Incremental Nuclei Segmentation from Histopathological Images via Future-class Awareness and Compatibility-inspired Distillation

Huyong Wang[1], Huisi Wu[1]*, Jing Qin[2]
[1] College of Computer Science and Software Engineering, Shenzhen University
[2] Centre for Smart Health, School of Nursing, The Hong Kong Polytechnic University
2210273011@email.szu.edu.cn, hswu@szu.edu.cn, harry.qin@polyu.edu.hk

Abstract

We present a novel semantic segmentation approach for incremental nuclei segmentation from histopathological images, which is a very challenging task as we have to incrementally optimize existing models to make them perform well in both old and new classes without using training samples of old classes. Yet, it is an indispensable component of computer-aided diagnosis systems. The proposed approach has two key techniques. First, we propose a new future-class awareness mechanism by separating some potential regions for future classes from background based on their similarities to both old and new classes in the representation space. With this mechanism, we can not only reserve more parameter space for future updates but also enhance the representation capability of learned features. We further propose an innovative compatibility-inspired distillation scheme to make our model take full advantage of the knowledge learned by the old model. We conducted extensive experiments on two famous histopathological datasets and the results demonstrate the proposed approach achieves much better performance than state-of-the-art approaches. The code is available at https://github.com/why19991/InSeg.

1. Introduction

Cellular nuclei segmentation aims to accurately delineate various cell nuclei from histopathological images, which is of great importance for cancer diagnosis, treatment, and prognosis prediction [37]. Recent years, deep learning models have achieved remarkable performance in this task [8, 20, 32, 34, 40], significantly promoting the development of computer-aided diagnosis systems. Despite these progresses, most existing models are incapable of maintaining the performance on previously learned classes when it is required to incrementally learn new classes, which is common in clinical practice [35]. This phenomenon is called *catastrophic forgetting* [24], which is a long-standing challenge and prohibits these models from being deployed in clinical settings. A common solution is to rebuild the training dataset containing annotations of all classes and retrain the model. However, the cost of collecting histopathological images with pixel-level annotations for segmentation is extremely laborious and time-consuming. Besides, access to previous data is usually limited by patient privacy. In this regard, we often have to incrementally optimize existing models to make them perform well in both old classes and new classes without using training samples of old classes.

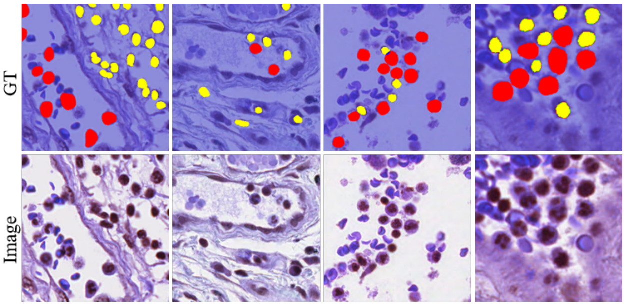


Figure 1. Pathological images typically contain numerous objects from different categories that have very similar appearances. For instance, epithelia (represented by red objects) and lymphocytes (represented by yellow objects) look quite alike. However, in the 1-1 setting at step 0, epithelia are labeled as new classes, whereas lymphocytes are seen as background. Our approach aims to mine such future classes in an unsupervised way and learn them in advance.

Incremental learning is proposed to solve the problem of catastrophic forgetting, and semantic segmentation based on incremental learning is known as incremental semantic segmentation (ISS). In scenarios where previous data can be reused, some studies [22, 45] select samples to help the model review and reinforce knowledge, greatly alleviating catastrophic forgetting. However, when privacy or storage issues make old data unavailable, which is common

*Corresponding Author

in clinical scenarios, most studies turn to employ distillation techniques [4, 13, 23, 25, 26, 30, 36] to constrain the model updating. These methods, while keeping the performance of old classes (stability) well, limit the model's capability in learning new classes (plasticity). In order to improve both stability and plasticity, a few studies [5, 39] propose to learn future classes in advance to prepare for learning new classes. Specifically, SSUL-M [5] applies a salient object detector to detect potential future classes while MicroSeg [39] utilizes mask proposals generated by Mask2Former [9] to detect future classes. Unfortunately, both the salient detector and mask proposals still heavily depend on a number of annotated data, which is, as mentioned, difficult to acquire in clinical practice due to privacy and cost. To the end, it is necessary to aware future classes in an unsupervised manner on the given data.

In this work, we present a novel ISS method for incremental nuclei segmentation from histopathological images, which has two key components: a new future-class awareness mechanism and an innovative compatibility-inspired distillation scheme. Histopathological images often contain numerous objects belonging to different categories with great similarity in appearance. As illustrated in Figure 1, some nuclei are learned as new classes or old classes, while others (future classes) are treated as background. Thus, we propose to separate some potential regions of future classes from background by calculating their similarities to both old and new classes in the representation space, and then we train the new model with both new and old classes, as well as with the separated regions. To the end, we cannot only reserve more parameter space for future updates but also enhance the representation capability of learned features. Furthermore, we design a compatibility-inspired distillation scheme to further improve both stability and plasticity. As the number of targeting classes of the updated model is always more than that of the old model, directly aligning them poses conflicts and background shifts [4]. In this regard, we propose to expand the old model's prediction range to match that of the new model by harnessing the representation space of the future classes. This scheme not only resolves alignment conflicts and background shifts but also enables the new model to effectively utilize the knowledge of the old model. Our contributions can be summarized as follows.

