MergeSlide: Continual Model Merging and Task-to-Class Prompt-Aligned Inference for Lifelong Learning on Whole Slide Images
Supplementary Material

Doanh C. Bui[1,*], Ba Hung Ngo[2], Hoai Luan Pham[1], Khang Nguyen[3], Maï K. Nguyen[4], Yasuhiko Nakashima[1]

[1]Nara Institute of Science and Technology, Japan
[2]Graduate School of Data Science, Chonnam National University, South Korea
[3]University of Information Technology, Viet Nam National University Ho Chi Minh City
[4]ETIS (UMR 8051), CY Cergy Paris University, ENSEA, CNRS, France
bui.cao_doanh.bd2@naist.ac.jp

Contents

A. Overview

In the supplementary material, we provide pseudocode for MergeSlide (Sec. B), implementation details of the backbone $f_{\mathcal{A}}$ (Sec. C), the full set of designed class-aware prompts per task (Sec. D), clarification of the class-incremental and task-incremental scenarios (Sec. E), definitions of evaluation metrics (Sec. F), and a discussion on comparison with the zero-shot approach (Sec. G). We further report additional metrics, including per-class Precision, Recall, AUC, and Macro/Weighted F1 (Sec. H). Additional comparisons are also included (Sec. I), such as task-wise performance drops, confidence score analysis, an ablation study on task-specific MLP-free fine-tuning, and performance after training on the final task.

B. Pseudocode

To provide a clearer description of MergeSlide, we present Algorithm 1, which comprehensively summarizes its processes and operations when training on a stream of T datasets.

C. Details of Slide Aggregators

In this section, we provide details of the slide aggregator $f_{\mathcal{A}}$ used in this study. For other rehearsal-based methods such as ER-ACE [3], AGEM [4], DER++ [2], ADaFGrad [1], and ConSlide [7], $f_{\mathcal{A}}$ is implemented using the HIT backbone introduced in the ConSlide study [7] to ensure a fair comparison. HIT is a hierarchical transformer designed to leverage pyramid multi-level images of a WSI. This network includes $L = 2$ Transformer blocks with a hidden dimension size of $d_{\text{model}} = 384$. Each Transformer block contains two self-attention layers for patch-level and region-level embeddings, respectively, with a cross-interaction mechanism to

Algorithm 1 MergeSlide

Input: Sequence of T training datasets $\mathcal{D}^{train} = \{D_t^{train}\}_{t=1}^{T}$, each with c_t cancer subtypes. Testing datasets: $\mathcal{D}^{test} = \{D_t^{test}\}_{t=1}^{T}$.
Init: Pre-trained slide aggregator $f_{\mathcal{A}}(\cdot, \theta_{base})$
Initialize controlling factor $\lambda_1 \leftarrow 1$
$\forall X_i \in \{D_t^{train}, D_t^{test}\}_{t=1}^{T}$, perform tissue segmentation, tiling, patch embedding, and sampling to obtain B_i'.
Fine-tuning and model merging:
for $t = 1$ to T **do**
  Fine-tuning for task t: ▷ Sec. 3.2.1
  Define class prompts $P^{\mathcal{C},t}$ and embeddings $E^{\mathcal{C},t}$
  Initialize weights: $\theta_t \leftarrow \theta_{base}$
  Train $f_{\mathcal{A}}$ on $B_i' \in D_t^{train}$:
    $Z_i \leftarrow f_{\mathcal{A}}(B_i', \theta_t)$
    $\hat{p}_i \leftarrow \{Z_i \cdot (e_j^{\mathcal{C},t})^\top \mid e_j^{\mathcal{C},t} \in E^{\mathcal{C},t}\}$
  Update: $\theta_t \leftarrow \arg\min_{\theta_t} \mathcal{L}_t(y_i, \hat{p}_i)$
  Continual model merging: ▷ Sec. 3.2.2
  if $t > 1$ **then**
    $\Delta\tilde{\theta}_{1:t-1} \leftarrow \tilde{\theta}_{1:t-1} - \theta_{base}$
    $\Delta\theta_t \leftarrow \theta_t - \theta_{base}$
    SVD: $\Delta\tilde{\theta}_{1:t-1} \leftarrow U\Sigma V^\top$
    Project: $G(\Delta\theta_t) \leftarrow U\big((U^\top \Delta\theta_t V) \odot M\big) V^\top$ ▷ Eq. (4)
    Compute $\lambda_t \leftarrow t \cdot \frac{\|\lambda_{t-1}\Delta\tilde{\theta}_{1:t-1} + G(\Delta\theta_t)\|_2}{\sum_{i=1}^{t}\|\theta_i\|_2}$ ▷ Eq. (5)
    Merge: $\tilde{\theta}_{1:t} \leftarrow \theta_{base} + \frac{\lambda_{t-1}\Delta\tilde{\theta}_{1:t-1} + G(\Delta\theta_t)}{\lambda_t}$ ▷ Eq. (6)
  else
    $\tilde{\theta}_{1:t} \leftarrow \theta_t$
  end if
end for
Inference on all test sets:
Compute task-level prompts: $E^{\mathcal{T},t} \leftarrow \frac{1}{c_t}\sum E^{\mathcal{C},t}$ for all t
$\mathcal{E}^{\mathcal{T}} \leftarrow \{E^{\mathcal{T},t}\}_{t=1}^{T}$
for $t = 1$ to T **do**
  for B_i' in D_t^{test} **do**
    $Z_i \leftarrow f_{\mathcal{A}}(B_i', \tilde{\theta}_{1:T})$
    CLASS-IL (naive):
    $\hat{p}_i \leftarrow \{\}$
    for $t = 1$ to T **do**
      $\hat{p}_i \leftarrow \hat{p}_i \cup \{Z_i \cdot (e_j^{\mathcal{C},t})^\top \mid e_j^{\mathcal{C},t} \in E^{\mathcal{C},t}\}$
    end for
    Predict: $\hat{y}_i \leftarrow \arg\max \hat{p}_i$
    CLASS-IL with **prompt-aligned inference**: ▷ Sec. 3.2.3
    $\hat{t} \leftarrow \arg\max_t \{Z_i \cdot (E^{\mathcal{T},t})^\top\}$
    Predict: $\hat{y}_i \leftarrow \arg\max_j \{Z_i \cdot (e_j^{\mathcal{C},\hat{t}})^\top\}$
    Task-IL:
    Predict: $\hat{y}_i \leftarrow \arg\max_j \{Z_i \cdot (e_j^{\mathcal{C},t})^\top\}$
  end for
end for
Return: Final unified model weights $\tilde{\theta}_{1:T}$

fuse these two levels. Specifically, region embeddings are added to each patch embedding. All patch embeddings are then passed through a convolutional layer and max-pooled before being added back to the region embedding. A region

comprises multiple corresponding patches. In this study, we use region images (of size 1024×1024) at $2.5\times$ magnification and patch images (of size 256×256) at $10\times$, meaning each region contains 4×4 patch images. For the Zero-shot setting and MergeSlide, since $f_{\mathcal{A}}$ should be pre-trained to contain strong base knowledge θ_{base}, we adopt the pre-trained Transformer-based network introduced in TITAN [6], which is a variant of the Vision Transformer that with $d_{\text{model}} = 768$. However, in MergeSlide, we sample only $K = 400$ patches per slide, whereas other methods utilize all patches to achieve the best performance.

D. How Class-aware Prompts Are Designed?

The class-aware prompts are designed based on the following format: class template + `[CLASSNAME]`, where the `[CLASSNAME]` tag is a short natural-language description of a cancer subtype. We leveraged 22 class template variants, as shown in Table D.1. For each cancer subtype, there are 3∼4 class tags, listed in Table D.2. For each i-th cancer subtype, variants of class-aware prompts is constructed by concatenating all 22 class templates with the 3∼4 class tags, resulting in approximately 88 class-aware prompts (denoted as $p_i^{\mathcal{C},t}$) for each cancer subtype. These prompts are then passed through the text encoder of TITAN to obtain 88 embedding vectors, which are averaged across dimensions to produce a class-aware embedding vector $e_i^{\mathcal{C},t} \in \mathbb{R}^{d_{text}}$. For each t-th dataset/task, there are c_t cancer subtypes (e.g., for TCGA-BRCA, there are two cancer subtypes: invasive ductal carcinoma and invasive lobular carcinoma). A set of class-aware prompts is then constructed as $E^{\mathcal{C},t} = \{e_i^{\mathcal{C},t}\}_{i=1}^{c_t}$, which is used to retrieve the final prediction by computing the dot-product similarity with the slide embedding.

#	Prompt template
❶	"`[CLASSNAME]`."
❷	"a photomicrograph showing `[CLASSNAME]`."
❸	"a photomicrograph of `[CLASSNAME]`."
❹	"an image of `[CLASSNAME]`."
❺	"an image showing `[CLASSNAME]`."
❻	"an example of `[CLASSNAME]`."
❼	"`[CLASSNAME]` is shown."
❽	"this is `[CLASSNAME]`."
❾	"there is `[CLASSNAME]`."
❿	"a histopathological image showing `[CLASSNAME]`."
⓫	"a histopathological image of `[CLASSNAME]`."
⓬	"a histopathological photograph of `[CLASSNAME]`."
⓭	"a histopathological photograph showing `[CLASSNAME]`."
⓮	"shows `[CLASSNAME]`."
⓯	"presence of `[CLASSNAME]`."
⓰	"`[CLASSNAME]` is present."
⓱	"an H&E stained image of `[CLASSNAME]`."
⓲	"an H&E stained image showing `[CLASSNAME]`."
⓳	"an H&E image showing `[CLASSNAME]`."
⓴	"an H&E image of `[CLASSNAME]`."
㉑	"`[CLASSNAME]`, H&E stain."
㉒	"`[CLASSNAME]`, H&E."

