1] T. C. Cornish, R. E. Swapp, and K. J. Kaplan, "Whole-slide imaging: routine pathologic diagnosis," *Advances in anatomic pathology*, vol. 19, no. 3, pp. 152–159, 2012.

[2] Z. Shao, H. Bian, Y. Chen, Y. Wang, J. Zhang, X. Ji *et al.*, "Transmil: Transformer based correlated multiple instance learning for whole slide image classification," *Proc. of Advances in Neural Information Processing Systems*, vol. 34, pp. 2136–2147, 2021.

[3] J. Tanizaki, H. Hayashi *et al.*, "Report of two cases of pseudoprogression in patients with non–small cell lung cancer treated with nivolumab—including histological analysis of one case after tumor regression," *Lung Cancer*, vol. 102, pp. 44–48, 2016.

[4] W. Lu, S. Graham, M. Bilal *et al.*, "Capturing cellular topology in multi-gigapixel pathology images." in *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition Workshops*, 2020, pp. 1049–1058.

[5] S.-C. Huang, C.-C. Chen, J. Lan *et al.*, "Deep neural network trained on gigapixel images improves lymph node metastasis detection in clinical settings," *Nature communications*, vol. 13, no. 1, p. 3347, 2022.

[6] S. Javed, A. Mahmood, N. Werghi *et al.*, "Multiplex cellular communities in multi-gigapixel colorectal cancer histology images for tissue phenotyping," *IEEE Trans. on Image Processing*, vol. 29, pp. 9204–9219, 2020.

[7] M. Y. Lu, D. F. Williamson, T. Y. Chen *et al.*, "Data-efficient and weakly supervised computational pathology on whole-slide images," *Nature biomedical engineering*, vol. 5, no. 6, pp. 555–570, 2021.

[8] H. Xu, N. Usuyama, J. Bagga, S. Zhang, R. Rao, T. Naumann, C. Wong, Z. Gero, J. González, Y. Gu *et al.*, "A whole-slide foundation model for digital pathology from real-world data," *Nature*, vol. 630, pp. 181–188, 2024.

[9] R. J. Chen, T. Ding, M. Y. Lu, D. F. Williamson, G. Jaume, A. H. Song, B. Chen, A. Zhang, D. Shao, M. Shaban *et al.*, "Towards a general-purpose foundation model for computational pathology," *Nature Medicine*, vol. 30, no. 3, pp. 850–862, 2024.

[10] R. J. Chen, C. Chen, Y. Li, T. Y. Chen, A. D. Trister, R. G. Krishnan, and F. Mahmood, "Scaling vision transformers to gigapixel images via hierarchical self-supervised learning," in *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, 2022, pp. 16 144–16 155.

[11] Y. Zheng, J. Li, J. Shi, F. Xie, and Z. Jiang, "Kernel attention transformer (kat) for histopathology whole slide image classification," in *Proc. of Intl. Conf. on Medical Image Computing and Computer Assisted Intervention*, 2022, pp. 283–292.

[12] Y. Guan, J. Zhang, K. Tian *et al.*, "Node-aligned graph convolutional network for whole-slide image representation and classification," in *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, 2022, pp. 18 813–18 823.

[13] Y. Shen, A. Sowmya, Y. Luo, X. Liang, D. Shen, and J. Ke, "A federated learning system for histopathology image analysis with an orchestral stain-normalization gan," *IEEE Trans. on Medical Imaging*, vol. 42, no. 7, pp. 1969–1981, 2023.

[14] C. S. Lee and A. Y. Lee, "Clinical applications of continual learning machine learning," *The Lancet Digital Health*, vol. 2, no. 6, pp. 279–281, 2020.

[15] M. M. Derakhshani, I. Najdenkoska, T. van Sonsbeek *et al.*, "Lifelonger: A benchmark for continual disease classification," in *Proc. of Intl. Conf. on Medical Image Computing and Computer Assisted Intervention*, 2022, pp. 314–324.

[16] V. Kaustaban, Q. Ba, I. Bhattacharya *et al.*, "Characterizing continual learning scenarios for tumor classification in histopathology images," in *International Workshop on Medical Optical Imaging and Virtual Microscopy Image Analysis*, 2022, pp. 177–187.

[17] J. Van der Laak, G. Litjens, and F. Ciompi, "Deep learning in histopathology: the path to the clinic," *Nature medicine*, vol. 27, no. 5, pp. 775–784, 2021.

[18] M. Perkonigg, J. Hofmanninger, C. J. Herold *et al.*, "Dynamic memory to alleviate catastrophic forgetting in continual learning with medical imaging," *Nature communications*, vol. 12, no. 1, p. 5678, 2021.

