

# Advancing Multiple Instance Learning with Continual Learning for Whole Slide Imaging

## Supplementary Material

### 7. Parameter Update in MIL

In this section, we derive (5) and (6). The bag-level feature vector is:

$$f_\phi = \phi^\top \mathbf{H} \mathbf{a} = \phi^\top \mathbf{z} = \sum_j \phi_j z_j, \quad (11)$$

and in the binary classification setting, the model's loss function is

$$L = \log(1 + e^{-yf_\phi}). \quad (12)$$

To obtain (5),

$$\frac{\partial L}{\partial \phi_j} = \frac{\partial L}{\partial f_\phi} \cdot \frac{\partial f_\phi}{\partial \phi_j} \quad (13)$$

$$= \left( \frac{1}{1 + e^{-yf_\phi}} e^{-yf_\phi} (-y) \right) (z_j) \quad (14)$$

$$= \sigma(-yf_\phi) (-yz_j). \quad (15)$$

Thus,

$$\left( \frac{\partial L}{\partial \phi_j} \right)^2 = \sigma^2(-yf_\phi) y^2 z_j^2. \quad (16)$$

For (6), we have

$$\frac{\partial L}{\partial a_i} = \frac{\partial L}{\partial f_\phi} \cdot \frac{\partial f_\phi}{\partial \mathbf{z}^\top} \cdot \frac{\partial \mathbf{z}}{\partial a_i} \quad (17)$$

$$= \left( \frac{1}{1 + e^{-yf_\phi}} e^{-yf_\phi} (-y) \right) (\phi^\top) (\mathbf{h}_i) \quad (18)$$

$$= \sigma(-yf_\phi) (-y\phi^\top \mathbf{h}_i). \quad (19)$$

Thus,

$$\left( \frac{\partial L}{\partial a_i} \right)^2 = \sigma^2(-yf_\phi) y^2 (\phi^\top \mathbf{h}_i)^2. \quad (20)$$

### 8. Experiment Details

We present the experimental details in this section.

#### 8.1. Dataset

The skin cancer dataset comprises WSIs representing six distinct types of cutaneous soft tissue neoplasms: leiomyoma, leiomyosarcoma, dermatofibroma, dermatofibrosarcoma, spindle-cell melanoma, and fibroxanthoma. While the original, full-resolution WSIs are not publicly accessible, a curated subset is available, consisting of 600 vectorized and labeled pathology images. The Camelyon16 dataset consists of 399 slides labeled as either normal or

tumor tissue. The TCGA-LUNG dataset provides data on two distinct types of lung cancer: LUAD with 534 slides, and LUSC with 512 slides. Finally, TCGA-RCC contains 940 renal cell carcinoma samples, divided among three subtypes: 121 from TCGA-KICH, 519 from TCGA-KIRC, and 300 from TCGA-KIRP.

#### 8.2. Data pre-processing

For the skin cancer dataset, we used their repo <https://github.com/cvblab/MICIL> to get the data. For the Camelyon-TCGA dataset, we followed [15], using the automated segmentation pipeline to get the tissue regions and crop  $256 \times 256$  patches at 20X magnification for each slide. We use the preset segmentation parameters for segmenting biopsy slides scanned at BWH for the Camelyon dataset and the parameters for TCGA slides for TCGA-LUNG and TCGA-RCC. The ResNet-50 model pre-trained with ImageNet is the fixed feature extractor that uses a global average pooling instead of the last convolutional module and converts each patch into a 1024-dimensional feature vector. Each task's dataset was split into 5 folds, with 4 folds used for training and the remaining fold reserved for testing. The training data is further randomly split into training and validation sets with a ratio of 4:1.

#### 8.3. Evaluation Metrics

Denote  $a_{t,j}$  as the accuracy on task  $j$  after CL training session  $t$ .

$$\text{AACC} = \frac{1}{T} \sum_{j=1}^T a_{T,j}, \quad (21)$$

where  $a_{T,j}$  is the test accuracy on task  $j$  after training on all  $T$  tasks.

