Continual Multiple Instance Learning with Enhanced Localization for Histopathological Whole Slide Image Analysis

Supplementary Material

S1. Experimental Setups

S1.1. Details on Datasets

We conducted tumor detection using the CAMELYON-16 (CM-16) [S1] and Pathology AI Platform (PAIP) [S8, S9, S18] datasets to evaluate continual and joint instance classification. For continual bag classification, we conducted tumor subtyping using The Cancer Genome Atlas (TCGA) [S17].

CAMELYON-16 (CM-16) is a WSI dataset for diagnosing breast cancer metastases in sentinel lymph nodes. It consists of 400 WSIs with corresponding pixel-level tumor annotations, officially split into 270 training slides and 130 test slides. Following [S3, S13], we merged the official training and test sets and performed three-times threefold cross-validation to ensure that each slide is used for both training and testing. This cross-validation strategy helps mitigate the impact of data partitioning and random seed selection on model evaluation. The numbers of tumor and non-tumor slides in CM-16 are summarized in Tab. S1.

Repository	Organ	# normal slides	# tumor slides
CM-16	Sentinel Lymph Nodes	241	159
PAIP	Liver	251	252
	Prostate	299	300
	Pancreas	207	207
	Colon	449	449

Table S1. Datasets for continual instance classification for WSI tumor detection, constructed by organ datasets from CAMELYON-16 (CM-16) [S1] and Pathology AI Platform (PAIP) [S8, S9, S18].

Pathology Artificial Intelligence Platform (PAIP) is a platform for developing learning-based models for WSI analysis, particularly for tumor diagnosis. PAIP consists of hundreds of WSIs across six different organs, each with corresponding pixel-level tumor annotations. Among the available organ datasets, we utilized the liver, prostate, pancreas, and colon datasets. However, these datasets only provide tumor slides, meaning that all slide-level labels correspond to the tumor class. Since application to a MIL setup requires to leverage both tumor and normal slide-level annotations as weak labels, we exploited the MIL formulation where each slide is treated as a bag of multiple instances (patches). Specifically, for each organ dataset in PAIP, we randomly split the slides into two halves — one half designated as tumor slides, and the other as normal slides. For the normal slides, we removed all tumor regions prior to patch extraction to ensure they only contain normal patches. The resulting numbers of tumor and normal slides for each organ in PAIP are summarized in Tab. S1. Similar to CAMELYON-16, we applied three-times threefold cross-validation to each organ dataset to mitigate the effect of data partitioning and random seed selection on model evaluation.

The Cancer Genome Atlas (TCGA) is a large-scale research project jointly conducted by the National Cancer Institute (NCI) and the National Human Genome Research Institute (NHGRI). It aims to systematically analyze genomic alterations in various types of cancer. TCGA provides WSIs from diverse organs along with information on corresponding tumor subtypes, enabling weakly-supervised tumor subtyping tasks. For continual bag classification, we conducted continual tumor subtyping tasks across four organs from TCGA like NSCLC, BRCA, RCC, and ESCA, following the setup of ConSlide [S6]. Detailed statistics for each dataset are summarized in Tab. S2.

Dataset	Tumor type	# slides
NSCLC	Lung adenocarcinoma (LUAD)	492
	Lung squamous cell carcinoma (LUSC)	466
BRCA	Invasive ductal (IDC)	726
	Invasive lobular carcinoma (ILC)	149
RCC	Clear cell renal cell carcinoma (CCRCC)	498
	Papillary renal cell carcinoma (PRCC)	289
ESCA	Esophageal adenocarcinoma (ESAD)	65
	Esophageal squamous cell carcinoma (ESCC)	89

Table S2. Datasets for continual bag classification for WSI tumor subtyping tasks, constructed by organ datasets from The Cancer Genome Atlas (TCGA) [S17].

S1.2. Details on Evaluation Metrics

S1.2.1. Continual Instance Classification.

Instance-level accuracy (Acc_{inst}) calculates the average instance-level accuracy across tasks after completing the training of all continual tasks. Let R^{inst}_{nl} denote the instance-level accuracy on the l-th task after training on the n-th task. For a total of N continual tasks, Acc_{inst} is computed as:

$$\text{Acc}_{\text{inst}} = \frac{1}{N}\sum_{l=1}^{N} R^{\text{inst}}_{Nl} \tag{S1}$$

Intersection over Union (IoU) and **Dice score** after the final task in continual MIL can be measured in a similar manner. Let IoU_{nl} and Dice_{nl} denote the IoU and Dice score, respectively, on the l-th task after training on the n-th task. Then, the IoU and Dice scores after completing all N continual tasks are measured as:

$$\text{IoU} = \frac{1}{N}\sum_{l=1}^{N}\text{IoU}_{Nl}, \quad \text{Dice} = \frac{1}{N}\sum_{l=1}^{N}\text{Dice}_{Nl} \tag{S2}$$

Forget on instance-level accuracy (Forget$_{\text{inst}}$) quantifies the degree of forgetting of the MIL model on previously learned knowledge as new tasks are introduced, measured in terms of instance-level accuracy. For its task, it measures the gap between the best performance of the task attained during training on sequential tasks and the final performance on the same task after training on all subsequent tasks. Then, it averages the measures over all tasks as:

$$\text{Forget}_{\text{inst}} = \frac{1}{N-1}\sum_{l=1}^{N-1}\max_{n\in\{l,\cdots,N-1\}} R^{\text{inst}}_{n,l} - R^{\text{inst}}_{N,l}. \tag{S3}$$

S1.2.2. Continual Bag Classification.

Bag-level accuracy (Acc$_{\text{bag}}$), similar to continual instance classification, measures the bag-level accuracy after training on all N continual tasks, which is defined as:

$$\text{Acc}_{\text{bag}} = \frac{1}{N}\sum_{l=1}^{N} R^{\text{bag}}_{Nl} \tag{S4}$$

where R^{bag}_{nl} is the bag-level accuracy on the l-th task after training on the n-th task.

Forget on bag-level accuracy (Forget$_{\text{bag}}$) is defined analogously to Forget$_{\text{inst}}$ as:

$$\text{Forget}_{\text{inst}} = \frac{1}{N-1}\sum_{l=1}^{N-1}\max_{n\in\{l,\cdots,N-1\}} R^{\text{bag}}_{n,l} - R^{\text{bag}}_{N,l}. \tag{S5}$$

Masked bag-level accuracy (M.Acc$_{\text{bag}}$) measures the average accuracy computed by restricting classification to only the classes within a task when the task index is given at test time. Let $R^{l,\text{bag}}_{Nl}$ be the bag-level accuracy of l-th task after training on N-th task measured within it class set, under the assumption that the test class is known to belong to the l-th task. Then, M.Acc$_{\text{bag}}$ can be represented as:

$$\text{M.Acc}_{\text{bag}} = \frac{1}{N}\sum_{l=1}^{N} R^{l,\text{bag}}_{Nl} \tag{S6}$$

S1.3. Further Implementation Details

We followed CLAM [S13] for patch and feature extraction. Instead of utilizing ResNet-50 like previous works [S7, S12, S15, S19], we utilized UNI [S4], the foundation model for computational pathology, as pre-trained feature extractor since it provides improved patch-wise representation for a WSI, resulting in overall enhanced localization results across all MIL and CL methods.