Table D.1. Prompt template variations.

#	Cancer Subtype	`[CLASSNAME]`
TCGA-BRCA	invasive ductal carcinoma	❶ "invasive ductal carcinoma" ❷ "breast invasive ductal carcinoma" ❸ "invasive ductal carcinoma of the breast" ❹ "invasive carcinoma of the breast, ductal pattern"
TCGA-BRCA	invasive lobular carcinoma	❶ 'invasive lobular carcinoma" ❷ "breast invasive lobular carcinoma" ❸ "invasive lobular carcinoma of the breast" ❹ "invasive carcinoma of the breast, lobular pattern"
TCGA-RCC	clear cell	❶ "clear cell renal cell carcinoma" ❷ "renal cell carcinoma, clear cell type" ❸ "renal cell carcinoma of the clear cell type" ❹ "clear cell rcc"
TCGA-RCC	papillary	❶ "papillary renal cell carcinoma" ❷ "renal cell carcinoma, papillary type" ❸ "renal cell carcinoma of the papillary type" ❹ "papillary rcc"
TCGA-RCC	chromophobe renal cell carcinoma	❶ "chromophobe renal cell carcinoma" ❷ "renal cell carcinoma, chromophobe type" ❸ "renal cell carcinoma of the chromophobe type" ❹ "chromophobe rcc"
TCGA-NSCLC	squamous cell carcinoma	❶ "squamous cell carcinoma" ❷ "lung squamous cell carcinoma" ❸ "squamous cell carcinoma of the lung" ❹ "lusc"
TCGA-NSCLC	adenocarcinoma	❶ "adenocarcinoma" ❷ "lung adenocarcinoma" ❸ "adenocarcinoma of the lung" ❹ "luad"
TCGA-ESCA	squamous cell carcinoma	❶ "squamous cell carcinoma" ❷ "esophageal squamous cell carcinoma" ❸ "squamous cell carcinoma of the esophagus" ❹ "essc"
TCGA-ESCA	adenocarcinoma	❶ "adenocarcinoma" ❷ "esophageal adenocarcinoma" ❸ "adenocarcinoma of the esophagus" ❹ "esad"
TCGA-TGCT	Sinoma	❶ "Sinoma" ❷ "testicular Sinoma" ❸ "Sinoma of the testis"
TCGA-TGCT	mixed germ cell tumor	❶ "mixed germ cell tumor" ❷ "testicular mixed germ cell tumor" ❸ "mixed germ cell tumor of the testis"
TCGA-CESC	squamous cell carcinoma	❶ "squamous cell carcinoma" ❷ "cervical squamous cell carcinoma" ❸ "squamous cell carcinoma of the cervix uteri"
TCGA-CESC	adenocarcinoma	❶ "adenocarcinoma" ❷ "cervical adenocarcinoma" ❸ "adenocarcinoma of the cervix uteri"

Table D.2. Class-aware prompts for each cancer subtype.

E. Explanations of CLASS-IL and TASK-IL

Under the CLASS-IL scenario, the model does not know the task identity (i.e., t is unknown) of an input WSI, whereas under the TASK-IL scenario, the model knows which task should be targeted (i.e., t is known).

E.1. Class-imcremental Learning (CLASS-IL)

Following the iCARL implementation [8], all continual learning methods take the argmax of the predicted logits $\hat{y} = \arg\max \hat{p}_i$, where $\hat{p} \in \mathbb{R}^C$ and C is the total number of classes seen so far from T tasks, i.e., $C = \sum_{t=1}^{T} c_t$, as the final prediction. For MergeSlide, in the naive setting, we compute all dot-product similarities between the input slide embedding and all class-aware prompt embeddings $\{E^{\mathcal{C},t}\}_{t=1}^{T}$ to obtain $\hat{p}_i$ (as shown in lines 31–34 of Algorithm 1). Under TCP inference, the model is first given the task identity $\hat{t}$ by selecting the task-level prompt embedding $\{E^{\mathcal{T},t}\}_{t=1}^{T}$ that yields the highest similarity. After obtain-

ing $\hat{t}$, the corresponding set of class-aware prompt embeddings $E^{\mathcal{C},\hat{t}}$ is used to compute the final prediction. *We note that this strategy is still valid under the CLASS-IL setting, where the model does not know the task identity and must consider all cancer subtype classes seen so far. The TCP inference strategy identifies the task using a vision-language similarity approach, thereby narrowing the task boundary. However, it does not assume access to the task identity at the beginning.*

E.2. Task-imcremental Learning (TASK-IL)

Under the TASK-IL scenario, the model is provided with the task identity, and therefore knows the task boundary. Specifically, given task t, the task boundary is defined as $s_t = \sum_{i=1}^{t-1} c_i, \quad e_t = \sum_{i=1}^{t} c_i$, where c_i is the number of classes of i-th task. Then, the prediction is computed by taking $\hat{y} = \arg\max_{j \in s_t,\ldots,e_t} \hat{p}_{i,j}$, where $\hat{p}_{i,j}$ is the j-th logit representing the j-th cancer subtype class after training on all tasks. For MergeSlide, as the model knows the task identity, we simply pick the corresponding set of class-aware prompts $E^{\mathcal{C},t}$ to retrieve the final prediction.

F. Explanations of Metrics

Alongside (1) **bACC** and (2) **Masked bACC**, we evaluate continual learning performance using three metrics: (3) **Mean Accuracy (mACC)**, (4) **Backward Transfer (BWT)**, (5) **Forgetting (FGT)**.

F.1. Balanced Accuracy (bACC)

Since each dataset in the six-TCGA benchmark is highly imbalanced, we report Balanced Accuracy as the main metric for both CLASS-IL and TASK-IL scenarios to better reflect performance on such test sets. bACC of the t-th task, which involves classifying c_t classes, is defined as:

$$bACC_t = \frac{1}{c_t} \sum_{i=1}^{c_t} \frac{TP_i}{TP_i + FN_i}, \tag{1}$$

where TP_i is the number of samples correctly classified as the i-th cancer subtype in the t-th task, and FN_i is the number of samples of the i-th cancer subtype misclassified as other subtypes. In other words, this corresponds to the macro-averaged recall across all cancer subtypes. The final bACC is computed as the mean bACC across all T tasks:

$$bACC = \frac{1}{T} \sum_{t=1}^{T} bACC_t. \tag{2}$$

F.2. Accuracy (ACC)

Accuracy is defined as the ratio of correctly classified samples to the total number of samples in a task, and it has been the primary evaluation metric in previous continual learning studies [1–4, 7]. To ensure fair comparison with these works, our study also adopts accuracy as the base metric for computing other continual learning measures.

Let T be the total number of tasks, and let $ACC_i(j)$ denote the accuracy on task j after training up to task i. The equations for **Mean ACC**, **BWT**, and **FGT** are provided in Secs. F.3, F.4, and F.5, respectively.

F.3. Mean Accuracy (Mean ACC)

Mean Accuracy is defined as the average of CLASS-IL accuracies over T sequences of incrementally learned tasks:

$$\text{mACC} = \frac{1}{T} \sum_{t=1}^{T} ACC_t, \tag{3}$$

where ACC_t denotes the accuracy under the CLASS-IL scenario on all tasks up to and including the t-th task.

F.4. Backward Transfer (BWT)

Backward Transfer measures how learning new tasks affects performance on old tasks:

$$\text{BWT} = \frac{1}{T-1} \sum_{t=1}^{T-1} \big(ACC_T(t) - ACC_t(t) \big). \tag{4}$$

F.5. Forgetting (FGT)

Forgetting measures the drop from each task's best achieved accuracy to its final accuracy after all training:

$$\text{FGT} = \frac{1}{T-1} \sum_{t=1}^{T-1} \Big(\max_{i \in \{t,\ldots,T\}} ACC_i(t) - ACC_T(t) \Big). \tag{5}$$

For ACC, Masked ACC, mACC, BWT, higher values indicate better performance, while for FGT, lower values are better. Specifically, BWT measures how tasks influence each other during the training process. Therefore, we expect the model to facilitate positive transfer between tasks; in other words, higher values are desirable. In contrast, FGT measures how much the model forgets old tasks after training on new ones, and this should be minimized. That is, lower values are better.

F.6. Precision, Recall, AUC and F1

To further investigate per-class performance across all models, we compute Precision, Recall, and AUC for each cancer subtype, as well as Weighted and Macro F1. For Precision, Recall, and AUC, we aggregate the predictions of all WSIs across T tasks and then compute the metrics to evaluate how well each cancer subtype can be recognized among all possible subtypes. Weighted and Macro F1 are also computed across all cancer subtypes in the sequence of T testing sets. The results are reported in Sec. H.