**BWT** measures how well the model retains knowledge from prior tasks as it learns new ones, with  $a_{T,j}$  indicating the accuracy on task  $j$  after training on the final task  $T$ ,

$$\text{BWT} = \frac{1}{T-1} \sum_{j=1}^{T-1} (a_{T,j} - a_{j,j}). \quad (22)$$

**IM** quantifies a model's inability to learn a new task effectively compared to an ideal scenario,

$$\text{IM} = \frac{1}{T} \sum_{j=1}^T (a_j^* - a_{j,j}), \quad (23)$$

where  $a_j^*$  denotes the joint training accuracy on task  $j$ .

#### 8.4. Training

For the skin cancer, we followed [5] and used their experiment settings for TransMIL. We changed the learning rate to

1e-4 and epoch to 50 for CLAM. For Camelyon-TCGA, following [15, 20], we use the Adam optimizer with a weight decay of 1e-5 and a learning rate of 2e-4. All the models are trained for 50 epochs with an early stop strategy. Training was conducted on an Nvidia RTX3090.

## 8.5. Results

More detailed experimental results are shown in this section. Table 6 shows the complete results including means and standard deviations. Figure 4 illustrates the actual forgetting process of the model on Camelyon-TCGA with a memory setting of 10 WSIs. Our method outperforms all other methods on both  $t = 2$  and  $t = 3$  by a wide margin, exhibiting significantly less degradation.

Figure 5 provides a comparative visualization of the tradeoff between two key metrics in continual learning: BWT and IM. BWT measures forgetting of previous tasks, with higher BWT indicating less forgetting. IM measures the model’s ability to learn new tasks, with lower values indicating better ability. For rehearsal-based methods, we only visualize results with a memory pool of 10 WSIs. Thus, proximity to the upper-left corner indicates strong resistance to forgetting and adaptability to new tasks, i.e., models closer to this ideal area maintain knowledge from previous tasks (high BWT) while acquiring new information efficiently (low IM). As our model is closest to the upper-left corner, our model achieves the best BWT-IM tradeoff with both MIL models.

## 8.6. Compare with ConSlide

ConSlide only released part of its code, and the WSI dataset preprocessing and partitioning steps are not released. Despite this, we implement the benchmark in their paper with the available information, and our results are presented in Tab. 7 – our method significantly outperforms in accuracy.

## 8.7. Visualization

Here we show more samples of visualization in Figure 6. The fine-tuning method exhibits limited effectiveness in maintaining the attention distribution – attention is maintained on only a few samples in stage  $t=2$ , while in stage  $t=3$ , attention in the tumor region significantly declined across all samples. In contrast, our method consistently preserved the desired attention distribution, even at stage  $t=3$ . This demonstrates that our approach is more effective in maintaining meaningful attention to critical regions that are discriminative for classification.

Table 6. CL performance of CLAM and TransMIL on Camelyon-TCGA dataset. The best performances are highlighted as bold.