- We propose a novel ISS method for incremental nuclei segmentation from histopathological images, which is able to incrementally optimize existing models to make them perform well in both old and new classes without using training samples of old classes.
- We propose a new future-class awareness mechanism to reserve more parameter space for incremental learning, and an innovative compatibility-inspired distillation scheme to make our model take full advantage of the knowledge learned by the old model.
- We demonstrate the proposed method achieve a better balance between stability and plasticity than state-of-the-art methods via extensive experiments on two famous public histopathological datasets, MoNuSAC [33] and CoNSeP [14].

2. Related Work

2.1. Nuclei Segmentation

There have been some preliminary works for nuclei segmentation based on traditional techniques and deep learning methods. Traditional techniques based on background subtraction and color threshold [27, 29] fail to generalize to complex scenarios such as overlap and occlusion in histopathology images. With the development of CNN, deep networks have been widely used in nuclei segmentation [31, 32, 34, 42], significantly improving the segmentation performance in the supervised setting. However, most of these networks are not incrementally designed and tend to forget previously learned classes when learning new types of cell nuclei continually.

2.2. Class Incremental Learning

Class incremental learning (CIL) aims to address the problem of catastrophic forgetting when learning knowledge for new classes. Existing CIL techniques can be summarized into regularization-based, replay-based, and architecture-based approaches. Regularization-based approaches utilize consistency constraints between new and old models to prevent significant changes to the knowledge associated with previously learned classes. The constraints can be applied on features [10, 12, 18], the output logits [4, 11, 21], the weights [1, 2, 19], or the gradients [6, 15]. Architecture-based approaches [16, 38, 44] dynamically grow the network to extend model capacity for new tasks. To better retain knowledge of old classes, replay-based approaches store and re-use a subset of previous training data, including raw images [3, 22, 45], features [17, 43] or generated data [7, 28].

2.3. Class Incremental Semantic Segmentation

Recently, several works have extended and combined existing incremental learning methods to semantic segmentation. MiB [4] proposes a novel unbiased function to model background shift. PLOP [13] adopts pseudo label to solve background shift and retain spatial dependencies by local pooling operation. CoNuSeg [35] maintains the relationships between the prototypes of old classes to preserve the semantic information of old classes. EWF [36] propose a weight fusion strategy to fuse parameters from new and old models. SSUL-M [5] and MicroSeg [39] further propose pre-learning future classes to prepare for the future.

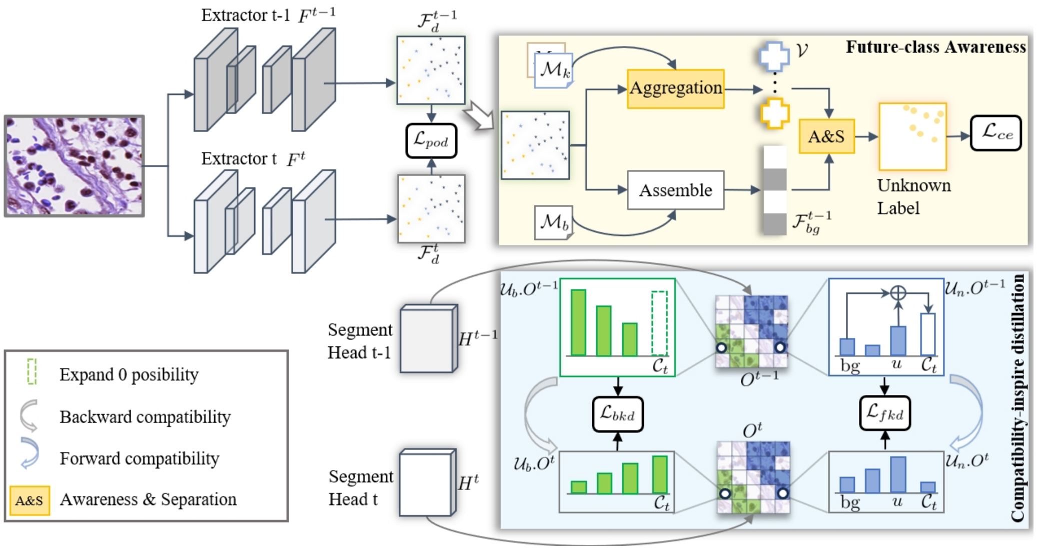


Figure 2. The overall framework. The old $(t-1)$ and new models (t) consist of the same feature extractor F and segmentation head H. During training, we freeze the old model and update the new model by minimizing the weighted sum of $\mathcal{L}_{ce}$, $\mathcal{L}_{pod}$, $\mathcal{L}_{bkd}$ and $\mathcal{L}_{fkd}$.

3. Method

3.1. Problem Formulation

Class incremental semantic segmentation involves conducting semantic segmentation in a series of steps, with the underlying assumption that there are a total of T steps. In step t, we have training set $D_t = \{(x, y)\}$ and label set $\mathcal{C}_t$, where x is an RGB image $\in \mathbb{R}^{C\times H\times W}$ and $y \in \mathcal{C}_t$ is the corresponding annotation. The label set $\mathcal{C}_t$ only contains newly added classes at step t, while previous learned classes $\mathcal{C}_0 \cup \mathcal{C}_1 \cup \ldots \cup \mathcal{C}_{t-1}$ and future classes $\mathcal{C}_{t+1:T}$ are considered as background, which can lead to catastrophic forgetting of the previously learned classes. Given a dataset $\mathcal{D}_t$, our goal is to achieve a model M with parameter θ_t that can perform well on all seen classes $\mathcal{C}_{0:t}$. Without loss of generality, M consists of a feature extractor F and a segmentation head H, which makes $M = F \circ H$, where $\circ$ denotes function composition.