For training continual MIL, we trained each task for 100 epochs. For optimization, we adopted Adam [S10] with an initial learning rate of 1×10^{-4} and a weight decay of 1×10^{-5}. To adjust the learning rate during training, we applied a cosine annealing schedule. Due to the large size of whole slide images (WSIs), we used a batch size of 1. All experiments for both continual and MIL approaches were conducted on a single NVIDIA A6000 GPU.

S2. Further Discussions on CoMEL

S2.1. Efficiency studies

Tab. S3 compares the computation and memory efficiency of GDAT, RRT-MIL, and smAP. Compared to RRT-MIL, GDAT achieves improved VRAM memory consumption, FLOPs, and inference latency, all with a comparable number of learnable parameters. Note that from Tab. 1, GDAT outperforms RRT-MIL not only in terms of MIL with localization performance but also in terms of memory and computational efficiency. While smAP exhibits lower memory consumption than GDAT, it suffers from significantly higher inference latency due to the computation of the adjacency map. Although smAP enhances localization performance as shown in Tab. 1, its lack of a scalable module hinders effective synergy with BPPL.

Methods	RRT-MIL	smAP	**GDAT**
Params (M)	6.00	0.60	**6.42**
VRAM (G)	10.7	3.6	**8.8**
FLOPs (G)	57.4	20.6	**23.4**
Latency (ms)	3.36	3526	**2.58**

Table S3. Efficiency analysis of GDAT and OWLoRA.

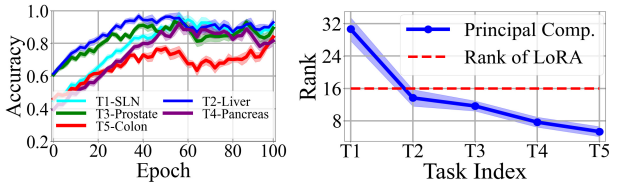
Tab. S4 compares the memory efficiency of OWLoRA with that of ER and ConSlide. ER and ConSlide are rehearsal-based approaches that require an additional memory buffer, whereas OWLoRA relies on additional learnable parameters. Then, it is evident that OWLoRA is more memory-efficient than the baselines comparing the memory size of their rehearsal buffers to the size of the additional parameters in OWLoRA.

Methods	ER/30	ConSlide/30	**OWLoRA**
Params (M)	-	-	**4.6×10^{-1}**
Buffer (M)	4.5×10^{2}	4.7×10^{2}	**-**

Table S4. Efficiency analysis of OWLoRA.

S2.2. Pseudo-label Accuracy of BPPL

To evaluate the robustness of BPPL in terms of pseudo-label quality, we measured the pseudo-label accuracy for each task during training continual instance classification. From Fig. S1(a), BPPL consistently improves the quality of pseudo-labels as training progresses, thereby enhancing localization performance.



(a) Pseudo-label accuracy (b) Rank study of OWLoRA

Figure S1. Pseudo-label accuracy of BPPL and analysis on low-rank property of MIL tasks, supporting the rationale of OWLoRA.

S2.3. Low-rank Property of MIL Tasks

Through extensive experiments on continual instance classification, we demonstrated that OWLoRA effectively mitigates forgetting in both bag- and instance-level classification. This is attributed to the low-rank nature of each task and OWLoRA's orthogonality-based regularization. Fig. S1(b) illustrates the number of new singular values required to capture 99% of the total norm of singular values (across 3 runs), confirming the inherent low-rankedness. Forgetting is further mitigated by enforcing orthogonal subspaces, as supported by Tab. S9 and Tab. S10.

S2.4. Ablation studies on GDAT

The term ηV in Eq. (4) is introduced to mitigate the reduction in diversity among instance features caused by the grouped attention mechanism. Here, η is a hyperparameter. Tab. S5 shows the performance of continual instance classification on the combined CM-16 and PAIP datasets with different values of η. We can see that smaller η resulted in worse instance-level accuracy since the reduced diversity among instance features leads to poorer instance discriminability. Meanwhile, when η becomes excessively large, the effect of attention-based feature refinement diminishes, ultimately resulting in degraded performance.

η	ACC_{inst} (↑)	$\text{Forget}_{\text{inst}}$ (↓)	ACC_{bag} (↑)
0	68.94 ± 2.49	14.96 ± 2.39	62.11 ± 1.57
0.01	69.32 ± 2.74	13.39 ± 2.52	61.72 ± 1.92
0.05	74.13 ± 2.39	12.94 ± 2.62	62.84 ± 1.61
0.1	$\mathbf{74.15 \pm 1.97}$	$\mathbf{13.05 \pm 2.31}$	$\mathbf{62.64 \pm 2.14}$
0.5	73.93 ± 1.72	14.33 ± 2.10	60.23 ± 1.75
1	73.24 ± 2.19	13.93 ± 2.38	58.18 ± 2.30
5	70.98 ± 1.89	15.47 ± 1.66	56.15 ± 2.08
10	69.17 ± 2.09	14.76 ± 2.39	53.21 ± 1.82

Table S5. Ablation studies of the effect of η in GDAT (3 runs). The row with **bold** represents the selected configurations.

We investigated the impact of different hyper-parameters on GDAT performance. Relevant hyper-parameters include the grouping window factor F and the number of attention layers in GDAT. When the grouping window factor is set to F, the number of instances within a group is F^2, and the number of groups is $m = M/F^2$. Tab. S6 shows that increasing F deteriorated the performance since unrelated instances could be grouped together and reduce the representation diversity for instances by the attention of GDAT. Meanwhile, increasing the number of attention layers in GDAT resulted in worse performance due to overfitting.

F	# of layers	ACC_{inst} (↑)	$\text{Forget}_{\text{inst}}$ (↓)	ACC_{bag} (↑)
8	2	74.20 ± 2.24	14.23 ± 2.04	62.92 ± 2.52
32	2	72.25 ± 2.56	14.57 ± 2.66	61.87 ± 2.40
64	2	68.23 ± 3.23	16.23 ± 2.43	56.24 ± 3.24
16	**2**	$\mathbf{74.15 \pm 1.97}$	$\mathbf{13.05 \pm 2.31}$	$\mathbf{62.64 \pm 2.14}$
16	1	72.43 ± 2.39	14.54 ± 2.46	60.83 ± 2.32
16	4	74.45 ± 2.08	13.33 ± 1.94	62.82 ± 1.72
16	8	73.20 ± 2.55	13.92 ± 2.72	61.33 ± 2.40
16	12	72.01 ± 2.04	14.90 ± 2.44	60.89 ± 2.06

Table S6. Ablation studies of the effect of grouping window factor F and the number of attention layers in GDAT (3 runs). The row with **bold** represents the selected configurations.