CL Type	Method	Memory Size	CLAM			TransMIL		
			AACC $\uparrow$	BWT $\uparrow$	IM $\downarrow$	AACC $\uparrow$	BWT $\uparrow$	IM $\downarrow$
Baselines	Joint training	-	0.858 $\pm$ 0.016	-	-	0.818 $\pm$ 0.027	-	-
	Fine-tuning	-	0.296 $\pm$ 0.014	-0.865 $\pm$ 0.013	-0.014 $\pm$ 0.019	0.290 $\pm$ 0.002	-0.751 $\pm$ 0.044	0.024 $\pm$ 0.036
Regularization	LwF	-	0.295 $\pm$ 0.013	-0.865 $\pm$ 0.006	-0.013 $\pm$ 0.015	0.296 $\pm$ 0.010	-0.747 $\pm$ 0.062	0.021 $\pm$ 0.045
	MICIL	-	0.295 $\pm$ 0.008	-0.866 $\pm$ 0.005	-0.013 $\pm$ 0.010	0.294 $\pm$ 0.011	-0.760 $\pm$ 0.056	0.014 $\pm$ 0.031
Rehearsal	ER	5 WSIs	0.294 $\pm$ 0.010	-0.864 $\pm$ 0.015	-0.012 $\pm$ 0.017	0.296 $\pm$ 0.008	-0.748 $\pm$ 0.053	0.020 $\pm$ 0.030
	DER++		0.301 $\pm$ 0.010	-0.857 $\pm$ 0.018	-0.013 $\pm$ 0.018	0.288 $\pm$ 0.021	-0.752 $\pm$ 0.050	0.025 $\pm$ 0.028
	MICIL w/ ER		0.289 $\pm$ 0.013	-0.869 $\pm$ 0.005	-0.010 $\pm$ 0.013	0.290 $\pm$ 0.009	-0.759 $\pm$ 0.047	0.018 $\pm$ 0.036
	Ours		<b>0.657<math>\pm</math>0.032</b>	-0.329 $\pm$ 0.051	-0.017 $\pm$ 0.010	<b>0.374<math>\pm</math>0.060</b>	-0.635 $\pm$ 0.103	0.018 $\pm$ 0.022
	ER	10 WSIs	0.352 $\pm$ 0.063	-0.785 $\pm$ 0.082	-0.017 $\pm$ 0.013	0.297 $\pm$ 0.012	-0.746 $\pm$ 0.062	0.021 $\pm$ 0.045
	DER++		0.372 $\pm$ 0.070	-0.760 $\pm$ 0.097	-0.020 $\pm$ 0.021	0.299 $\pm$ 0.015	-0.738 $\pm$ 0.049	0.024 $\pm$ 0.038
	MICIL w/ ER		0.298 $\pm$ 0.006	-0.863 $\pm$ 0.005	-0.015 $\pm$ 0.010	0.302 $\pm$ 0.016	-0.746 $\pm$ 0.049	0.016 $\pm$ 0.042
	Ours		<b>0.729<math>\pm</math>0.041</b>	-0.217 $\pm$ 0.048	-0.059 $\pm$ 0.036	<b>0.394<math>\pm</math>0.085</b>	-0.595 $\pm$ 0.146	0.024 $\pm$ 0.030
	ER	30 WSIs	0.494 $\pm$ 0.058	-0.565 $\pm$ 0.081	-0.011 $\pm$ 0.016	0.308 $\pm$ 0.011	-0.728 $\pm$ 0.057	0.021 $\pm$ 0.043
	DER++		0.449 $\pm$ 0.048	-0.645 $\pm$ 0.070	-0.020 $\pm$ 0.016	0.315 $\pm$ 0.028	-0.739 $\pm$ 0.071	0.006 $\pm$ 0.044
	MICIL w/ ER		0.308 $\pm$ 0.021	-0.835 $\pm$ 0.015	-0.006 $\pm$ 0.009	0.298 $\pm$ 0.019	-0.739 $\pm$ 0.052	0.024 $\pm$ 0.037
	Ours		<b>0.754<math>\pm</math>0.029</b>	-0.177 $\pm$ 0.041	-0.058 $\pm$ 0.035	<b>0.489<math>\pm</math>0.059</b>	-0.460 $\pm$ 0.059	0.018 $\pm$ 0.042