3.2. Framework Overview

The proposed incremental semantic segmentation framework, depicted in Figure 2, includes an old and new model. Under the incremental setting, both old and future classes are treated as background. To correct pixels of old classes, we adopt the pseudo-labeling strategy to assign them pseudo-labels. For pixels of potential future classes, we propose a new future-class awareness mechanism (detailed in Section. 3.3) to assign them unknown labels. At the feature level, we apply multi-scale local distillation with Local POD (described in Section. 3.5) between the corresponding layers of the two models to avoid a large feature discrepancy. At the output level, we design an innovative compatibility-inspired distillation (depicted in Section. 3.4) between the model's predictions to achieve better stability and plasticity. After training, we fuse the parameters from the two models by using the Endpoints Weight Fusion strategy (EWF) [36]. Specific details about EWF can be found in the supplementary materials.

3.3. Future-class Awareness

Our future-class awareness mechanism corrects the background pixels that are considered potential future classes, i.e., pixels that are visually similar to a new or old class. We propose to measure the similarity S by utilizing the class-centroids $\mathcal{V}$ in the representation space of the old model. Subsequently, we assign unknown labels to these pixels for joint learning with both new and old classes. This approach enhances the representation capability of learned features and better prepares the model for the future.

Image-wise Centroids Aggregation. Class-centroids (prototypes) are representatives of each class in the representation space and serve as an effective tool to measure similarities between classes [30]. Given an image as input, we first generate pseudo-labels $\hat{y}(i, c)$ for old classes in the background using predictions O^{t-1} of the old model and the ground-truth label $y(i, c)$ at step t :

$$\hat{y}(i,c) = \begin{cases} 1 & \text{if } y(i,c) = 1 \\ 1 & \text{if } y(i,c) = 0 \text{ and } c = \operatorname{argmax} O^{t-1}(i,c) \end{cases} \tag{1}$$

This strategy can effectively re-assign labels for old classes in the background but fails to separate potential future classes hidden in the background. Next, we generate class-aware mask $\mathcal{M}_c$ for class c under the guidance of pseudo

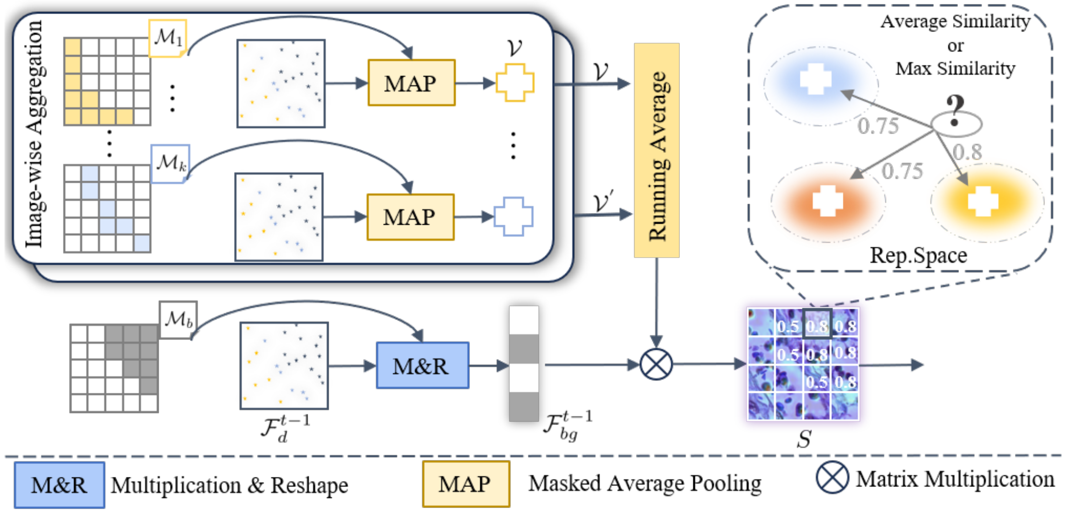


Figure 3. Illustration of future-class awareness. Centroids representing both old and new categories are aggregated in representation space, then a class-aware matrix $\mathcal{S}$ is computed between all centroids and background features to aware future classes.

label $\hat{y}(i, c)$:

$$\mathcal{M}_c(i) = \begin{cases} 1 & \text{if } \hat{y}(i,c) = 1 \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

Then, for a certain class k, we average the deepest input feature $\mathcal{F}_d^{t-1}$ with the class-aware mask to obtain class-centroid $\mathcal{V}_k$:

$$\begin{aligned} \mathcal{V}_k &= MAP(\mathcal{M}_k, \mathcal{F}_d^{t-1}) \\ &= \frac{\sum_i \mathcal{F}_d^{t-1}(i)\mathbb{1}[\mathcal{M}_k(i) = 1]}{\sum_i \mathbb{1}[\mathcal{M}_k(i) = 1]} \end{aligned} \tag{3}$$

where MAP represents the masked average pooling operation. After that, we obtain a centroid set $\mathcal{V} = \{\mathcal{V}_1, ..., \mathcal{V}_{\mathcal{C}_t}\}$ containing old and new class-centeroids, which will be utilized to detect future classes.