S2.5. Ablation studies on BPPL

We investigated the impact of different hyper-parameters on BPPL performance. Relevant hyper-parameters include temperature T and coefficient τ. Tab. S7 shows that increasing T too much deteriorates the performance due to small probability values, while decreasing T too much deteriorates the performance due to large probability values. Additionally, increasing τ too much deteriorated the performance due to strict pseudo-labeling, while decreasing τ deteriorated the performance due to inaccurate pseudo-labels.

T	τ	ACC_{inst} (↑)	$\text{Forget}_{\text{inst}}$ (↓)	ACC_{bag} (↑)
0.01	0.35	60.32 ± 3.52	13.02 ± 3.09	60.92 ± 2.10
0.05	0.35	68.23 ± 2.69	12.43 ± 2.72	61.52 ± 1.93
0.1	0.35	73.23 ± 2.46	13.49 ± 2.42	62.04 ± 2.01
1	0.35	73.12 ± 2.98	13.92 ± 2.19	62.40 ± 1.92
5	0.35	62.23 ± 2.82	14.92 ± 2.39	61.24 ± 1.82
0.5	**0.35**	$\mathbf{74.15 \pm 1.97}$	$\mathbf{13.05 \pm 2.31}$	$\mathbf{62.64 \pm 2.14}$
0.5	0.05	61.23 ± 3.22	15.12 ± 2.72	61.23 ± 2.92
0.5	0.1	64.23 ± 2.32	17.23 ± 1.98	61.63 ± 1.62
0.5	0.25	70.37 ± 2.23	14.62 ± 2.93	61.73 ± 1.40
0.5	0.45	71.24 ± 2.42	14.34 ± 3.09	63.20 ± 2.20

Table S7. Ablation studies of the effect of T and τ in BPPL (3 runs). The row with **bold** represents the selected configurations.

We also investigated the impact of additional hyper-parameters, including the regularization coefficients λ_1 and λ_2. Tab. S8 shows that increasing λ_1 too much deteriorated the performance due to placing too much weight on instance localization, while decreasing λ_1 deteriorated the performance on instance classification due to less weight on training for localization. Furthermore, increasing λ_2 too much

deteriorated the performance due to excessive bag separation, while decreasing λ_2 deteriorated the performance due to insufficient bag separation.

λ_1	λ_2	ACC_{inst} (↑)	$\text{Forget}_{\text{inst}}$ (↓)	ACC_{bag} (↑)
0.1	1	72.32 ± 2.33	13.89 ± 2.87	62.47 ± 1.98
1	1	73.84 ± 2.93	13.24 ± 2.52	62.92 ± 2.52
5	1	67.23 ± 1.92	11.23 ± 1.52	57.24 ± 1.62
10	1	61.23 ± 2.32	9.23 ± 2.24	54.23 ± 2.20
0.5	**1**	$\mathbf{74.15 \pm 1.97}$	$\mathbf{13.05 \pm 2.31}$	$\mathbf{62.64 \pm 2.14}$
0.5	0.1	70.92 ± 2.60	17.34 ± 2.72	62.82 ± 2.12
0.5	0.3	71.72 ± 2.72	15.94 ± 2.11	62.52 ± 1.82
0.5	3	70.24 ± 1.82	16.68 ± 1.43	61.24 ± 2.67
0.5	10	62.23 ± 2.53	12.35 ± 3.19	55.23 ± 3.06

Table S8. Ablation studies of the effect of λ_1 and λ_2 in BPPL (3 runs). The row with **bold** represents the selected configurations.

S2.6. Ablation studies on OWLoRA

We investigated the impact of ϵ for first task and the rank d for subsequent tasks in OWLoRA. Tab. S9 shows decreasing ϵ deteriorated the performance on instance classification. It is attributed to the forgetting of the first task. Meanwhile, decreasing the rank d resulted in worse instance classification by the reduced adaptability to new tasks.

ϵ	d	ACC_{inst} (↑)	$\text{Forget}_{\text{inst}}$ (↓)	ACC_{bag} (↑)
0.5	16	68.23 ± 2.30	19.47 ± 2.73	56.10 ± 2.49
0.7	16	70.42 ± 2.29	16.78 ± 2.19	59.35 ± 2.44
0.9	16	71.23 ± 2.49	15.04 ± 2.59	61.23 ± 2.87
0.999	16	74.22 ± 2.05	12.99 ± 2.10	62.24 ± 2.19
0.99	**16**	$\mathbf{74.15 \pm 1.97}$	$\mathbf{13.05 \pm 2.31}$	$\mathbf{62.64 \pm 2.14}$
0.99	2	67.23 ± 1.88	20.44 ± 1.84	58.38 ± 1.61
0.99	4	70.04 ± 1.71	17.05 ± 1.79	59.92 ± 2.04
0.99	8	72.86 ± 2.65	14.98 ± 1.94	61.04 ± 1.98
0.99	32	74.19 ± 2.04	13.37 ± 2.46	63.33 ± 2.46

Table S9. Ablation studies of the impact of ϵ for first task and d for subsequent tasks in OWLoRA for continual instance classification (3 runs). The row with **bold** represents the selected configurations.

We further investigated the impact of the hyperparameter λ_3 on the performance of OWLoRA. Tab. S10 shows that increasing λ_3 deteriorated the performance because it excessively enforces the orthogonality on the basis for mitigating forgetting, which hinders to learn new tasks. Meanwhile, decreasing λ_3 too much also deteriorated the performance because of the inability to mitigate forgetting.

S2.7. Additional Structural Ablation Studies

We further performed additional structural ablation studies to investigate the design choices of GDAT and BPPL. From Tab. S11, GDAT (double attention) achieves competitive performance to single attention with better scalability. From Tab. S12, BPPL fails to learn discriminative prototypes without $\mathcal{L}_{\text{sep}}$, reducing performance. Since three filtering factors in BPPL are crucial for high-quality pseudo-labels, removing any of them leads to degradation in localization.

λ_3	ACC_{inst} (↑)	$\text{Forget}_{\text{inst}}$ (↓)	ACC_{bag} (↑)
0.1	70.32 ± 2.91	17.50 ± 2.72	55.33 ± 3.24
0.5	72.24 ± 2.33	14.88 ± 2.29	60.23 ± 2.57
1	$\mathbf{74.15 \pm 1.97}$	$\mathbf{13.05 \pm 2.31}$	$\mathbf{62.64 \pm 2.14}$
5	74.21 ± 2.04	13.08 ± 2.46	61.94 ± 2.22
10	71.39 ± 1.64	16.42 ± 1.82	60.92 ± 1.50

Table S10. Ablation studies of the effect of λ_3 in OWLoRA (3 runs). The row with **bold** represents the selected configurations.