F.7. Davies-Bouldin Index (DBI)

In Qualitative Analyses provided in the main manuscript (Sec. 5), we also compute DBI to measure how good each slide embeddings clustered to its task/class spaces.

$$\text{DBI} = \frac{1}{K}\sum_{i=1}^{K}\max_{j\neq i}\left(\frac{S_i + S_j}{M_{ij}}\right), \tag{6}$$

where $S_i = \frac{1}{|C_i|}\sum_{Z\in C_i}\|Z-\mu_i\|^2$, where Z is the slide embedding of a WSI aggregated by $f_{\mathcal{A}}$, and C_i is the i-th cluster, i.e., the set of WSIs belonging to the same task or cancer subtyping class. μ_i is the centroid of cluster i, and $M_{ij} = \|\mu_i - \mu_j\|^2$ denotes the squared distance between the i-th and j-th centroids. Although t-SNE shows points in a projected 2D space, DBI is still computed on the original d-dimensional slide embedding.

G. Discussion For Zero-shot Approach

For zero-shot evaluation, we simply obtain the embedding from the pre-trained $f_{\mathcal{A}}$ without any task-specific training, and compute the dot-product similarity with all class-aware prompt embeddings in $\mathcal{E}^{\mathcal{C}}$. The resulting scores are then normalized using the softmax function, and the class with the highest score is selected as the final prediction. Let Z_i be the slide embedding produced by $f_{\mathcal{A}}$, and let the initial prediction vector be $\hat{p}_i = \{\varnothing\}$. This process can be expressed as:

$$\begin{aligned}\hat{p}_i &= \hat{p}_i \cup \{Z_i \cdot (e_j^{\mathcal{C},t})^\top \mid \forall e_j^{\mathcal{C},t} \in E^{\mathcal{C},t}\}, \quad \text{for } t = 1,\ldots,T\\ \hat{y}_i &= \arg\max\left(\text{softmax}(\hat{p}_i)\right)\end{aligned} \tag{7}$$

We note that a head-to-head comparison with the zero-shot setting is important, as MergeSlide relies on pathology VLMs. One might expect that zero-shot inference could achieve good performance without any task-specific training, leveraging vision-language alignment conditioned on well-crafted class-aware prompts. However, we show that simply relying on this approach does not outperform rehearsal-based continual learning methods, although it remains competitive. In contrast, our study proposes training an MLP-free slide aggregator $f_{\mathcal{A}}$, performing prediction using class-aware prompt embeddings over a few epochs, and then merging the results into a unified model. This approach significantly outperforms the zero-shot setting while maintaining a favorable trade-off between efficiency and effectiveness.

H. Extended Metrics

According to Tab. 1, the datasets are highly imbalanced. For instance, TCGA-BRCA and TCGA-NSCLC are dominated by invasive ductal carcinoma (IDC) and squamous cell carcinoma (SCC), respectively. Therefore, additional results of weighted and macro F1, as well as precision, recall, and AUC per class, should be investigated. We consider three evaluation scenarios: forward and reverse sequences of datasets in the **in-domain setting**, and the forward sequence in the **out-of-domain setting**. These results are reported in Tabs H.1, H.2, and H.3, respectively. All of these results are reported using a buffer size of 30 WSIs for all baseline continual learning methods. As a result, there are no consistent winners across methods in terms of class-wise precision, recall, and AUC. However, our MergeSlide consistently achieves the highest values among all methods. In particular, for rare classes, MergeSlide tends to outperform others in terms of precision and/or recall. For example, in Tab. H.1 (forward sequence, in-domain setting), considering the IDC of TCGA-BRCA, MergeSlide outperforms all other methods in both precision ($\geq$0.066) and AUC ($\geq$0.008). In the more challenging out-of-domain setting, where test slides are not acquired from the same sites as the training slides, this observation remains unchanged. MergeSlide maintains the highest precision and AUC for ILC, with values $\geq$0.034 and $\geq$0.013, respectively. A similar trend is observed for the adenocarcinoma class of TCGA-NSCLC, where MergeSlide outperforms all methods in precision, recall, and AUC, except for the precision in the out-of-domain setting. Moreover, for the ChRCC class of TCGA-RCC, MergeSlide consistently achieves the highest recall and AUC across all settings.

I. Additional Analyses

I.1. Time complexity

Regarding the formal time complexity of the orthogonal merging strategy, it consists of six steps (lines 15–20) in Algorithm 1 in the Supplementary: (1) computing the task vector $\Delta\tilde{\theta}_{1:t-1}$ between $\tilde{\theta}_{1:t-1}$ and θ_{base}, (2) computing the task vector $\Delta\theta_t$ between θ_t and θ_{base}, (3) performing SVD on $\tilde{\theta}_{1:t}$, (4) computing the projection $G(\Delta\theta_t)$, (5) computing λ_t, and (6) computing θ_t. Given each set of parameters $\theta \in \mathbb{R}^{m\times n}$, the time complexity of steps (1), (2), (5), and (6) is $\mathcal{O}(m\cdot n)$, as they involve only addition, subtraction, or scalar–matrix operations. (3) involves SVD decomposition, for which we use a QR-based algorithm (see Algorithm 1a in [5]). The time complexity, in the case $m > n$, is $\mathcal{O}(m\cdot n^2)$. The projection in (4) is dominated by matrix multiplications with U and $V^\intercal$, which also has a time complexity of $\mathcal{O}(m\cdot n^2)$. Obviously, these six steps are dominated by (3) and (4), then the time complexity for all steps is $\mathcal{O}(m\cdot n^2)$. Because the slide aggregator $f_{\mathcal{A}}$ includes L layers, each layer has set of paramters $\theta_l \in \mathbb{R}^{m_l\times n_l}$, the total cost for orthogonal merging over all L layers is $\mathcal{O}(\sum_{l=1}^{L} m_l n_l^2)$. For $|\mathcal{D}|$ datasets/tasks, the overall time complexity becomes $\mathcal{O}\left(|\mathcal{D}|\sum_{l=1}^{L} m_l n_l^2\right)$.