Cross Image-wise Centroids Aggregation. Different cell nuclei are typically distributed across different images, which suggests that potential future classes may not coexist in the same image as their similar cell nuclei. Therefore, aggregating class centroids in an image-wise manner may not include centroids of classes that are similar to future classes. To address this issue, we propose to update class centroids using a running average approach. The final class centroids are computed as follows:

$$\mathcal{V}_k = \frac{1}{N}(\mathcal{V}_k^{'} \cdot N_k^{'} + \sum_i \mathcal{F}_d^{t-1}(i)\mathbb{1}[\mathcal{M}_k(i) = 1]) \tag{4}$$

$$N = N_k^{'} + \sum_i \mathbb{1}[\mathcal{M}_k(i) = 1] \tag{5}$$

where $\mathcal{V}_k^{'}$ and $N_k^{'}$ represent the accumulated result and total pixel number of class k from the initial update to the last update, respectively.

Awareness & Separation. To aware future classes, we first assemble the background features $\mathcal{F}_{bg} \in \mathbb{R}^{C \times N_b}$ via the masked multiplication on the deepest feature $\mathcal{F}_d^{t-1}$ with the background mask $\mathcal{M}_b$. We then calculate a pixel-to-centroid affinity matrix $\mathcal{A}$ between pixels of the reshaped background feature $\mathcal{F}_{bg}$ and class-centroids $\mathcal{V}$ via a matrix multiplication operation Mat:

$$\mathcal{A} = Mat(Norm(\mathcal{F}_{bg})^T, Norm(\mathcal{V})) \tag{6}$$

where $Norm$ indicates $l2$-normalization. The size of $\mathcal{A}$ is $\mathbb{R}^{N \times P}$ and $P = |\mathcal{V}|$ is the cardinality of class-centroid set. Through the affinity matrix, we construct a semantic graph for each pixel-centroid pair, where each vertex represents the semantic structure and the edge represents the similarity relationship, thus a fine-grained similarity can be measured by utilizing the information in the representation space.

We obtain the final class-aware matrix $\mathcal{S}$ by taking the average or maximum across the second dimension of affinity matrix $\mathcal{A}$. To effectively pre-learn knowledge of future classes, only background pixels that are sufficiently similar are re-assigned unknown labels. The final corrected ground truth $y(i, c)$ of pixel i is defined as:

$$y(i,c) = \begin{cases} u & \text{if } \mathcal{M}_b(i) = 1 \text{ and } \mathcal{S}(i) < \tau_u \\ \hat{y}(i,c) & \text{otherwise} \end{cases} \tag{7}$$

where u represents the unknown label assigned to future classes, and τ_u is a threshold controlling the confidence of future classes. In other words, for background pixels filtered by Equation 1, we re-label them with an unknown label u on the guidance of the class-aware matrix S. Otherwise, we directly copy their pseudo-labels (for old classes) or true labels (for new classes).

3.4. Compatibility-inspired Distillation

Considering that the previous model only produces predictions O^{t-1} for $|\mathcal{C}_{0:t-1}|$ classes, while the new model can produce predictions O^t for $|\mathcal{C}_{0:t}|$ ($|\mathcal{C}_{0:t}| > |\mathcal{C}_{0:t-1}|$) classes, directly aligning the outputs between the two models is contradictory. Treating the model's predictions as interfaces of software, we introduce new interfaces into the old model to be compatible with the new model (i.e., expanding O^{t-1} to match O^t, and padding the additional channels with 0). The expanded prediction O^{t-1} becomes:

$$O^{t-1}(i,c) = \begin{cases} 0 & \text{if } c \ \in \mathcal{C}_t \\ O^{t-1}(i,c) & \text{otherwise} \end{cases} \tag{8}$$

Then we enforce constraints to pixels of new class set $\mathcal{U}_n$ and background set $\mathcal{U}_b$ from the view of backward compatibility and forward compatibility respectively. It is worth noting that the background here includes old classes, future classes, and true background pixels.

Forward compatibility. For pixels i from the new class set, we transfer the old model's background prediction $O^{t-1}(i,b)$ to the prediction corresponding to the ground-truth to address background shift [4]. Next, with the future class-aware mechanism, the old model has pre-learned knowledge about the new classes (which are future classes to the old model). Thus we also add the old model's future class prediction $O^{t-1}(i,u)$ to the prediction corresponding to the ground truth, thereby naturally inheriting the knowledge that the old model has prepared for future learning (forward compatibility). After that, we reset the predictions of background and future classes to 0 while keeping other predictions unchanged. The corrected prediction of the old model is defined as:

$$O^{t-1}(i,c) = \begin{cases} O^{t-1}(i,u) \ + \ O^{t-1}(i,b) \quad \text{if } c \ \in \mathcal{C}_t \\ O^{t-1}(i,c) \quad \text{if } c \ \in \mathcal{C}_{0:t-1} \\ 0 \quad \text{otherwise} \end{cases} \tag{9}$$

where $i \in \mathcal{U}_n$ and u indicates the future class at step $t-1$.