Model	ACC_{bag} (↑)	ACC_{inst} (↑)	**IoU** (↑)	**Dice** (↑)
Single Attn	73.53 ± 1.64	80.87 ± 1.82	50.69 ± 2.74	62.32 ± 2.35
GDAT	$\mathbf{72.94 \pm 1.28}$	$\mathbf{80.55 \pm 2.34}$	$\mathbf{50.35 \pm 3.43}$	$\mathbf{61.70 \pm 2.87}$

Table S11. Comparison of single attention and GDAT (double attention) when **combined with BPPL** on merged dataset (3 runs).

Ablating Comp.	ACC_{inst} (↑)	$\text{Forget}_{\text{inst}}$ (↓)	**Dice** (↑)	ACC_{bag} (↑)
CoMEL	$\mathbf{74.15 \pm 1.97}$	$\mathbf{13.05 \pm 2.31}$	$\mathbf{52.27 \pm 2.42}$	$\mathbf{62.64 \pm 2.14}$
w/o $\mathcal{L}_{\text{sep}}$	69.82 ± 2.29	17.86 ± 2.43	45.60 ± 2.56	62.84 ± 2.18
w/o $\mathbb{1}(\hat{Y} = Y)$	66.2 ± 2.1	15.0 ± 1.9	31.6 ± 2.4	60.4 ± 1.6
w/o $\mathbb{1}(\hat{p}_m > \tau_1)$	69.28 ± 1.76	15.84 ± 1.96	37.51 ± 2.65	61.40 ± 1.66
w/o $\mathbb{1}(Y_{\text{bag}} \in \mathcal{Y}_{\text{pos}})$	66.54 ± 1.72	15.81 ± 1.73	34.77 ± 2.23	60.89 ± 2.32

Table S12. Ablation studies of loss components and filtering factors in BPPL on continual instance classification (3 runs).

S3. Additional Quantitative Results

S3.1. Results on Continual Instance Classification

Tab. S13 presents additional quantitative results for tumor detection across the sequential datasets of combined CM16 and PAIP, but the sequence is reversed from the experiment in the main text. That is, the sequence is: colon, pancreas, prostate, liver, and sentinel lymph nodes. From the results for ACC_{bag} and ACC_{inst}, we observed that while rehearsal-based approaches mitigated forgetting in bag classification with better performance than regularization-based approaches, they still suffered from substantial forgetting in instance classification similar to our experiment in the main text. On the other hand, InfLoRA outperformed them in both ACC_{inst} and $\text{Forget}_{\text{inst}}$ with competitive performance in terms of ACC_{bag} also similar to our initial experiment. Upon the LoRA-based CL approach, our CoMEL achieved the best performance across all metrics. In particular, CoMEL outperformed in terms of IoU and Dice by a large margin, demonstrating its effectiveness in preserving localization in the continual MIL even when the sequence is reversed.

S3.2. Results on Continual Bag Classification

To further evaluate the performance on continual bag classification, we compared CoMEL against the same base-

IL Type	Method	ACC_{inst} (↑)	$\text{Forget}_{\text{inst}}$ (↓)	IoU (↑)	Dice (↑)	ACC_{bag} (↑)
Upper Bound	Joint (Full label)	90.32 ± 3.34	-	67.72 ± 2.04	77.23 ± 2.39	75.97 ± 3.71
	Joint (Weak label)	79.50 ± 2.72	-	51.67 ± 2.19	61.09 ± 2.64	72.28 ± 3.49
Lower Bound	Finetune	56.00 ± 2.17	29.12 ± 3.37	10.82 ± 2.30	18.71 ± 2.39	9.42 ± 3.58
Regularization-based	EWC	60.65 ± 2.88	23.79 ± 2.37	13.81 ± 3.00	20.75 ± 2.52	15.37 ± 2.79
	LwF	61.83 ± 2.87	24.30 ± 2.26	16.29 ± 2.92	23.96 ± 3.19	18.82 ± 3.23
Rehearsal-based	A-GEM/30	63.19 ± 3.33	22.78 ± 3.39	18.48 ± 2.61	25.23 ± 3.64	38.65 ± 2.57
	ER/30	63.27 ± 2.83	23.01 ± 3.72	17.16 ± 3.11	24.80 ± 3.08	43.70 ± 2.31
	ER/100	67.09 ± 3.28	$\underline{19.49 \pm 2.84}$	21.29 ± 3.02	29.12 ± 3.30	44.82 ± 3.43
	DER++/30	65.83 ± 2.74	21.76 ± 3.00	21.03 ± 3.11	28.06 ± 2.52	45.79 ± 3.02
	ER-ACE/30	66.64 ± 3.58	20.84 ± 3.10	17.64 ± 3.40	24.83 ± 3.19	48.02 ± 3.63
	ConSlide/30	65.52 ± 3.12	21.27 ± 3.13	19.97 ± 2.56	25.67 ± 3.60	52.92 ± 3.29
Prompt-tuning-based	QPMIL-VL [S5]	66.37 ± 2.21	20.51 ± 2.17	25.15 ± 2.35	34.28 ± 2.19	$\underline{54.83 \pm 2.39}$
LoRA-based	LoRA finetune	61.27 ± 3.03	22.95 ± 2.45	21.44 ± 3.02	28.65 ± 3.72	28.01 ± 3.45
	InfLoRA	$\underline{67.60 \pm 2.84}$	20.37 ± 3.65	$\underline{30.33 \pm 3.42}$	$\underline{39.54 \pm 3.14}$	51.53 ± 2.61
	CoMEL (Ours)	$\mathbf{70.65 \pm 2.84}$	$\mathbf{16.93 \pm 2.79}$	$\mathbf{38.81 \pm 3.28}$	$\mathbf{48.22 \pm 3.69}$	$\mathbf{57.72 \pm 3.26}$

Table S13. Additional quantitative results of CL methods on instance classification in the reversed continual MIL setup. The best and second best results are marked as **bold** and underline. Each experiment consisted of 10 runs. The experiments were conducted on five sequential organ datasets from combined CM-16 and PAIP. For baselines, we applied the CL approaches upon our GDAT+BPPL, except for ConSlide. All metrics are reported in percentages. CoMEL achieved the highest performance across all metrics while minimizing the forgetting.