Metric \ Method	Subtype	MergeSlide (TCP)	MergeSlide (naive)	ADaFGrad	ConSlide	DER++	AGEM	ER-ACE	Joint	Naive Finetuning
Balanced ACC (%)		**87.929 (±2.110)**	80.668 (±1.860)	69.034 (±3.861)	64.622 (±1.755)	55.560 (±2.746)	42.076 (±4.287)	58.556 (±5.654)	81.309 (±3.064)	25.673 (±2.571)
Masked Balanced ACC (%)		**92.087 (±1.740)**	**92.087 (±1.740)**	89.704 (±2.082)	88.346 (±1.977)	86.863 (±2.810)	80.364 (±2.206)	75.255 (±5.161)	89.688 (±1.687)	86.597 (±1.933)
Weighted F1		**0.899 (±0.013)**	0.790 (±0.021)	0.725 (±0.029)	0.684 (±0.015)	0.629 (±0.015)	0.488 (±0.042)	0.641 (±0.043)	0.849 (±0.020)	0.272 (±0.027)
Macro F1		**0.860 (±0.019)**	0.793 (±0.020)	0.688 (±0.037)	0.644 (±0.016)	0.544 (±0.026)	0.398 (±0.039)	0.561 (±0.058)	0.827 (±0.024)	0.222 (±0.037)
Precision per class	TCGA-BRCA IDC	**0.947 (±0.033)**	0.944 (±0.035)	0.960 (±0.039)	0.969 (±0.031)	0.881 (±0.064)	0.703 (±0.236)	0.820 (±0.079)	0.918 (±0.041)	0.000 (±0.000)
	TCGA-BRCA ILC	0.730 (±0.133)	**0.745 (±0.102)**	0.684 (±0.153)	0.613 (±0.104)	0.672 (±0.291)	0.018 (±0.054)	0.401 (±0.418)	0.816 (±0.132)	0.633 (±0.425)
	TCGA-RCC CC	**0.997 (±0.008)**	0.968 (±0.027)	0.983 (±0.016)	0.918 (±0.054)	0.961 (±0.027)	0.933 (±0.074)	0.915 (±0.055)	0.973 (±0.018)	0.990 (±0.021)
	TCGA-RCC P	0.879 (±0.043)	0.895 (±0.042)	0.927 (±0.048)	0.926 (±0.056)	**0.936 (±0.040)**	0.925 (±0.126)	0.835 (±0.105)	0.913 (±0.044)	0.958 (±0.044)
	TCGA-RCC ChRCC	0.576 (±0.061)	0.809 (±0.081)	0.834 (±0.108)	0.973 (±0.054)	**0.936 (±0.088)**	0.779 (±0.227)	0.849 (±0.291)	0.921 (±0.066)	0.580 (±0.477)
	TCGA-NSCLC SCC	**0.957 (±0.029)**	0.952 (±0.038)	0.823 (±0.045)	0.805 (±0.076)	0.842 (±0.059)	0.827 (±0.055)	0.579 (±0.115)	0.868 (±0.053)	0.719 (±0.147)
	TCGA-NSCLC A	**0.879 (±0.029)**	0.848 (±0.034)	0.824 (±0.080)	0.764 (±0.063)	0.821 (±0.056)	0.514 (±0.197)	0.830 (±0.084)	0.904 (±0.060)	0.232 (±0.223)
	TCGA-ESCA SCC	**0.943 (±0.026)**	0.748 (±0.063)	0.805 (±0.281)	0.761 (±0.281)	0.400 (±0.490)	0.597 (±0.421)	0.529 (±0.234)	0.803 (±0.066)	0.367 (±0.458)
	TCGA-ESCA A	**0.907 (±0.037)**	0.410 (±0.035)	0.477 (±0.414)	0.280 (±0.368)	0.000 (±0.000)	0.000 (±0.000)	0.427 (±0.312)	0.707 (±0.118)	0.000 (±0.000)
	TCGA-TGCT S	0.905 (±0.033)	0.852 (±0.058)	0.890 (±0.074)	**0.909 (±0.085)**	0.000 (±0.000)	0.386 (±0.403)	0.545 (±0.197)	0.871 (±0.064)	0.030 (±0.090)
	TCGA-TGCT MGCT	**0.761 (±0.081)**	0.749 (±0.147)	0.623 (±0.346)	0.721 (±0.324)	0.100 (±0.300)	0.292 (±0.375)	0.307 (±0.415)	0.843 (±0.153)	0.000 (±0.000)
	TCGA-CESC SCC	**0.649 (±0.049)**	0.594 (±0.075)	0.458 (±0.066)	0.414 (±0.063)	0.338 (±0.064)	0.308 (±0.063)	0.026 (±0.077)	0.777 (±0.092)	0.133 (±0.040)
	TCGA-CESC A	**0.959 (±0.011)**	0.824 (±0.039)	0.634 (±0.040)	0.598 (±0.019)	0.555 (±0.025)	0.453 (±0.037)	0.697 (±0.080)	0.821 (±0.051)	0.372 (±0.026)
Recall per class	TCGA-BRCA IDC	**0.938 (±0.029)**	0.918 (±0.027)	0.787 (±0.052)	0.677 (±0.051)	0.875 (±0.063)	0.786 (±0.264)	0.879 (±0.076)	0.934 (±0.044)	0.000 (±0.000)
	TCGA-BRCA ILC	0.746 (±0.183)	**0.875 (±0.133)**	0.882 (±0.131)	0.917 (±0.087)	0.489 (±0.301)	0.100 (±0.300)	0.318 (±0.366)	0.775 (±0.191)	0.176 (±0.188)
	TCGA-RCC CC	0.821 (±0.044)	0.902 (±0.035)	0.913 (±0.047)	0.897 (±0.051)	**0.929 (±0.032)**	0.647 (±0.256)	0.925 (±0.034)	0.951 (±0.024)	0.301 (±0.156)
	TCGA-RCC P	0.940 (±0.038)	**0.974 (±0.024)**	0.910 (±0.044)	0.906 (±0.080)	0.875 (±0.089)	0.451 (±0.194)	0.849 (±0.080)	0.955 (±0.055)	0.341 (±0.166)
	TCGA-RCC ChRCC	**1.000 (±0.000)**	0.963 (±0.046)	0.963 (±0.084)	0.645 (±0.258)	0.870 (±0.117)	0.681 (±0.293)	0.625 (±0.304)	0.935 (±0.059)	0.165 (±0.177)
	TCGA-NSCLC SCC	0.861 (±0.041)	0.878 (±0.052)	0.889 (±0.109)	0.911 (±0.033)	0.877 (±0.051)	0.761 (±0.122)	**0.976 (±0.024)**	0.922 (±0.044)	0.506 (±0.216)
	TCGA-NSCLC A	**0.960 (±0.026)**	0.911 (±0.044)	0.853 (±0.055)	0.849 (±0.072)	0.881 (±0.081)	0.297 (±0.244)	0.687 (±0.149)	0.871 (±0.056)	0.050 (±0.053)
	TCGA-ESCA SCC	**0.869 (±0.055)**	0.859 (±0.047)	0.193 (±0.091)	0.176 (±0.122)	0.017 (±0.023)	0.052 (±0.047)	0.500 (±0.332)	0.752 (±0.100)	0.031 (±0.042)
	TCGA-ESCA A	**0.961 (±0.018)**	0.787 (±0.049)	0.050 (±0.059)	0.018 (±0.031)	0.000 (±0.000)	0.000 (±0.000)	0.200 (±0.244)	0.532 (±0.163)	0.000 (±0.000)
	TCGA-TGCT S	0.891 (±0.043)	0.887 (±0.074)	0.729 (±0.148)	0.516 (±0.157)	0.000 (±0.000)	0.056 (±0.071)	**0.902 (±0.083)**	0.940 (±0.065)	0.011 (±0.033)
	TCGA-TGCT MGCT	**0.784 (±0.075)**	0.585 (±0.142)	0.327 (±0.252)	0.313 (±0.172)	0.009 (±0.027)	0.062 (±0.086)	0.190 (±0.285)	0.587 (±0.224)	0.000 (±0.000)
	TCGA-CESC SCC	**0.790 (±0.054)**	0.575 (±0.087)	0.755 (±0.088)	0.785 (±0.092)	0.770 (±0.078)	0.785 (±0.098)	0.090 (±0.270)	0.660 (±0.109)	0.800 (±0.063)
	TCGA-CESC A	0.913 (±0.020)	0.514 (±0.080)	0.966 (±0.021)	0.962 (±0.016)	**0.966 (±0.016)**	0.966 (±0.017)	0.685 (±0.241)	0.891 (±0.042)	0.968 (±0.018)
AUC per class	TCGA-BRCA IDC	0.940 (±0.032)	**0.951 (±0.035)**	0.908 (±0.052)	0.912 (±0.047)	0.858 (±0.049)	0.627 (±0.140)	0.789 (±0.103)	0.935 (±0.044)	0.393 (±0.134)
	TCGA-BRCA ILC	0.945 (±0.030)	**0.965 (±0.032)**	0.948 (±0.044)	0.957 (±0.029)	0.944 (±0.032)	0.757 (±0.100)	0.895 (±0.074)	0.957 (±0.037)	0.915 (±0.057)
	TCGA-RCC CC	0.974 (±0.009)	**0.988 (±0.008)**	0.984 (±0.009)	0.962 (±0.019)	0.980 (±0.015)	0.926 (±0.028)	0.971 (±0.015)	0.991 (±0.007)	0.865 (±0.083)
	TCGA-RCC P	0.983 (±0.011)	**0.990 (±0.007)**	0.987 (±0.009)	0.981 (±0.015)	0.981 (±0.012)	0.949 (±0.049)	0.968 (±0.022)	0.989 (±0.009)	0.938 (±0.062)
	TCGA-RCC ChRCC	0.979 (±0.010)	0.995 (±0.003)	0.994 (±0.005)	**0.996 (±0.004)**	0.995 (±0.004)	0.977 (±0.015)	0.992 (±0.008)	0.995 (±0.004)	0.932 (±0.072)
	TCGA-NSCLC SCC	**0.975 (±0.015)**	0.965 (±0.026)	0.969 (±0.016)	0.972 (±0.017)	0.969 (±0.013)	0.952 (±0.022)	0.961 (±0.019)	0.972 (±0.015)	0.935 (±0.043)
	TCGA-NSCLC A	**0.975 (±0.015)**	0.952 (±0.023)	0.954 (±0.031)	0.948 (±0.024)	0.957 (±0.021)	0.715 (±0.196)	0.966 (±0.013)	0.974 (±0.010)	0.454 (±0.190)
	TCGA-ESCA SCC	**0.977 (±0.012)**	0.971 (±0.012)	0.841 (±0.076)	0.575 (±0.149)	0.576 (±0.090)	0.701 (±0.110)	0.838 (±0.096)	0.886 (±0.046)	0.705 (±0.049)
	TCGA-ESCA A	**0.977 (±0.012)**	0.967 (±0.012)	0.914 (±0.022)	0.904 (±0.036)	0.924 (±0.020)	0.927 (±0.022)	0.918 (±0.034)	0.941 (±0.020)	0.918 (±0.031)
	TCGA-TGCT S	0.875 (±0.054)	0.914 (±0.048)	0.911 (±0.061)	0.907 (±0.073)	0.935 (±0.051)	**0.941 (±0.055)**	0.791 (±0.113)	0.929 (±0.045)	0.930 (±0.056)
	TCGA-TGCT MGCT	0.884 (±0.053)	0.802 (±0.091)	0.817 (±0.093)	0.790 (±0.128)	0.812 (±0.123)	0.822 (±0.112)	**0.925 (±0.049)**	0.934 (±0.051)	0.691 (±0.226)
	TCGA-CESC SCC	0.952 (±0.016)	0.915 (±0.041)	0.969 (±0.010)	0.958 (±0.018)	0.968 (±0.010)	**0.970 (±0.010)**	0.363 (±0.263)	0.966 (±0.017)	0.962 (±0.014)
	TCGA-CESC A	0.951 (±0.017)	0.878 (±0.031)	0.946 (±0.017)	0.953 (±0.019)	0.964 (±0.015)	**0.967 (±0.011)**	0.866 (±0.184)	0.944 (±0.019)	0.960 (±0.017)

Table H.1. Extended evaluation metrics in the sequence of B→R→N→E→T→C in the **in-domain setting**.