Backward compatibility. For pixels $i \in \mathcal{U}_b$ from the background set, similar to regular knowledge distillation, we only need to align the outputs between the two models to ensure that the new model does not deviate significantly from the old model during updates. The complete compatibility-inspired distillation loss is defined as:

$$\mathcal{L}_{cpd} = \alpha_n \cdot \mathcal{L}_{fkd}(x, \mathcal{U}_n) + \alpha_b \cdot \mathcal{L}_{bkd}(x, \mathcal{U}_b) \tag{10}$$

where α_n and α_b control the contribution of the new classes set $\mathcal{U}_n$ and background set $\mathcal{U}_b$, which are discussed in Section. 4.3; $\mathcal{L}(x, \mathcal{U})$ is a common knowledge distillation loss:

$$\mathcal{L}(x, \mathcal{U}) = -\frac{1}{|\mathcal{U}|} \sum_{i \in \mathcal{U}} \sum_{c}^{\mathcal{C}_{0:t}} O_x^{t-1}(i,c) \log O_x^t(i,c) \tag{11}$$

We emphasize that our distillation approach supplements and corrects the contents extended in the interfaces of the old model, thus addressing alignment conflict and background shift, as well as facilitating the new model to effectively leverage the knowledge learned by the old model.

3.5. Feature-based Distillation

To prevent catastrophic forgetting in feature space, we also utilize multi-scale local distillation with Local POD [13] to retain knowledge by preserving multi-scale spatial information. The Local POD loss is formulated as:

$$\mathcal{L}_{pod} = \frac{1}{L} \sum_{l=1}^{L} |\phi(\mathcal{F}_l^t) - \phi(\mathcal{F}_l^{t-1})|^2 \tag{12}$$

where $\mathcal{F}_l^t$ is the feature of the l-th layer and $\phi(.)$ is a function that captures multi-scale spatial statistics.

Finally, we train the model with the total loss as follows:

$$\mathcal{L}_{all} = \mathcal{L}_{ce} + \mathcal{L}_{cpd} + \lambda \cdot \mathcal{L}_{pod} \tag{13}$$

where λ denotes the weight of LocalPOD, which is set to 0.0001 in our experiments.

4. Experiment

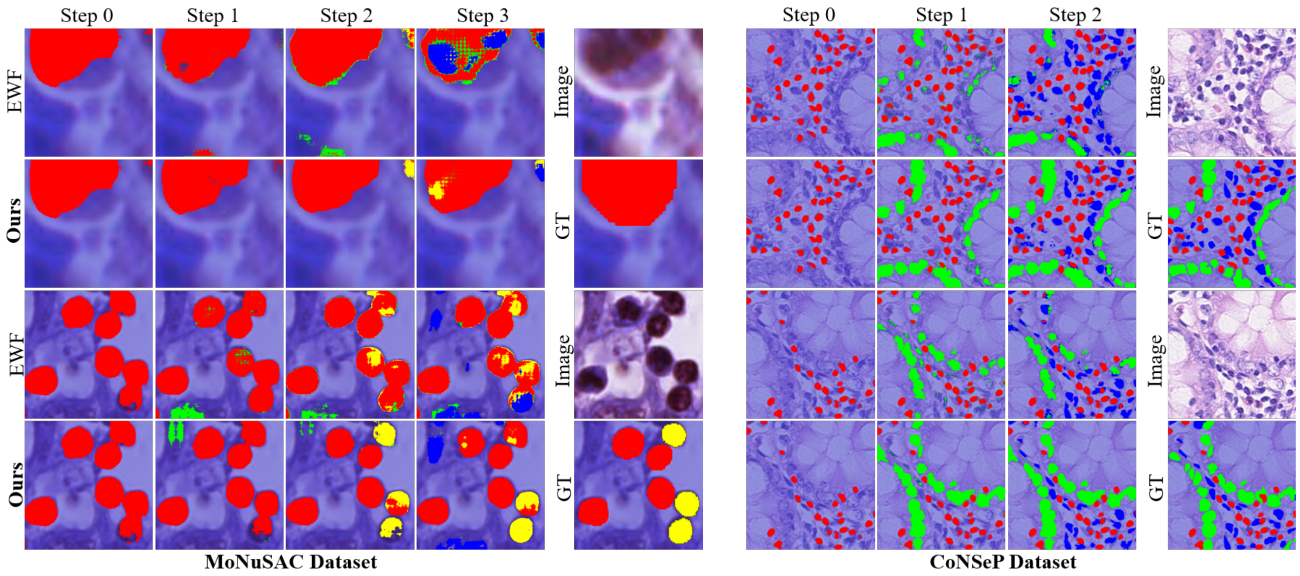
4.1. Datasets, Protocols, and Evaluation

Datasets. We conduct all the experiments on two publicly available nuclei datasets, MoNuSAC and CoNSeP. MoNuSAC contains H&E stained tissue images of four organs, along with annotations for multiple cell types, including epithelial (Epith), lymphocyte (Lymph), macrophage (Macro), and neutrophil (Neutr). We randomly split 20% of the training set for validation. Besides, all images in the training, testing, and validation sets have been uniformly cropped to a resolution of 320×320. Finally, the MoNuSAC contains 1177 training images, 295 validation images and 651 test images. CoNSeP contains H&E stained image tiles with annotations of 7 cell types, where we divide them into 3 classes: epithelial, spindle-shaped and others. With the same pre-processing, we obtain a training set with 345, a validation set with 87, and a test set with 224 images.

Protocols. There are two common settings in ISS benchmarks: disjoint and overlapped. In the disjoint setting, D_t only contains images of new classes and old classes. In the overlapped setting, D_t can contain images of all classes, including old classes, new classes, and future classes. Regarding of setting, only new classes are annotated, while others are treated as background. The overlapped setting is considered more realistic since it allows future classes to appear in the current dataset, reflecting real-world scenarios more accurately. In this work, we only focus on the overlapped setting.