Model	ACC_{bag} (↑)	AUC_{bag} (↑)	F1_{bag} (↑)	ACC_{inst} (↑)	IoU (↑)	Dice (↑)
ABMIL [S13]	91.13 ± 1.34	94.21 ± 1.07	86.42 ± 1.39	77.53 ± 2.28	38.22 ± 3.01	49.33 ± 2.71
DS-MIL [S12]	90.91 ± 1.59	93.83 ± 1.24	86.11 ± 1.48	72.18 ± 2.17	28.59 ± 2.64	37.81 ± 2.42
TransMIL [S14]	91.51 ± 1.33	94.02 ± 1.19	87.25 ± 1.36	76.03 ± 2.51	35.11 ± 2.41	45.08 ± 2.29
RRT-MIL [S16]	$\underline{93.23 \pm 1.08}$	$\underline{96.14 \pm 1.13}$	89.53 ± 0.94	73.44 ± 1.98	32.14 ± 1.97	41.01 ± 2.14
smAP [S2]	91.42 ± 1.21	94.06 ± 1.38	87.63 ± 1.31	$\underline{86.01 \pm 1.95}$	$\underline{47.12 \pm 2.51}$	$\underline{59.11 \pm 2.67}$
GDAT (Ours)	$\mathbf{93.21 \pm 1.31}$	$\mathbf{96.56 \pm 1.14}$	$\mathbf{90.02 \pm 0.97}$	78.98 ± 2.36	39.91 ± 2.74	51.61 ± 2.53
GDAT+BPPL (Ours)	93.04 ± 1.42	95.74 ± 1.25	$\underline{89.62 \pm 1.27}$	$\mathbf{87.73 \pm 2.61}$	$\mathbf{53.21 \pm 3.04}$	$\mathbf{64.17 \pm 3.23}$

Table S14. Comparison of different MIL models on a single dataset CM-16. We evaluated bag classification using ACC_{bag}, AUC_{bag}, and F1_{bag}. For instance classification, We evaluated using ACC_{inst}, IoU, and Dice score. The best and second-best results are marked as **bold** and underline, respectively. Each experiment consisted of 10 runs. Our proposed method achieved the best performance across all metrics, demonstrating its superiority in instance-level classification.

Method	ACC_{bag} (↑)	$\text{Forget}_{\text{bag}}$ (↓)	$\text{M.ACC}_{\text{bag}}$ (↑)
Joint	91.18 ± 2.31	-	93.42 ± 2.53
Finetune	24.41 ± 3.74	66.13 ± 3.87	79.36 ± 3.43
EWC	25.02 ± 4.11	64.58 ± 3.98	84.02 ± 3.66
LwF	26.61 ± 3.73	62.35 ± 3.40	87.91 ± 3.94
A-GEM/30	45.62 ± 4.03	48.42 ± 4.14	87.44 ± 4.38
ER/30	67.79 ± 2.94	25.48 ± 3.07	89.33 ± 5.28
ER/100	70.66 ± 2.83	23.61 ± 2.98	90.21 ± 3.38
DER++/30	68.94 ± 4.02	24.11 ± 4.23	89.76 ± 3.97
ConSlide/30	76.21 ± 2.93	17.66 ± 3.13	90.01 ± 3.24
QPMIL-VL	79.92 ± 3.54	$\underline{13.24 \pm 3.28}$	$\underline{90.80 \pm 2.95}$
LoRA finetune	45.89 ± 4.22	38.52 ± 3.80	87.07 ± 3.21
InfLoRA	$\underline{80.31 \pm 4.14}$	13.59 ± 3.64	88.46 ± 3.72
CoMEL (Ours)	$\mathbf{83.27 \pm 3.63}$	$\mathbf{11.51 \pm 3.78}$	$\mathbf{91.42 \pm 3.17}$

Table S15. Additional comparison of methods for slide-level classification on the TCGA dataset with reversed task order. The best and second-best results are marked as **bold** and underline. Each experiment consisted of 10 runs.

lines in the continual instance classification, but with the sequence reversed. That is, the sequence of organ for continual tasks is: ESCA, RCC, BRCA, and NSCLC. Tab. S15 illustrates the performance of CL baselines and CoMEL for continual bag classification on the TCGA datasets with reverse sequence. Just like in the initial experiment, CoMEL achieved the highest performance in terms of ACC_{bag} and $\text{Forget}_{\text{bag}}$, demonstrating its strong performance while effectively preserving previously learned knowledge. Consistently, CoMEL also achieved the highest masked bag-level accuracy ($\text{M.ACC}_{\text{bag}}$) compared to the baselines. Rehearsal-based approaches such as ER and ConSlide demonstrated their effectiveness in mitigating catastrophic forgetting for continual bag classification, while regularization-based methods suffered from severe forgetting.

S3.3. Results on Single Dataset

We further evaluated our method's bag and instance classification ability on the single dataset CM-16. From Tab. S14, we can see that GDAT achieved the best performance on bag classification compared to the baselines, demonstrating its effectiveness. Furthermore, the removal of BPPL from GDAT results in a notable performance drop in instance

classification for localization. This indicates that BPPL is an important component for effective instance localization.

S3.4. Additional Metrics with Various Backbones

Tab. S16 presents the results of additional metrics that were omitted from Tab. 4. It demonstrates that CoMEL performs well across various backbones for the missing metrics.

Model	ResNet50	PLIP	CONCH	UNI
IoU (↑)	36.19 ± 2.24	39.34 ± 1.87	37.16 ± 2.31	42.78 ± 1.93
ACC$_{\text{bag}}$ (↑)	55.29 ± 1.47	59.22 ± 2.04	54.20 ± 1.49	61.59 ± 2.11
AUC$_{\text{bag}}$ (↑)	88.86 ± 1.21	89.38 ± 1.03	88.27 ± 1.86	90.22 ± 1.74

Table S16. Additional metrics with different feature extractors on continual instance classification.

S4. Additional Qualitative Results

We provide additional qualitative results for instance classification under continual MIL setup. From Fig. S2 to Fig. S5, we can see that CoMEL can preserve the localized tumor region after training on all five organ datasets.

S5. Limitations

In this work, we considered only a fixed sequence of disjoint tasks for the continual MIL setup. For example, the MIL models learn from datasets of distinct organs or subtypes for WSI analysis in our experiments. However, in real-world hospital settings, WSIs collected over a certain period include a mixture of various organs and tumor subtypes. In continual learning, such configuration has already been studied under the concept of blurry tasks [S11], where task boundaries are ambiguous. In this work, we did not consider such blurry tasks, leaving it as an interesting future work. Bag Prototypes-based Pseudo-Labeling (BPPL) module heavily relies on the accuracy of attention scores as pseudo-labels. Our ablation studies on BPPL hyperparameters in Tab. S7 and Tab. S8 indicate that localization performance is highly sensitive to hyperparameter selection which influences pseudo-label quality. The Orthogonal Weighted Low-Rank Adaptation (OWLoRA) effectively mitigated catastrophic forgetting in previous tasks. However, the basis of new tasks cannot be introduced indefinitely, as the maximum rank of a matrix is upper-bounded by its dimension. Therefore, OWLoRA has inherent limitations in learning an infinite sequence of sequential tasks.