Metric \ Method	Subtype	MergeSlide (TCP)	MergeSlide (naive)	ADaFGrad	ConSlide	DER++	AGEM	ER-ACE	Joint	Naive Finetuning
Balanced ACC (%)		**85.112 (±1.570)**	77.785 (±3.940)	61.836 (±5.562)	61.336 (±5.791)	57.951 (±3.115)	41.864 (±4.186)	60.808 (±4.769)	76.101 (±3.098)	23.259 (±5.204)
Masked Balanced ACC (%)		**89.575 (±2.440)**	**89.575 (±2.440)**	86.110 (±3.217)	85.945 (±4.393)	83.469 (±2.902)	76.207 (±3.793)	77.125 (±4.800)	86.716 (±3.018)	82.06 (±3.905)
Weighted F1		**0.897 (±0.009)**	0.853 (±0.010)	0.768 (±0.056)	0.750 (±0.063)	0.769 (±0.052)	0.473 (±0.112)	0.739 (±0.029)	0.851 (±0.028)	0.198 (±0.080)
Macro F1		**0.830 (±0.012)**	0.741 (±0.012)	0.600 (±0.071)	0.587 (±0.071)	0.569 (±0.040)	0.341 (±0.059)	0.555 (±0.044)	0.753 (±0.026)	0.156 (±0.068)
Precision per class	TCGA-BRCA IDC	0.941 (±0.015)	0.941 (±0.029)	0.942 (±0.044)	**0.964 (±0.016)**	0.898 (±0.062)	0.467 (±0.382)	0.876 (±0.067)	0.907 (±0.054)	0.000 (±0.000)
	TCGA-BRCA ILC	0.714 (±0.087)	**0.759 (±0.086)**	0.677 (±0.201)	0.607 (±0.160)	0.712 (±0.170)	0.192 (±0.293)	0.725 (±0.259)	0.768 (±0.099)	0.488 (±0.489)
	TCGA-RCC CC	**0.997 (±0.005)**	0.977 (±0.012)	0.928 (±0.040)	0.908 (±0.039)	0.917 (±0.044)	0.892 (±0.096)	0.924 (±0.051)	0.935 (±0.050)	0.786 (±0.393)
	TCGA-RCC P	0.868 (±0.024)	0.847 (±0.045)	**0.918 (±0.039)**	0.909 (±0.040)	0.902 (±0.062)	0.736 (±0.317)	0.829 (±0.138)	0.903 (±0.071)	0.546 (±0.451)
	TCGA-RCC ChRCC	0.607 (±0.108)	0.788 (±0.126)	0.796 (±0.113)	0.879 (±0.126)	0.841 (±0.138)	**0.897 (±0.207)**	0.756 (±0.284)	0.819 (±0.141)	0.974 (±0.048)
	TCGA-NSCLC SCC	**0.946 (±0.047)**	0.929 (±0.059)	0.810 (±0.106)	0.794 (±0.107)	0.831 (±0.099)	0.719 (±0.155)	0.699 (±0.136)	0.859 (±0.088)	0.733 (±0.209)
	TCGA-NSCLC A	0.842 (±0.065)	**0.856 (±0.055)**	0.772 (±0.125)	0.741 (±0.120)	0.769 (±0.086)	0.517 (±0.218)	0.856 (±0.060)	0.821 (±0.119)	0.333 (±0.166)
	TCGA-ESCA SCC	**0.967 (±0.022)**	0.651 (±0.217)	0.531 (±0.443)	0.564 (±0.395)	0.408 (±0.429)	0.343 (±0.433)	0.348 (±0.336)	0.652 (±0.268)	0.200 (±0.400)
	TCGA-ESCA A	**0.891 (±0.055)**	0.385 (±0.203)	0.217 (±0.350)	0.200 (±0.400)	0.125 (±0.301)	0.000 (±0.000)	0.262 (±0.209)	0.584 (±0.229)	0.000 (±0.000)
	TCGA-TGCT S	**0.899 (±0.030)**	0.737 (±0.096)	0.749 (±0.291)	0.802 (±0.126)	0.604 (±0.406)	0.562 (±0.394)	0.455 (±0.170)	0.811 (±0.083)	0.000 (±0.000)
	TCGA-TGCT MGCT	0.443 (±0.356)	**0.455 (±0.390)**	0.375 (±0.437)	0.451 (±0.405)	0.175 (±0.354)	0.100 (±0.300)	0.056 (±0.096)	0.698 (±0.322)	0.000 (±0.000)
	TCGA-CESC SCC	**0.652 (±0.094)**	0.469 (±0.154)	0.395 (±0.174)	0.352 (±0.178)	0.401 (±0.215)	0.221 (±0.120)	0.119 (±0.299)	0.726 (±0.135)	0.039 (±0.019)
	TCGA-CESC A	**0.962 (±0.014)**	0.741 (±0.062)	0.513 (±0.172)	0.518 (±0.164)	0.526 (±0.169)	0.309 (±0.080)	0.708 (±0.184)	0.756 (±0.120)	0.172 (±0.062)
Recall per class	TCGA-BRCA IDC	**0.930 (±0.015)**	0.904 (±0.022)	0.779 (±0.090)	0.675 (±0.086)	0.838 (±0.081)	0.511 (±0.419)	0.851 (±0.057)	0.907 (±0.043)	0.000 (±0.000)
	TCGA-BRCA ILC	0.748 (±0.061)	0.829 (±0.054)	0.825 (±0.100)	**0.890 (±0.042)**	0.626 (±0.208)	0.390 (±0.471)	0.582 (±0.279)	0.704 (±0.175)	0.083 (±0.116)
	TCGA-RCC CC	0.809 (±0.055)	0.884 (±0.075)	0.885 (±0.045)	0.883 (±0.043)	**0.908 (±0.052)**	0.768 (±0.188)	0.855 (±0.199)	0.928 (±0.042)	0.221 (±0.193)
	TCGA-RCC P	0.945 (±0.022)	**0.951 (±0.024)**	0.861 (±0.065)	0.882 (±0.038)	0.806 (±0.162)	0.370 (±0.367)	0.884 (±0.046)	0.902 (±0.070)	0.211 (±0.232)
	TCGA-RCC ChRCC	**0.993 (±0.015)**	0.926 (±0.075)	0.817 (±0.092)	0.605 (±0.242)	0.712 (±0.143)	0.436 (±0.221)	0.671 (±0.255)	0.830 (±0.113)	0.224 (±0.140)
	TCGA-NSCLC SCC	0.846 (±0.021)	0.838 (±0.039)	0.857 (±0.076)	0.869 (±0.070)	0.853 (±0.050)	0.731 (±0.247)	**0.951 (±0.023)**	0.856 (±0.092)	0.440 (±0.277)
	TCGA-NSCLC A	**0.959 (±0.021)**	0.907 (±0.032)	0.803 (±0.095)	0.799 (±0.067)	0.845 (±0.049)	0.317 (±0.356)	0.712 (±0.124)	0.881 (±0.051)	0.055 (±0.036)
	TCGA-ESCA SCC	**0.831 (±0.070)**	0.817 (±0.049)	0.144 (±0.176)	0.069 (±0.062)	0.091 (±0.097)	0.034 (±0.043)	0.357 (±0.394)	0.689 (±0.182)	0.008 (±0.017)
	TCGA-ESCA A	**0.973 (±0.019)**	0.803 (±0.081)	0.026 (±0.060)	0.012 (±0.030)	0.008 (±0.016)	0.000 (±0.000)	0.488 (±0.362)	0.517 (±0.241)	0.000 (±0.000)
	TCGA-TGCT S	0.845 (±0.094)	**0.911 (±0.139)**	0.306 (±0.339)	0.498 (±0.329)	0.328 (±0.267)	0.168 (±0.251)	0.863 (±0.111)	0.955 (±0.053)	0.000 (±0.000)
	TCGA-TGCT MGCT	**0.536 (±0.222)**	0.440 (±0.415)	0.254 (±0.333)	0.232 (±0.224)	0.020 (±0.041)	0.050 (±0.150)	0.350 (±0.391)	0.476 (±0.217)	0.000 (±0.000)
	TCGA-CESC SCC	**0.810 (±0.062)**	0.547 (±0.098)	0.757 (±0.079)	0.784 (±0.128)	0.763 (±0.107)	0.808 (±0.053)	0.052 (±0.121)	0.540 (±0.208)	0.800 (±0.111)
	TCGA-CESC A	0.904 (±0.053)	0.497 (±0.073)	0.962 (±0.027)	0.951 (±0.037)	0.965 (±0.023)	**0.966 (±0.025)**	0.486 (±0.276)	0.834 (±0.135)	0.968 (±0.028)
AUC per class	TCGA-BRCA IDC	0.941 (±0.014)	**0.946 (±0.009)**	0.891 (±0.039)	0.915 (±0.036)	0.847 (±0.041)	0.677 (±0.196)	0.822 (±0.105)	0.912 (±0.027)	0.468 (±0.098)
	TCGA-BRCA ILC	0.945 (±0.013)	**0.965 (±0.011)**	0.949 (±0.016)	0.952 (±0.022)	0.927 (±0.027)	0.801 (±0.121)	0.919 (±0.045)	0.951 (±0.015)	0.888 (±0.034)
	TCGA-RCC CC	0.977 (±0.024)	**0.980 (±0.019)**	0.958 (±0.02)	0.947 (±0.025)	0.956 (±0.020)	0.927 (±0.032)	0.964 (±0.026)	0.972 (±0.026)	0.819 (±0.108)
	TCGA-RCC P	0.982 (±0.006)	**0.986 (±0.005)**	0.975 (±0.016)	0.974 (±0.018)	0.970 (±0.018)	0.932 (±0.041)	0.963 (±0.035)	0.983 (±0.011)	0.916 (±0.045)
	TCGA-RCC ChRCC	0.982 (±0.019)	**0.988 (±0.013)**	0.976 (±0.011)	0.976 (±0.017)	0.979 (±0.013)	0.939 (±0.062)	0.983 (±0.015)	0.986 (±0.013)	0.861 (±0.102)
	TCGA-NSCLC SCC	**0.972 (±0.005)**	0.957 (±0.009)	0.943 (±0.027)	0.954 (±0.021)	0.949 (±0.012)	0.927 (±0.028)	0.956 (±0.015)	0.957 (±0.016)	0.928 (±0.029)
	TCGA-NSCLC A	**0.973 (±0.004)**	0.941 (±0.011)	0.917 (±0.038)	0.926 (±0.034)	0.927 (±0.018)	0.673 (±0.201)	0.961 (±0.010)	0.955 (±0.024)	0.414 (±0.183)
	TCGA-ESCA SCC	0.980 (±0.020)	**0.989 (±0.011)**	0.494 (±0.127)	0.556 (±0.112)	0.414 (±0.209)	0.565 (±0.176)	0.839 (±0.228)	0.863 (±0.058)	0.647 (±0.196)
	TCGA-ESCA A	**0.980 (±0.020)**	0.943 (±0.056)	0.930 (±0.053)	0.912 (±0.036)	0.911 (±0.047)	0.939 (±0.032)	0.921 (±0.072)	0.956 (±0.033)	0.943 (±0.027)
	TCGA-TGCT S	0.696 (±0.232)	0.775 (±0.227)	0.743 (±0.247)	0.774 (±0.223)	0.723 (±0.254)	**0.806 (±0.199)**	0.787 (±0.186)	0.783 (±0.223)	0.818 (±0.192)
	TCGA-TGCT MGCT	0.704 (±0.225)	0.607 (±0.354)	**0.777 (±0.131)**	0.777 (±0.205)	0.505 (±0.182)	0.657 (±0.180)	0.814 (±0.191)	0.838 (±0.113)	0.681 (±0.169)
	TCGA-CESC SCC	0.959 (±0.013)	0.926 (±0.032)	**0.967 (±0.014)**	0.962 (±0.015)	0.959 (±0.024)	0.964 (±0.023)	0.465 (±0.243)	0.972 (±0.017)	0.958 (±0.021)
	TCGA-CESC A	0.955 (±0.014)	0.895 (±0.034)	0.958 (±0.010)	0.956 (±0.017)	0.955 (±0.018)	**0.964 (±0.017)**	0.826 (±0.098)	0.928 (±0.022)	0.960 (±0.013)