Method	1-1 (4 tasks)			2-1 (3 tasks)			2-2 (2 tasks)			3-1 (2 tasks)		
	Old	New	Mean	Old	New	Mean	Old	New	Mean	Old	New	Mean
ILT [25]	33.46	53.91	48.80	50.00	73.30	61.65	46.00	68.18	57.09	62.11	73.27	64.90
MiB [4]	59.95	64.57	63.41	52.51	74.67	63.59	57.33	76.87	67.10	63.85	75.38	66.73
SDR [26]	55.50	67.30	64.35	53.47	74.37	63.92	57.08	78.03	67.56	65.70	77.31	68.60
PLOP [13]	60.14	65.60	64.23	56.85	75.87	66.36	56.74	75.88	66.31	66.16	77.43	68.97
REMINDER [30]	62.41	66.49	65.47	56.57	75.89	66.23	59.89	76.03	67.96	68.22	77.18	70.46
CoNuSeg [35]	65.37	67.92	67.28	57.85	76.09	66.97	61.70	77.11	69.40	66.47	77.33	69.18
IDEC [41]	62.16	69.41	67.59	58.63	76.03	67.33	63.22	75.41	69.31	67.36	77.12	69.80
EWF [36]	64.57	67.89	67.06	60.27	76.23	68.26	**65.90**	75.94	70.92	69.06	77.62	71.20
Ours	**68.11**	**69.86**	**69.44**	**63.79**	**76.53**	**70.16**	65.43	**77.21**	**71.32**	**70.46**	**78.03**	**72.36**
Joint	72.01	75.29	74.47	70.36	78.58	74.47	70.36	78.58	74.47	72.65	79.93	74.47

Table 1. Incremental semantic segmentation results (mDice) on MoNuSAC. The best is in **bold**.



MoNuSAC Dataset

CoNSeP Dataset

Figure 4. Visualization of the proposed method and EWF at different incremental steps on two datasets.

Evaluation. The ISS benchmark configurations are represented as n-m where n and m correspond to the number of new classes to be learned during the initial step and each subsequent step, respectively. We set several ISS settings for each dataset, e.g., on MoNuSAC 3-1, 2-1, 2-2, and 1-1 respectively include learning 3 classes then 1 class (2 steps), 2 classes then 1 class (3 steps), 2 classes then 2 classes (2 steps) and 1 class at initial step then 1 class at subsequent step (4 steps). We emphasized that the 1-1 setting is the most challenging task due to its small amount of data and high number of steps. Similarly, on CoNSeP 1-1 (3 steps) and 2-1 (2 steps). After training all steps, we use mean Dice (mDice) to evaluate the performance.

Implementation Details. We use the ResNet-101 pretrained on ImageNet as the feature encoder for all experiments. During training, the initial learning rate is set to 0.05 for the first step and 0.01 for subsequent incremental steps. The number of epochs for all steps is set to 100. We employ an SGD optimizer with a batch size of 12, distributed across three 2080Ti GPUs, to train the model.

Comparison Results. To demonstrate the advantages of our method, we employ the state-of-the-art approach (EWF) as our baseline for comparison. Table 1 displays our approach's comparative results against some strong ISS methods on the MoNuSAC dataset. Evidently, the performance of ILT is the worst in all incremental setups, since it lacks the capacity to address background shift. Integrating background correction methods, such as MiB and

PLOP, leads to marked improvements for both old and new classes in all settings. By further introducing innovative distillation mechanisms into the background correction method, approaches like REMINDER, CoNuSeg, and IDEC have achieved additional performance gains. EWF receives higher performance after applying the Endpoints Weight Fusion strategy to PLOP.

Despite their great progress, our approach surpasses them by a large margin on all settings. In particular, on the most challenging setup (1-1), our method outperforms the EWF by 3.54 % in old classes and IDEC by 0.45% in new classes. While the gains in new classes are not that large, we believe that the performance is already quite better compared to that of old classes. As presented in Figure 5, our model achieves all the highest mDice scores from incremental step 1 to step 3. This also indicates that the stability and plasticity of our model are superior to those of other strong models. Experiments in Table 2 further demonstrate the strong generalization ability of our method.

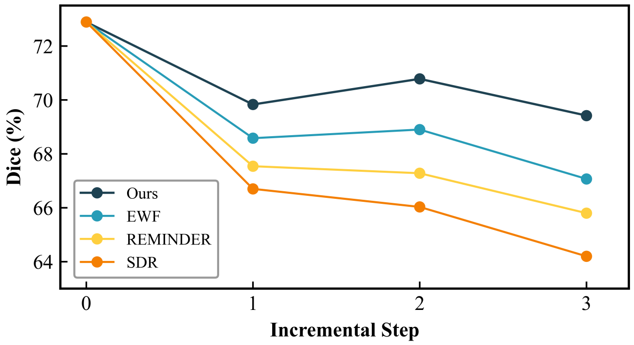


Figure 5. Results of mDice at different incremental steps.