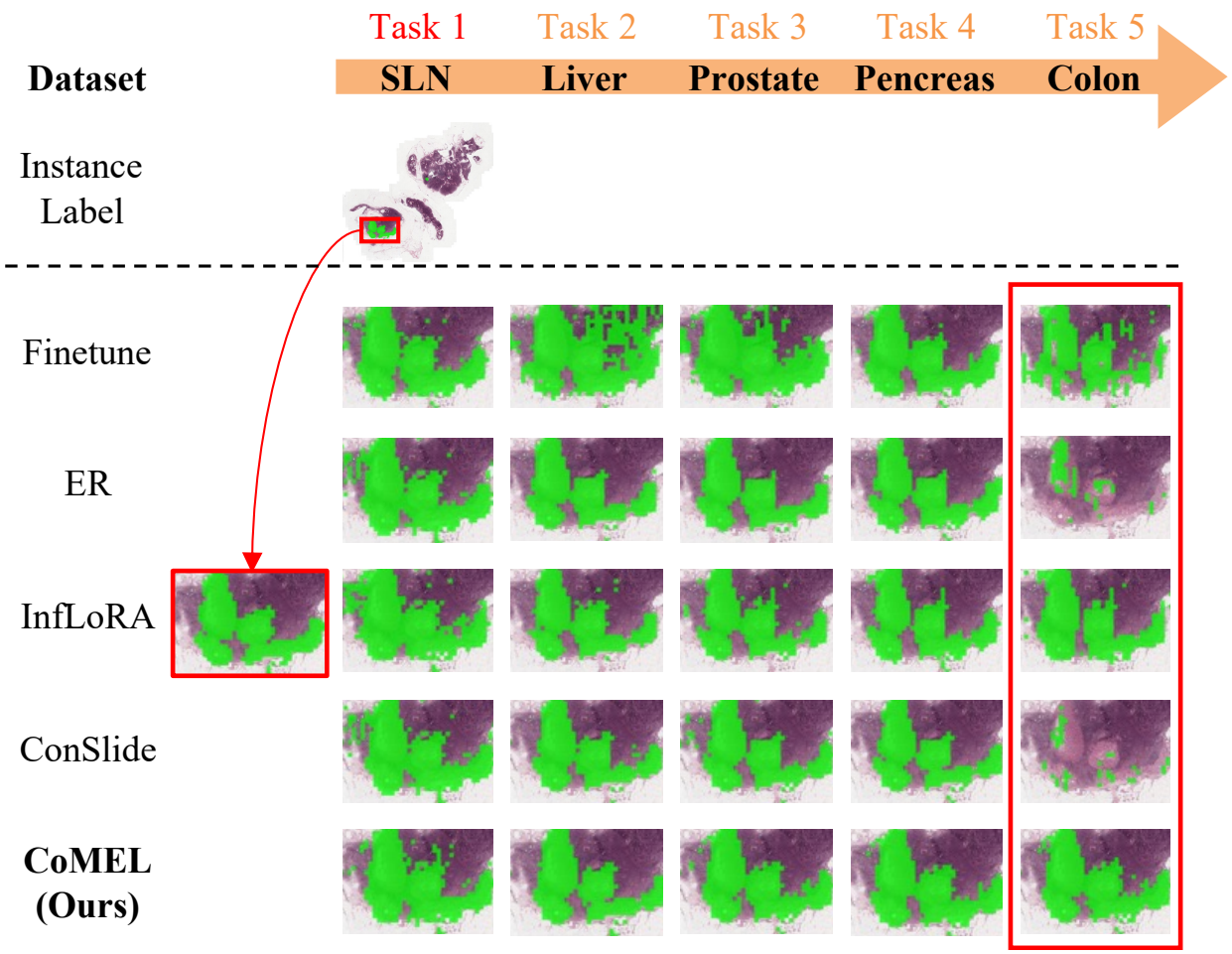


Figure S2. Additional qualitative results of localization across sequential organ datasets under the continual MIL setup. Each column is the localization performance on Task 1 as the learned organ changes over sequential tasks. Each row corresponds to CL methods including CoMEL. CoMEL successfully preserved the localization quality across all tasks, while baselines increase false positives or false negatives.