Table H.2. Extended evaluation metrics on the sequence of B→R→N→E→T→C in the **out-of-domain setting**.

Metric \ Method	Subtype	MergeSlide (TCP)	MergeSlide (naive)	ADaFGrad	ConSlide	DER++	AGEM	ER-ACE	Joint	Naive Finetuning
Balanced ACC (%)		**87.930 (±2.112)**	80.636 (±1.865)	77.096 (±4.474)	70.821 (±2.180)	61.037 (±3.319)	43.961 (±3.009)	54.613 (±4.323)	81.151 (±2.283)	24.976 (±3.33)
Balanced Masked ACC		**92.109 (±1.700)**	**92.109 (±1.700)**	90.216 (±1.677)	88.896 (±2.442)	84.758 (±1.929)	79.458 (±3.215)	79.428 (±3.682)	90.138 (±1.854)	86.614 (±3.543)
Weighted F1		**0.899 (±0.013)**	0.789 (±0.022)	0.793 (±0.038)	0.683 (±0.033)	0.623 (±0.066)	0.410 (±0.054)	0.648 (±0.026)	0.843 (±0.021)	0.177 (±0.028)
Macro F1		**0.860 (±0.019)**	0.793 (±0.021)	0.782 (±0.035)	0.684 (±0.020)	0.604 (±0.044)	0.417 (±0.026)	0.543 (±0.050)	0.824 (±0.022)	0.234 (±0.034)
Precision per class	TCGA-BRCA	IDC	**0.947 (±0.033)**	0.944 (±0.035)	0.719 (±0.129)	0.433 (±0.045)	0.433 (±0.448)	0.000 (±0.000)	0.000 (±0.000)	0.768 (±0.137)	0.000 (±0.000)
		ILC	0.730 (±0.133)	0.745 (±0.102)	0.871 (±0.057)	**0.920 (±0.052)**	0.830 (±0.058)	0.691 (±0.110)	0.726 (±0.068)	0.823 (±0.060)	0.000 (±0.000)
	TCGA-RCC	CC	**0.997 (±0.008)**	0.968 (±0.027)	0.835 (±0.111)	0.836 (±0.090)	0.793 (±0.288)	0.551 (±0.281)	0.495 (±0.495)	0.825 (±0.082)	0.000 (±0.000)
		P	0.879 (±0.043)	**0.898 (±0.040)**	0.748 (±0.146)	0.667 (±0.166)	0.422 (±0.331)	0.058 (±0.126)	0.200 (±0.400)	0.883 (±0.120)	0.000 (±0.000)
		ChRCC	0.576 (±0.061)	0.809 (±0.081)	0.850 (±0.065)	0.726 (±0.329)	**0.867 (±0.175)**	0.389 (±0.348)	0.632 (±0.083)	0.771 (±0.101)	0.300 (±0.458)
	TCGA-NSCLC	SCC	**0.957 (±0.029)**	0.952 (±0.038)	0.675 (±0.108)	0.415 (±0.063)	0.408 (±0.170)	0.109 (±0.134)	0.456 (±0.088)	0.740 (±0.094)	0.150 (±0.320)
		A	0.879 (±0.029)	0.845 (±0.029)	0.898 (±0.046)	**0.938 (±0.021)**	0.899 (±0.047)	0.186 (±0.373)	0.772 (±0.113)	0.869 (±0.044)	0.000 (±0.000)
	TCGA-ESCA	SCC	**0.943 (±0.026)**	0.747 (±0.065)	0.871 (±0.076)	0.941 (±0.035)	0.905 (±0.056)	0.551 (±0.363)	0.803 (±0.098)	0.894 (±0.038)	0.100 (±0.300)
		A	0.907 (±0.037)	0.412 (±0.038)	**0.971 (±0.028)**	0.954 (±0.047)	0.935 (±0.057)	0.916 (±0.060)	0.704 (±0.096)	0.968 (±0.025)	0.910 (±0.075)
	TCGA-TGCT	S	**0.905 (±0.033)**	0.852 (±0.058)	0.870 (±0.040)	0.873 (±0.108)	0.780 (±0.104)	0.828 (±0.084)	0.736 (±0.108)	0.928 (±0.043)	0.729 (±0.188)
		MGCT	0.761 (±0.081)	0.743 (±0.151)	0.806 (±0.108)	0.666 (±0.062)	0.853 (±0.163)	0.786 (±0.166)	**1.000 (±0.000)**	0.914 (±0.099)	0.840 (±0.159)
	TCGA-CESC	SCC	0.649 (±0.049)	0.595 (±0.081)	0.662 (±0.131)	0.555 (±0.038)	0.430 (±0.061)	0.277 (±0.030)	**0.728 (±0.150)**	0.913 (±0.049)	0.155 (±0.011)
		A	**0.959 (±0.011)**	0.822 (±0.041)	0.808 (±0.081)	0.640 (±0.158)	0.707 (±0.151)	0.697 (±0.169)	0.416 (±0.428)	0.821 (±0.120)	0.644 (±0.139)
Recall per class	TCGA-BRCA	IDC	**0.938 (±0.029)**	0.918 (±0.027)	0.585 (±0.110)	0.586 (±0.069)	0.045 (±0.047)	0.000 (±0.000)	0.000 (±0.000)	0.575 (±0.199)	0.000 (±0.000)
		ILC	0.746 (±0.183)	**0.875 (±0.133)**	0.740 (±0.114)	0.448 (±0.133)	0.456 (±0.222)	0.344 (±0.273)	0.770 (±0.070)	0.896 (±0.051)	0.000 (±0.000)
	TCGA-RCC	CC	0.821 (±0.044)	0.902 (±0.035)	0.848 (±0.182)	**0.915 (±0.048)**	0.462 (±0.406)	0.679 (±0.345)	0.125 (±0.220)	0.962 (±0.024)	0.000 (±0.000)
		P	0.940 (±0.038)	**0.974 (±0.024)**	0.715 (±0.152)	0.694 (±0.153)	0.456 (±0.291)	0.160 (±0.332)	0.083 (±0.167)	0.672 (±0.075)	0.000 (±0.000)
		ChRCC	**1.000 (±0.000)**	0.963 (±0.046)	0.641 (±0.129)	0.069 (±0.045)	0.183 (±0.105)	0.331 (±0.319)	0.655 (±0.140)	0.710 (±0.162)	0.010 (±0.016)
	TCGA-NSCLC	SCC	0.861 (±0.041)	**0.878 (±0.052)**	0.537 (±0.112)	0.771 (±0.102)	0.611 (±0.231)	0.261 (±0.336)	0.497 (±0.131)	0.489 (±0.182)	0.011 (±0.024)
		A	**0.960 (±0.026)**	0.909 (±0.043)	0.835 (±0.056)	0.741 (±0.045)	0.796 (±0.063)	0.099 (±0.199)	0.911 (±0.071)	0.918 (±0.028)	0.000 (±0.000)
	TCGA-ESCA	SCC	**0.869 (±0.055)**	0.862 (±0.046)	0.812 (±0.105)	0.693 (±0.069)	0.734 (±0.052)	0.116 (±0.120)	0.789 (±0.091)	0.882 (±0.062)	0.002 (±0.007)
		A	0.961 (±0.018)	0.787 (±0.049)	0.899 (±0.051)	0.887 (±0.044)	0.941 (±0.035)	0.880 (±0.084)	**0.973 (±0.018)**	0.955 (±0.026)	0.583 (±0.268)
	TCGA-TGCT	S	0.891 (±0.043)	0.887 (±0.074)	0.940 (±0.038)	**0.947 (±0.055)**	0.921 (±0.042)	0.734 (±0.136)	0.888 (±0.082)	0.940 (±0.048)	0.525 (±0.217)
		MGCT	0.784 (±0.075)	0.585 (±0.142)	0.935 (±0.083)	**0.958 (±0.048)**	0.897 (±0.088)	0.771 (±0.177)	0.552 (±0.337)	0.945 (±0.084)	0.734 (±0.227)
	TCGA-CESC	SCC	0.790 (±0.054)	0.570 (±0.090)	**0.972 (±0.020)**	0.963 (±0.019)	0.960 (±0.023)	0.951 (±0.027)	0.850 (±0.082)	0.938 (±0.041)	0.938 (±0.072)
		A	**0.913 (±0.020)**	0.514 (±0.084)	0.718 (±0.172)	0.757 (±0.157)	0.783 (±0.163)	0.745 (±0.180)	0.263 (±0.301)	0.802 (±0.156)	0.809 (±0.156)
AUC per class	TCGA-BRCA	IDC	0.940 (±0.032)	0.951 (±0.034)	0.944 (±0.037)	**0.953 (±0.017)**	0.535 (±0.132)	0.641 (±0.148)	0.914 (±0.052)	0.963 (±0.019)	0.298 (±0.102)
		ILC	0.945 (±0.030)	**0.965 (±0.032)**	0.939 (±0.024)	0.826 (±0.083)	0.872 (±0.038)	0.821 (±0.172)	0.889 (±0.047)	0.946 (±0.020)	0.897 (±0.022)
	TCGA-RCC	CC	0.974 (±0.009)	**0.988 (±0.008)**	0.922 (±0.052)	0.938 (±0.032)	0.834 (±0.144)	0.867 (±0.053)	0.840 (±0.117)	0.926 (±0.043)	0.594 (±0.259)
		P	0.983 (±0.011)	**0.990 (±0.007)**	0.942 (±0.052)	0.948 (±0.051)	0.780 (±0.129)	0.864 (±0.072)	0.681 (±0.193)	0.924 (±0.051)	0.791 (±0.162)
		ChRCC	0.979 (±0.010)	**0.995 (±0.003)**	0.894 (±0.042)	0.801 (±0.069)	0.809 (±0.079)	0.836 (±0.117)	0.930 (±0.020)	0.937 (±0.022)	0.499 (±0.093)
	TCGA-NSCLC	SCC	**0.975 (±0.015)**	0.965 (±0.026)	0.850 (±0.050)	0.907 (±0.031)	0.819 (±0.083)	0.677 (±0.215)	0.778 (±0.092)	0.885 (±0.047)	0.700 (±0.089)
		A	**0.975 (±0.015)**	0.952 (±0.023)	0.952 (±0.025)	0.954 (±0.010)	0.926 (±0.021)	0.826 (±0.097)	0.955 (±0.028)	0.972 (±0.012)	0.718 (±0.125)
	TCGA-ESCA	SCC	**0.977 (±0.012)**	0.971 (±0.012)	0.953 (±0.027)	0.925 (±0.017)	0.908 (±0.042)	0.773 (±0.114)	0.960 (±0.019)	0.970 (±0.008)	0.834 (±0.039)
		A	0.977 (±0.012)	0.967 (±0.012)	**0.985 (±0.009)**	0.985 (±0.011)	0.982 (±0.012)	0.970 (±0.021)	0.980 (±0.012)	0.991 (±0.005)	0.904 (±0.077)
	TCGA-TGCT	S	0.875 (±0.054)	0.914 (±0.048)	**0.988 (±0.007)**	0.988 (±0.010)	0.985 (±0.011)	0.970 (±0.014)	0.983 (±0.007)	0.991 (±0.006)	0.949 (±0.025)
		MGCT	0.884 (±0.053)	0.801 (±0.092)	0.994 (±0.006)	**0.996 (±0.004)**	0.995 (±0.004)	0.976 (±0.028)	0.991 (±0.013)	0.997 (±0.004)	0.967 (±0.042)
	TCGA-CESC	SCC	0.952 (±0.016)	0.915 (±0.041)	0.949 (±0.042)	**0.955 (±0.034)**	0.937 (±0.049)	0.937 (±0.042)	0.634 (±0.167)	0.936 (±0.041)	0.944 (±0.046)
		A	0.951 (±0.017)	0.878 (±0.031)	0.958 (±0.035)	**0.958 (±0.032)**	0.943 (±0.043)	0.947 (±0.036)	0.798 (±0.161)	0.956 (±0.040)	0.946 (±0.039)