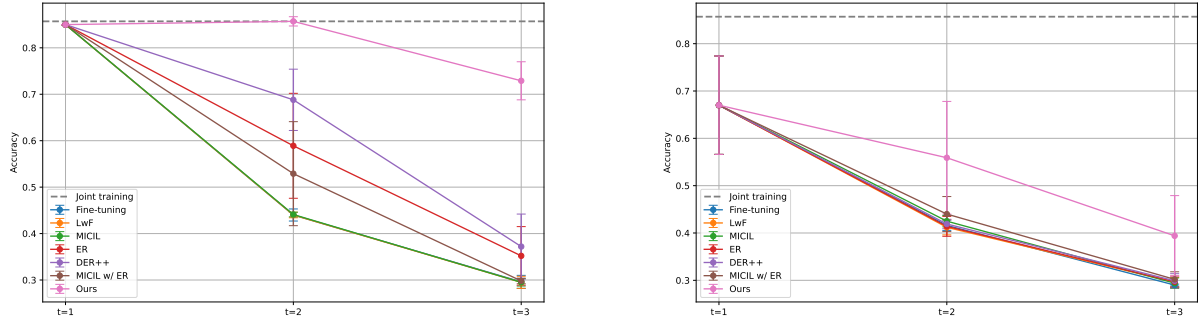


Figure 4. AACC performance of (left) CLAM and (right) TransMIL on Camelyon-TCGA as the number of tasks increases, i.e., the CL session  $t$  increases. The gray dotted line indicates the AACC performance of joint training.

Method	Buffer Size	ACC $\uparrow$	BWT $\uparrow$
ConSlide [8]	1100 regions $\approx$ 5 WSIs	0.553 $\pm$ 0.033	-0.066 $\pm$ 0.023
Ours	5 WSIs	0.763 $\pm$ 0.011	-0.222 $\pm$ 0.034
ConSlide [8]	2200 regions $\approx$ 10 WSIs	0.594 $\pm$ 0.053	-0.092 $\pm$ 0.026
Ours	10 WSIs	0.803 $\pm$ 0.030	-0.171 $\pm$ 0.035
ConSlide [8]	6600 regions $\approx$ 30 WSIs	0.659 $\pm$ 0.022	-0.075 $\pm$ 0.030
Ours	30 WSIs	0.868 $\pm$ 0.028	-0.071 $\pm$ 0.026

Table 7. CL performance of CLAM on TCGA dataset (NSCLC  $\rightarrow$  BRCA  $\rightarrow$  RCC  $\rightarrow$  ESCA).

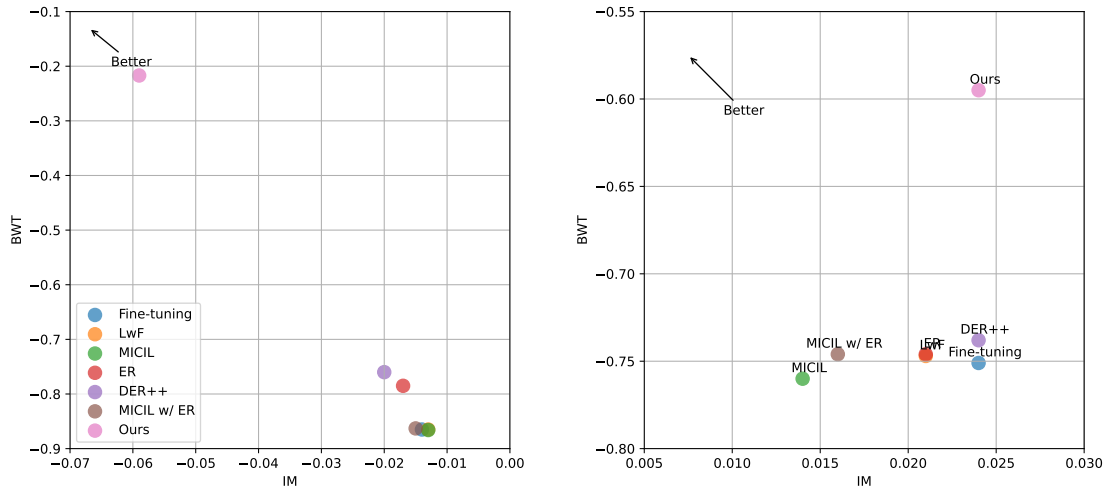


Figure 5. The tradeoff between BWT and IM on Camelyon-TCGA for (left) CLAM and (right) TransMIL. BWT measures the amount of forgetting of previous tasks, with higher BWT indicating less forgetting. IM measures the ability to learn new tasks, with lower IM indicating better ability. Thus, better methods are closer to the upper-left corner. The points represent different methods, with our method achieving the best trade-off with both models.