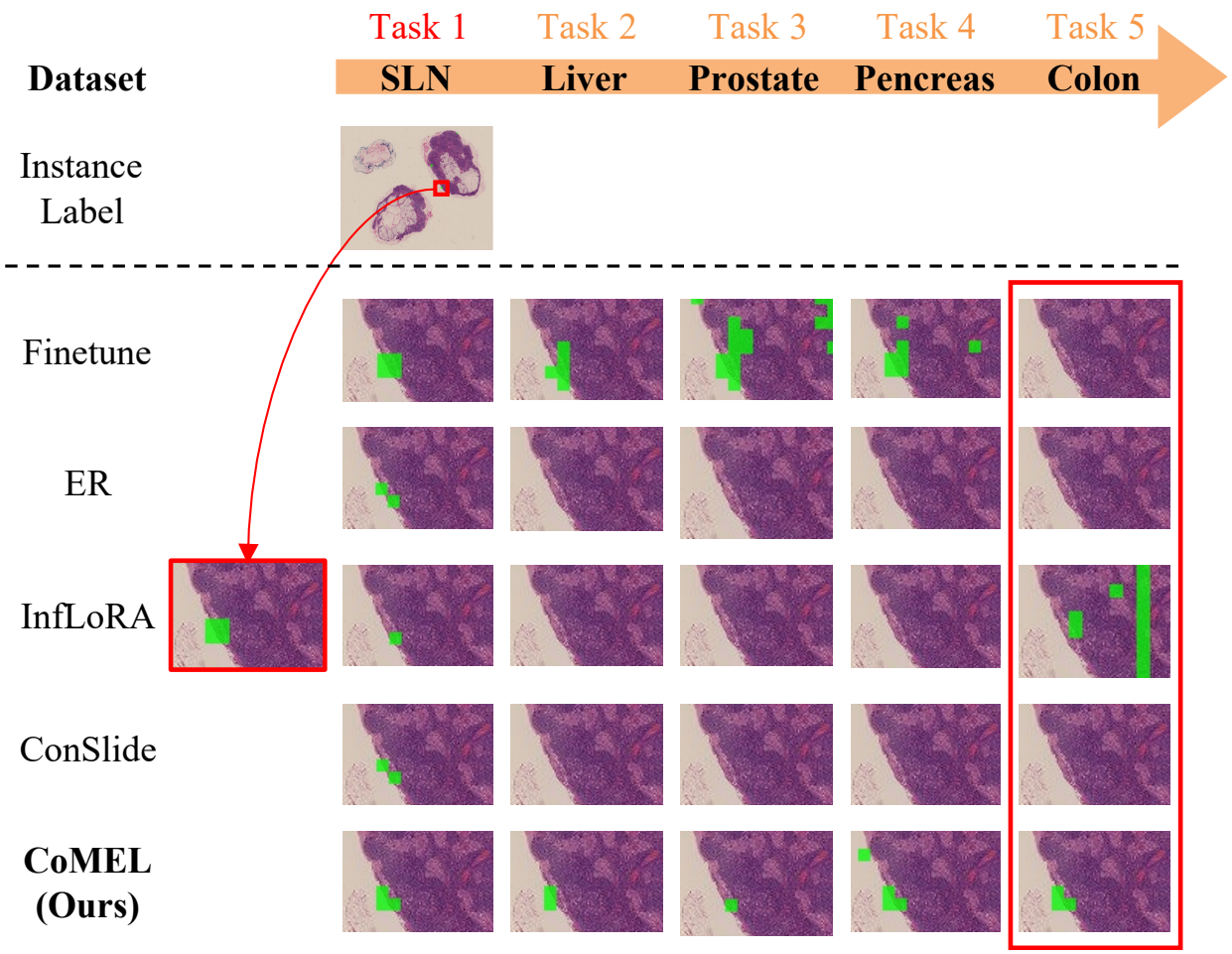


Figure S3. Additional qualitative results of localization across sequential organ datasets under the continual MIL setup. Each column is the localization performance on Task 1 as the learned organ changes over sequential tasks. Each row corresponds to CL methods including CoMEL. CoMEL successfully preserved the localization quality across all tasks, while baselines increase false positives or false negatives.

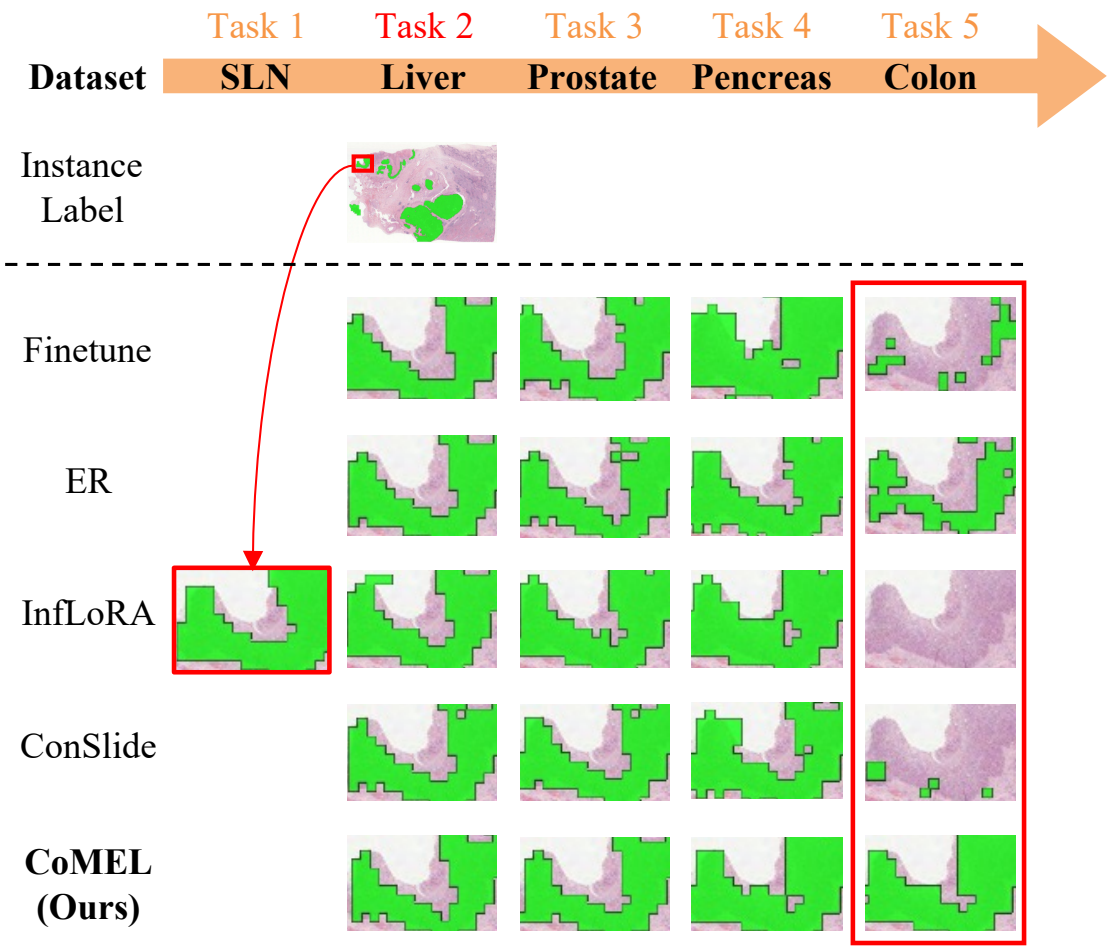


Figure S4. Additional qualitative results of localization across sequential organ datasets under the continual MIL setup. Each column is the localization performance on Task 2 as the learned organ changes over sequential tasks. Each row corresponds to CL methods including CoMEL. CoMEL successfully preserved the localization quality across all tasks, while baselines increase false positives or false negatives.