Table H.3. Extended evaluation metrics on the reversed sequence of C→T→E→N→R→B in the **in-domain setting**.

I.2. Performance Drop Comparison

When new tasks are added, the performance on previously learned tasks often degrades. We provide Fig. I.1 to examine the performance drop of each task as new tasks are sequentially introduced to the first five tasks. Overall, MergeSlide maintains the most stable performance across tasks. Notably, several tasks experience significant performance drops: TCGA-NSCLC, TCGA-ESCA, and TCGA-TGCT. For TCGA-NSCLC, DER++ shows a substantial drop after training on the last task (TCGA-CESC), while other methods appear more stable. For TCGA-ESCA, all models except MergeSlide show significant drops after training on both the fifth task (TCGA-TGCT) and the last task (TCGA-CESC). For TCGA-TGCT, although MergeSlide initially performs worse than others when the task is first introduced, it shows greater stability afterward. In contrast, other models suffer a noticeable decline in TCGA-TGCT performance after the final task is added.

I.3. Confidence Score Study

To further explore model behavior in its predictions, we provide Fig. I.2, which shows the average confidence scores for the ground-truth cancer subtyping class positions. For instance, given the prediction logits $\hat{p}_i \in \mathbb{R}^C$ for each WSI, where C is the total number of cancer subtypes so far, we take the j-th logit value $\hat{p}_{i,j}$ to visualize, where j indicates the ground-truth cancer subtype. We observe the following: while other methods tend to show decreased confidence on recent tasks (e.g., ESCA, TGCT) and a significant shift toward the last task, CESC, MergeSlide maintains stable confidence across all six tasks, with no significant degradation on any of them.

I.4. What Happens When the MLP is Also Trained?

As described, only the backbone $f_{\mathcal{A}}$ is trained when a new task is added, in order to obtain the corresponding weight set θ_t. We do not train the MLP as a classifier but instead use class-aware prompt embeddings $E^{\mathcal{C},t}$. Results in Tab. I.1 show that the MLP-free strategy not only avoids performance degradation but also slightly improves both Masked ACC and Mean ACC compared to trainable MLPs, regardless of whether `TCP` is used. The only exception is overall ACC when using `TCP`, which exhibits a very marginal decrease.

Method	MLP Setting	ACC	Masked ACC	Mean ACC
MergeSlide (naive)	Trainable MLP	78.613 (±1.710)	93.534 (±0.831)	90.075 (±1.145)
	MLP-free (ours)	**78.797 (±1.916)**	**93.617 (±1.098)**	**90.640 (±1.247)**
MergeSlide `w/ TCP`	Trainable MLP	**91.935 (±0.897)**	93.534 (±0.831)	92.137 (±1.287)
	MLP-free (ours)	91.914 (±0.980)	**93.617 (±1.098)**	**92.686 (±0.899)**

Table I.1. Comparison between two strategies for MLP: Trainable MLP and MLP-Free on the sequence of B→R→N→E→T→C.

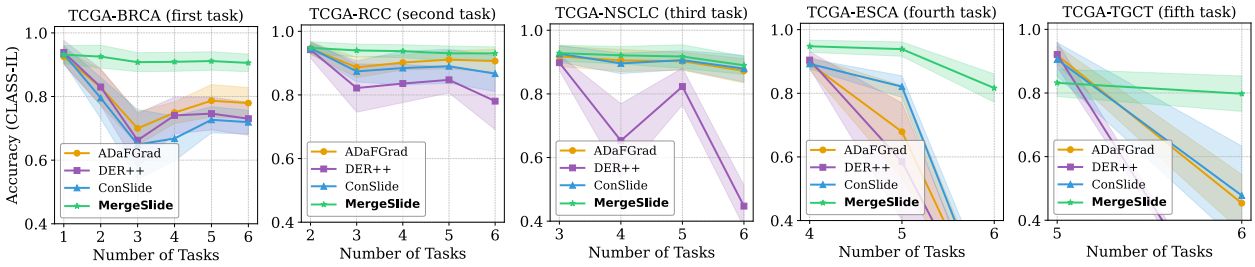

Figure I.1. Performance drop comparisons across different methods as new tasks are added to the first five datasets: TCGA-BRCA, TCGA-RCC, TCGA-NSCLC, TCGA-ESCA, and TCGA-TGCT (sequence: B→R→N→E→T→C).