[19] M. Boschini, L. Bonicelli, A. Porrello *et al.*, "Transfer without forgetting," in *Proc. of European Conf. on Computer Vision*, 2022, pp. 692–709.

[20] M. De Lange, R. Aljundi, M. Masana *et al.*, "A continual learning survey: Defying forgetting in classification tasks," *IEEE Trans. on Pattern Anal. and Mach. Intell.*, vol. 44, no. 7, pp. 3366–3385, 2021.

[21] T. Lesort, V. Lomonaco, A. Stoian *et al.*, "Continual learning for robotics: Definition, framework, learning strategies, opportunities and challenges," *Information fusion*, vol. 58, pp. 52–68, 2020.

[22] D. Lopez-Paz and M. Ranzato, "Gradient episodic memory for continual learning," *Proc. of Advances in Neural Information Processing Systems*, vol. 30, 2017.

[23] J. Kirkpatrick, R. Pascanu, N. Rabinowitz *et al.*, "Overcoming catastrophic forgetting in neural networks," *Proceedings of the national academy of sciences*, vol. 114, no. 13, pp. 3521–3526, 2017.

[24] S.-A. Rebuffi, A. Kolesnikov, G. Sperl *et al.*, "ICARL: Incremental classifier and representation learning," in *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, 2017, pp. 2001–2010.

[25] R. Kurle, B. Cseke, A. Klushyn *et al.*, "Continual learning with bayesian neural networks for non-stationary data," in *Proc. of International Conference on Learning Representations*, 2019.

[26] Z. Li and D. Hoiem, "Learning without forgetting," *IEEE Trans. on Pattern Anal. and Mach. Intell.*, vol. 40, no. 12, pp. 2935–2947, 2017.

[27] J. Bonato, F. Pelosin, L. Sabetta, and A. Nicolosi, "Mind: Multi-task incremental network distillation," in *Proc. of the AAAI Conf. on Artificial Intelligence*, vol. 38, no. 10, 2024, pp. 11 105–11 113.

[28] M. Zajac, T. Tuytelaars, and G. M. van de Ven, "Prediction error-based classification for class-incremental learning," *arXiv preprint arXiv:2305.18806*, 2023.

[29] Y.-S. Liang and W.-J. Li, "Inflora: Interference-free low-rank adaptation for continual learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 23 638–23 647.

[30] Q. Pham, C. Liu, and S. Hoi, "Dualnet: Continual learning, fast and slow," *Proc. of Advances in Neural Information Processing Systems*, vol. 34, pp. 16 131–16 144, 2021.

[31] A. Douillard, A. Ramé, G. Couairon *et al.*, "Dytox: Transformers for continual learning with dynamic token expansion," in *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, 2022, pp. 9285–9295.

[32] Y. Li, Q. Li, H. Wang, R. Li, W. Zhong, and G. Zhang, "Towards efficient replay in federated incremental learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 12 820–12 829.

[33] J. Kim, H. Cho, J. Kim, Y. Y. Tiruneh, and S. Baek, "Sddgr: Stable diffusion-based deep generative replay for class incremental object detection," *arXiv preprint arXiv:2402.17323*, 2024.

[34] P. Buzzega, M. Boschini, A. Porrello *et al.*, "Dark experience for general continual learning: a strong, simple baseline," *Proc. of Advances in Neural Information Processing Systems*, vol. 33, pp. 15 920–15 930, 2020.

[35] Y. Huang, W. Zhao, S. Wang, Y. Fu, Y. Jiang, and L. Yu, "ConSlide: Asynchronous hierarchical interaction transformer with breakup-reorganize rehearsal for continual whole slide image analysis," in *Proc. of IEEE Intl. Conf. on Computer Vision*, 2023, pp. 21 349–21 360.

[36] L. Le Bescond, M. Lerousseau, F. Andre, and H. Talbot, "Sparsexmil: Leveraging spatial context for classifying whole slide images in digital pathology," 2024.

[37] D. Tellez, G. Litjens, P. Bándi, W. Bulten, J.-M. Bokhorst, F. Ciompi, and J. Van Der Laak, "Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology," *Medical image analysis*, vol. 58, p. 101544, 2019.

[38] L. Berrada, A. Zisserman, and P. Mudigonda, "Smooth loss functions for deep top-k classification," in *International Conference on Learning Representations (ICLR), 2018*, 2018.

[39] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[40] X. Wang, S. Yang, J. Zhang, M. Wang, J. Zhang, W. Yang, J. Huang, and X. Han, "Transformer-based unsupervised contrastive learning for histopathological image classification," *Medical image analysis*, vol. 81, p. 102559, 2022.