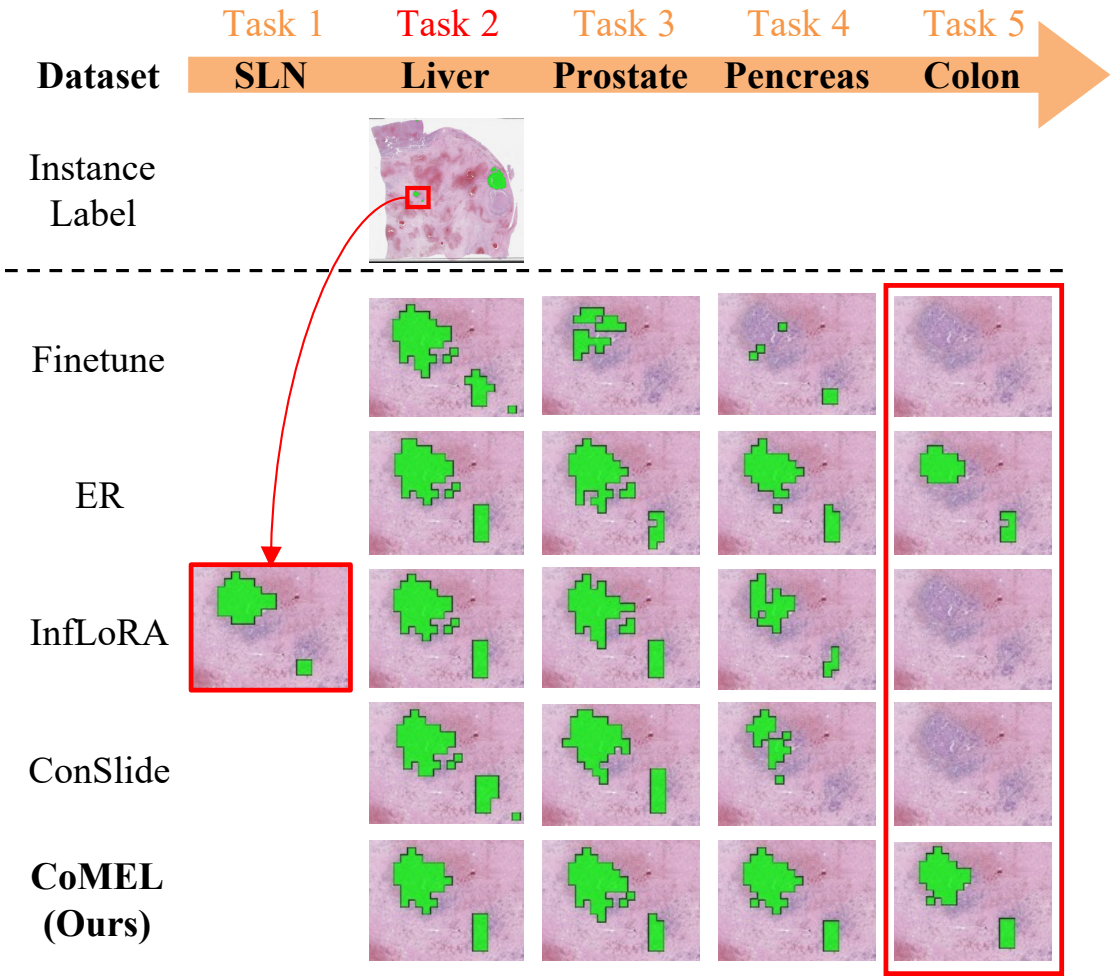


Figure S5. Additional qualitative results of localization across sequential organ datasets under the continual MIL setup. Each column is the localization performance on Task 2 as the learned organ changes over sequential tasks. Each row corresponds to CL methods including CoMEL. CoMEL successfully preserved the localization quality across all tasks, while baselines increase false positives or false negatives.

References

[S1] Babak Ehteshami Bejnordi, Mitko Veta, Paul Johannes Van Diest, Bram Van Ginneken, Nico Karssemeijer, Geert Litjens, Jeroen AWM Van Der Laak, Meyke Hermsen, Quirine F Manson, Maschenka Balkenhol, et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *Jama*, 318(22):2199–2210, 2017. 1

[S2] Francisco M Castro-Macías, Pablo Morales-Alvarez, Yunan Wu, Rafael Molina, and Aggelos Katsaggelos. Sm: enhanced localization in multiple instance learning for medical imaging classification. *NeurIPS*, 2024. 5

[S3] Richard J Chen, Chengkuan Chen, Yicong Li, Tiffany Y Chen, Andrew D Trister, Rahul G Krishnan, and Faisal Mahmood. Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. *CVPR*, 2022. 1

[S4] Richard J Chen, Tong Ding, Ming Y Lu, Drew FK Williamson, Guillaume Jaume, Andrew H Song, Bowen Chen, Andrew Zhang, Daniel Shao, Muhammad Shaban, et al. Towards a general-purpose foundation model for computational pathology. *Nature Medicine*, 30(3):850–862, 2024. 2

[S5] Jiaxiang Gou, Luping Ji, Pei Liu, and Mao Ye. Queryable prototype multiple instance learning with vision-language models for incremental whole slide image classification. *arXiv preprint arXiv:2410.10573*, 2024. 5

[S6] 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. *ICCV*, 2023. 1

[S7] Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based deep multiple instance learning. *ICML*, 2018. 2

[S8] Kyungmo Kim, Kyoungbun Lee, Sungduk Cho, Dong Un Kang, Seongkeun Park, Yunsook Kang, Hyunjeong Kim, Gheeyoung Choe, Kyung Chul Moon, Kyu Sang Lee, et al. Paip 2020: Microsatellite instability prediction in colorectal cancer. *Medical Image Analysis*, 89:102886, 2023. 1

[S9] Yoo Jung Kim, Hyungjoon Jang, Kyoungbun Lee, Seongkeun Park, Sung-Gyu Min, Choyeon Hong, Jeong Hwan Park, Kanggeun Lee, Jisoo Kim, Wonjae Hong, et al. Paip 2019: Liver cancer segmentation challenge. *Medical image analysis*, 67:101854, 2021. 1

[S10] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *ICLR*, 2015. 2

[S11] Hyunseo Koh, Dahyun Kim, Jung-Woo Ha, and Jonghyun Choi. Online continual learning on class incremental blurry task configuration with anytime inference. In *ICLR*, 2022. 6

[S12] Bin Li, Yin Li, and Kevin W Eliceiri. Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. *CVPR*, 2021. 2, 5

[S13] 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. 1, 2, 5

[S14] Zhuchen Shao, Hao Bian, Yang Chen, Yifeng Wang, Jian Zhang, Xiangyang Ji, et al. Transmil: Transformer based correlated multiple instance learning for whole slide image classification. *NeurIPS*, 2021. 5

[S15] Wenhao Tang, Sheng Huang, Xiaoxian Zhang, Fengtao Zhou, Yi Zhang, and Bo Liu. Multiple instance learning framework with masked hard instance mining for whole slide image classification. *ICCV*, 2023. 2

[S16] Wenhao Tang, Fengtao Zhou, Sheng Huang, Xiang Zhu, Yi Zhang, and Bo Liu. Feature re-embedding: Towards foundation model-level performance in computational pathology. *CVPR*, 2024. 5

[S17] John N Weinstein, Eric A Collisson, Gordon B Mills, Kenna R Shaw, Brad A Ozenberger, Kyle Ellrott, Ilya Shmulevich, Chris Sander, and Joshua M Stuart. The cancer genome atlas pan-cancer analysis project. *Nature genetics*, 45(10):1113–1120, 2013. 1

[S18] WISEPAIP. Wisepaip: Whole-slide image standard evaluation for pathology ai platform. http://www.wisepaip.org/. 1

[S19] Hongrun Zhang, Yanda Meng, Yitian Zhao, Yihong Qiao, Xiaoyun Yang, Sarah E Coupland, and Yalin Zheng. Dtfd-mil: Double-tier feature distillation multiple instance learning for histopathology whole slide image classification. *CVPR*, 2022. 2