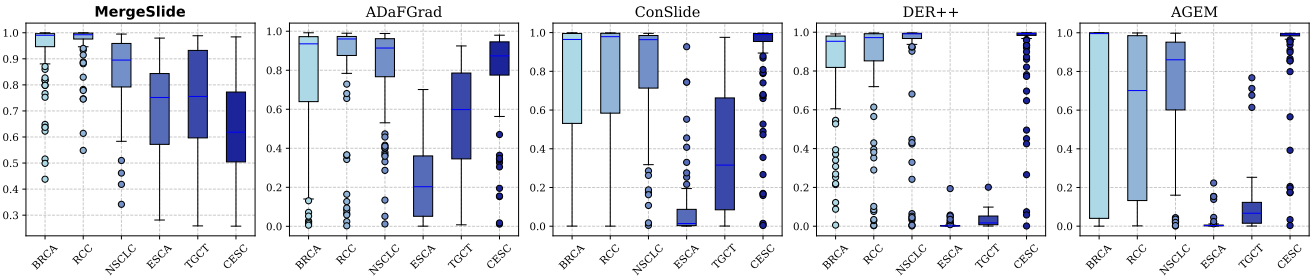

Figure I.2. Comparison of average confidence scores for target cancer subtyping after training on the last task, TCGA-CESC (sequence: B→R→N→E→T→C).

I.5. Task-wise Performance

We provide a task-wise performance comparison between the top three methods under the CLASS-IL and TASK-IL scenarios in Tabs. I.3 and I.4, respectively, after training on the last task, i.e., TCGA-CESC. As a result, under the CLASS-IL scenario, MergeSlide still outperforms the others on five out of six tasks. Even without TCP, it achieves better performance than the others on four out of six tasks. However, we also observe that rehearsal-based methods tend to perform poorly on earlier tasks with limited WSIs, such as TCGA-ESCA and TCGA-TGCT. Indeed, after training on the final task (TCGA-CESC), the performance on this dataset becomes much higher compared to those with fewer samples. Nonetheless, MergeSlide maintains stable performance across tasks.

I.6. Task Addition

Final accuracy after merging weights reflects the end performance; however, performance during the addition of new tasks should also be examined to ensure stability. The comparison of CLASS-IL ACC for MergeSlide against other methods is reported in Tab. I.2, where the number of accumulated tasks $T \in \{2, 3, 4, 5\}$ is evaluated. The results show that some methods exhibit unstable performance *while MergeSlide demonstrates consistent performances as the number of tasks increases.* $\sigma < 1\%$ further confirms the stability. We further compare performance drops on the two earliest tasks, TCGA-BRCA and TCGA-RCC, across top methods and MergeSlide without TCP (Fig. I.1). MergeSlide maintains stable performance as new tasks arrive, while others show significant and moderate drops on TCGA-BRCA and TCGA-RCC, respectively.

Method	Number of increasing tasks				σ
	T=2	*T=3*	*T=4*	*T=5*	
ER-ACE	90.045 (±1.968)	81.210 (±2.793)	79.297 (±4.234)	74.849 (±5.461)	5.525
AGEM	84.839 (±9.500)	70.843 (±4.327)	75.236 (±6.225)	75.169 (±4.972)	5.121
DER++	92.244 (±3.177)	86.693 (±2.215)	88.104 (±1.323)	82.307 (±2.946)	3.549
ConSlide	86.733 (±2.228)	80.643 (±3.226)	83.929 (±1.968)	85.257 (±2.068)	2.250
ADaFGrad	90.945 (±3.214)	86.002 (±2.225)	87.382 (±1.908)	85.252 (±2.271)	2.187
MergeSlide	**92.907 (±1.337)**	**92.571 (±0.713)**	**93.309 (±0.493)**	**92.345 (±0.990)**	**0.364**

Table I.2. Closer look at MergeSlide compared to other continual learning methods as the number of tasks increases on the sequence of B→R→N→E→T→C (Metric: CLASS-IL ACC). σ shows the standard deviation of results across $T \in \{2, 3, 4, 5\}$.

I.7. Sensitivity of number of patches K

To further examine the sensitivity of the sampling patch number K, we report the bACC of MergeSlide under both CLASS-IL and TASK-IL settings for $K \in 50, 100, 200, 300, 400$ in Tab. I.3. In the CLASS-IL scenario, results without TCP remain stable, while with TCP the performance gap between $K = 50$ and $K = 400$ is only 4.997%. In the TASK-IL scenario, performance is also stable across different K values. Notably, even the lowest setting ($K = 50$) still outperforms all comparative baselines.

Method	BRCA	RCC	NSCLC
Zero-shot	89.922 (±2.981)	83.474 (±2.891)	83.874 (±2.766)
DER++ [2] ($\|\mathcal{B}_r\| = 30$)	73.026 (±5.194)	78.055 (±8.936)	44.730 (±6.570)
ConSlide [7] ($\|\mathcal{B}_r\| = 30$)	71.940 (±3.949)	86.700 (±5.921)	87.933 (±4.409)
ADaFGrad [1] ($\|\mathcal{B}_r\| = 30$)	77.938 (±4.871)	90.647 (±3.636)	87.252 (±3.843)
MergeSlide	90.784 (±3.311)	**93.418 (±2.351)**	**93.657 (±2.633)**

Method	ESCA	TGCT	CESC
Zero-shot	62.388 (±3.704)	79.771 (±5.316)	17.578 (±2.928)
DER++ [2] ($\|\mathcal{B}_r\| = 30$)	2.836 (±2.707)	0.526 (±1.053)	93.516 (±1.310)
ConSlide [7] ($\|\mathcal{B}_r\| = 30$)	8.657 (±6.814)	45.371 (±9.516)	**93.531 (±2.209)**
ADaFGrad [1] ($\|\mathcal{B}_r\| = 30$)	6.269 (±5.247)	47.787 (±15.559)	93.594 (±1.148)
MergeSlide	**94.925 (±2.331)**	**81.343 (±5.616)**	92.031 (±1.389)

Figure I.3. Task-wise performances on stream of datasets B→R→N→E→T→C under CLASS-IL scenario.

Method	BRCA	RCC	NSCLC
Zero-shot	91.643 (±2.969)	83.591 (±2.914)	89.917 (±1.857)
DER++ [2] ($\|\mathcal{B}_r\| = 30$)	84.783 (±5.802)	89.113 (±5.243)	88.952 (±2.989)
ConSlide [7] ($\|\mathcal{B}_r\| = 30$)	87.604 (±4.920)	92.574 (±4.466)	92.265 (±2.503)
ADaFGrad [1] ($\|\mathcal{B}_r\| = 30$)	86.582 (±4.129)	**93.652 (±2.434)**	91.458 (±2.817)
MergeSlide	**93.359 (±3.353)**	93.645 (±2.100)	**92.939 (±2.539)**

Method	ESCA	TGCT	CESC
Zero-shot	88.060 (±2.111)	81.862 (±5.855)	**95.859 (±1.108)**
DER++ [2] ($\|\mathcal{B}_r\| = 30$)	87.185 (±3.379)	**92.153 (±3.806)**	93.913 (±1.201)
ConSlide [7] ($\|\mathcal{B}_r\| = 30$)	87.797 (±4.649)	91.641 (±7.755)	94.099 (±2.119)
ADaFGrad [1] ($\|\mathcal{B}_r\| = 30$)	88.152 (±3.832)	90.799 (±7.266)	94.362 (±1.164)
MergeSlide	**95.075 (±1.772)**	91.066 (±2.433)	94.297 (±1.443)

Figure I.4. Task-wise performances on stream of datasets B→R→N→E→T→C under TASK-IL scenario.

K	bACC (CLASS-IL)		Masked bACC (TASK-IL)
	MergeSlide w/o TCP	MergeSlide w/ TCP	MergeSlide
50	78.810 (±1.942)	82.932 (±1.802)	90.935 (±1.870)
100	80.167 (±1.752)	84.284 (±2.478)	91.610 (±1.535)
200	80.078 (±1.952)	86.269 (±1.975)	91.761 (±1.907)
300	79.849 (±1.944)	87.416 (±2.015)	92.083 (±1.514)
400	**80.668 (±1.860)**	**87.929 (±2.110)**	**92.087 (±1.740)**

Table I.3. Sensitivity of K sampling patches under the CLASS-IL and TASK-IL scenarios on B→R→N→E→T→C.

References

[1] Doanh C. Bui, Hoai Luan Pham, Vu Trung Duong Le, Tuan Hai Vu, Van Duy Tran, Khang Nguyen, and Yasuhiko Nakashima. Lifelong whole slide image analysis: Online vision-language adaptation and past-to-present gradient distillation. *IEEE Access*, pages 1–1, 2025. 1, 4, 9

[2] 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. 1, 9

[3] Lucas Caccia, Rahaf Aljundi, Nader Asadi, Tinne Tuytelaars, Joelle Pineau, and Eugene Belilovsky. New insights on reducing abrupt representation change in online continual learning. *arXiv preprint arXiv:2104.05025*, 2021. 1

[4] Arslan Chaudhry, Marc'Aurelio Ranzato, Marcus Rohrbach, and Mohamed Elhoseiny. Efficient lifelong learning with a-gem. *arXiv preprint arXiv:1812.00420*, 2018. 1, 4

[5] Alan Kaylor Cline and Inderjit S Dhillon. Computation of the singular value decomposition. In *Handbook of linear algebra*, pages 45–1. Chapman and Hall/CRC, 2006. 5

[6] Tong Ding, Sophia J Wagner, Andrew H Song, Richard J Chen, Ming Y Lu, Andrew Zhang, Anurag J Vaidya, Guillaume Jaume, Muhammad Shaban, Ahrong Kim, et al. Multimodal whole slide foundation model for pathology. *arXiv preprint arXiv:2411.19666*, 2024. 3

[7] 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, 2023. 1, 4, 9

[8] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. icarl: Incremental classifier and representation learning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 2001–2010, 2017. 3