Method	1-1 (3 tasks)			2-1 (2 tasks)		
	Old	New	Mean	Old	New	Mean
ILT [25]	35.23	63.19	44.55	40.82	63.61	48.41
MiB [4]	67.02	67.06	67.04	70.99	61.96	67.98
SDR [26]	67.05	67.29	67.21	70.37	62.92	67.88
PLOP [13]	67.06	67.78	67.54	72.50	63.15	69.38
CoNuSeg [35]	**67.52**	68.59	68.23	72.80	63.41	69.67
IDEC [41]	67.20	68.28	67.92	72.16	62.78	69.03
EWF [36]	67.18	70.03	69.08	74.16	63.00	70.44
Ours	66.64	**70.84**	**69.44**	**74.29**	**64.65**	**71.08**

Table 2. Incremental semantic segmentation results (mDice) on the CoNSep dataset.

Visual Results. We visualize the segmentation results of our method and EWF on two datasets to highlight the comparative effectiveness. As shown in Figure 4, our method yields segmentation results that are more complete and accurate than those produced by EWF. For example, for the second image (rows 3-4, left), both methods gradually forget class epithelial (the red ones) and macrophage (the green ones) from step 0 to step 2. Nevertheless, our method still yields satisfactory segmentation results after step 3. Similar results can be observed in other images. Figure 6 explicitly presents the Dice score for each class at every step.

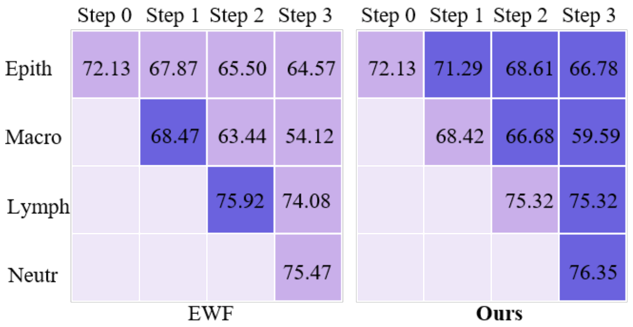


Figure 6. Values of Dice metric for each class at different incremental steps on MoNuSAC 1-1.

4.2. Ablation Study

We verify the effectiveness of the proposed method using 1-1 and 3-1 setups on MoNuSAC. From Table 3, we can find that the performance of PLOP + EWF surpasses that of PLOP. Only adding future-class awareness (FCA) can further boost the performance of EWF across different setups. After applying compatibility-inspired distillation (CID), we abtain the highest performance on MoNuSAC. In particular, the performance of PLOP + EWF + FCA + CID outperforms that of PLOP + EWF in 1-1 settings, with a great improvement of 3.54% in old classes and 1.97% in new classes. These results effectively demonstrate that both proposed modules can effectively improve the plasticity and stability of the model, thereby improving the overall performance.

Method	1-1 (4 tasks)			3-1 (2 tasks)		
	Old	New	Mean	Old	New	Mean
PLOP	60.14	65.60	64.23	66.16	77.43	68.97
+ EWF	64.57	67.89	67.06	69.06	77.62	71.20
+ FCA	67.42	69.00	68.60	70.11	77.93	72.07
+ CID	**68.11**	**69.86**	**69.42**	**70.46**	**78.03**	**72.36**

Table 3. Ablation study results (mDice) on the MoNuSAC dataset.

4.3. Further Analysis

Future-class Awareness. To better understand how pre-learning of future classes enhances the results, we visualize t-SNE distribution of features generated from the baseline and baseline + FCA. As shown in Figure 7, the baseline tends to generate more ambiguous features while the baseline + FCA is able to learn more discriminative features. These robust features then help the model to resist catastrophic forgetting better. Similar conclusions can be drawn from the feature map (Figure 8). The baseline model exhibits a lower response to target features and is prone to generating erroneous high responses in the background, whereas the baseline with FCA generates more accurate and higher responses to target features. This phenomenon further shows that training with FCA can enhance the representation power to alleviate knowledge forgetting and reserve more parameter space to be updated.

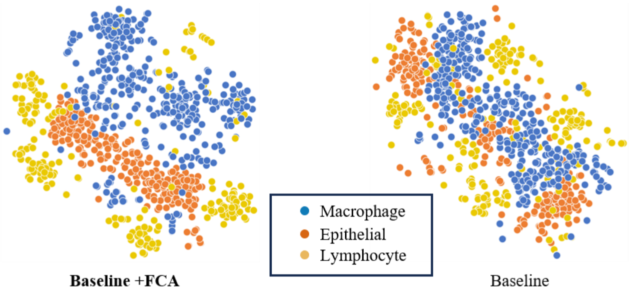


Figure 7. T-SNE visualization of features learned by Baseline and Baseline + FCA at final incremental step on MoNuSAC 1-1.

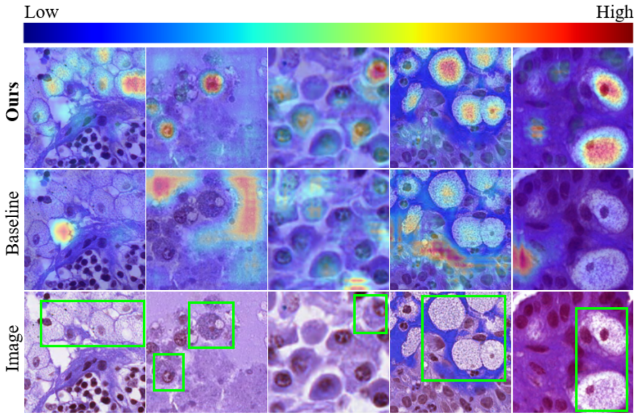


Figure 8. Feature maps of the Baseline and Baseline + FCA at final incremental step on MoNuSAC 1-1.