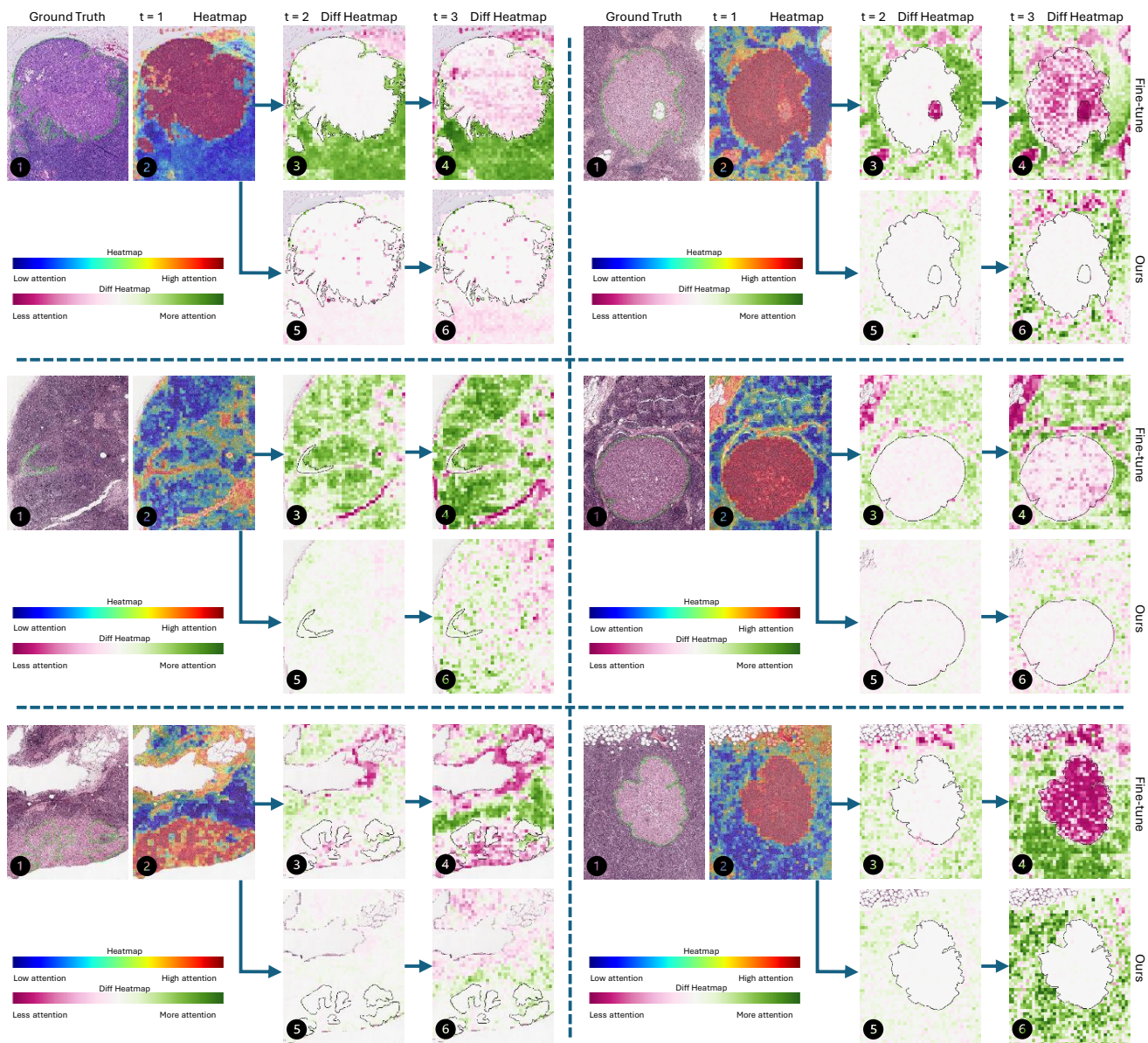


Figure 6. Additional visualizations illustrating the attention distributions across different CL sessions ( $t = 1, 2, 3$ ), using the same format as Figure 1.