Compatibility-inspired Distillation. As shown in Table 4, slightly increasing both α_n and α_b to 1 can improve the performance of both new and old classes. However, a large value of α_n, e.g., $\alpha_n = 5$ leads the model to focus excessively on new classes, compromising the performance of old classes and thus negatively affecting overall performance. Similarly, when we set α_b a high value 3, the model tends to overly remember old classes and overlook the new classes. In our experiments, we achieve the best performance when α_n and α_b are set to 3 and 1, respectively. For additional ablation studies and detailed analytical experiments, please refer to the supplementary materials.

α_n	α_b	Background	Old	New	All
0	0	94.58	67.42	69.00	73.80
1	1	94.53	67.71	69.31	74.03
3	1	94.43	68.11	**69.86**	**74.42**
5	1	94.51	65.91	69.18	73.59
3	0	94.70	67.61	69.25	74.01
3	2	**94.74**	68.08	69.41	74.21
3	3	94.53	**68.18**	68.34	73.54

Table 4. Effect of the weighted factors α_n and α_b.

5. Limitation

We develop our method based on the attributes of histopathological images, where future classes in the dataset exhibit visual similarities to the learned classes. Therefore, our approach may not perform optimally in datasets where future classes significantly differ from the learned classes, which necessitates further research. Nevertheless, considering the vast diversity of cancer cell types, our method shows promise in enhancing computer-aided diagnosis systems with more robust incremental updating capabilities.

6. Conclusion

In this work, we present a novel method for incremental semantic segmentation from histopathological images, aiming to solve catastrophic forgetting without storing samples of previous data. First, we propose a new future-class awareness approach that detects future classes in an unsupervised way. Second, drawing inspiration from software engineering, we introduce an innovative compatibility-inspired knowledge distillation to make the new model take full advantage of the knowledge learned by the old model. Finally, comparative experiments and ablation studies clearly demonstrate the effectiveness of the proposed method.

Acknowledgments

This work was supported partly by National Natural Science Foundation of China (No. 62273241), Natural Science Foundation of Guangdong Province, China (No. 2024A1515011946), and the Theme-based Research Scheme from Hong Kong Research Grants Council (project no. T45-401/22-N).

References

[1] Rahaf Aljundi, Punarjay Chakravarty, and Tinne Tuytelaars. Expert gate: Lifelong learning with a network of experts. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3366–3375, 2017. 2

[2] Rahaf Aljundi, Francesca Babiloni, Mohamed Elhoseiny, Marcus Rohrbach, and Tinne Tuytelaars. Memory aware synapses: Learning what (not) to forget. In *Proceedings of the European conference on computer vision (ECCV)*, pages 139–154, 2018. 2

[3] Jihwan Bang, Heesu Kim, YoungJoon Yoo, Jung-Woo Ha, and Jonghyun Choi. Rainbow memory: Continual learning with a memory of diverse samples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8218–8227, 2021. 2

[4] Fabio Cermelli, Massimiliano Mancini, Samuel Rota Bulo, Elisa Ricci, and Barbara Caputo. Modeling the background for incremental learning in semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9233–9242, 2020. 2, 5, 6, 7

[5] Sungmin Cha, YoungJoon Yoo, Taesup Moon, et al. Ssul: Semantic segmentation with unknown label for exemplar-based class-incremental learning. *Advances in neural information processing systems*, 34:10919–10930, 2021. 2

[6] Arslan Chaudhry, Marc'Aurelio Ranzato, Marcus Rohrbach, and Mohamed Elhoseiny. Efficient lifelong learning with a-gem. *arXiv preprint arXiv:1812.00420*, 2018. 2

[7] Jingfan Chen, Yuxi Wang, Pengfei Wang, Xiao Chen, Zhaoxiang Zhang, Zhen Lei, and Qing Li. Diffusepast: Diffusion-based generative replay for class incremental semantic segmentation. *arXiv preprint arXiv:2308.01127*, 2023. 2

[8] Yiqi Chen, Xuanya Li, Kai Hu, Zhineng Chen, and Xieping Gao. Nuclei segmentation in histopathology images using rotation equivariant and multi-level feature aggregation neural network. In *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 549–554. IEEE, 2020. 1

[9] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1290–1299, 2022. 2

[10] Prithviraj Dhar, Rajat Vikram Singh, Kuan-Chuan Peng, Ziyan Wu, and Rama Chellappa. Learning without memorizing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5138–5146, 2019. 2

[11] Jiahua Dong, Wenqi Liang, Yang Cong, and Gan Sun. Heterogeneous forgetting compensation for class-incremental learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11742–11751, 2023. 2

[12] Arthur Douillard, Matthieu Cord, Charles Ollion, Thomas Robert, and Eduardo Valle. Podnet: Pooled outputs distillation for small-tasks incremental learning. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16*, pages 86–102. Springer, 2020. 2

[13] Arthur Douillard, Yifu Chen, Arnaud Dapogny, and Matthieu Cord. Plop: Learning without forgetting for continual semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4040–4050, 2021. 2, 5, 6, 7

[14] Simon Graham, Quoc Dang Vu, Shan E Ahmed Raza, Ayesha Azam, Yee Wah Tsang, Jin Tae Kwak, and Nasir Rajpoot. Hover-net: Simultaneous segmentation and classification of nuclei in multi-tissue histology images. *Medical Image Analysis*, 58:101563, 2019. 2

[15] Yiduo Guo, Wenpeng Hu, Dongyan Zhao, and Bing Liu. Adaptive orthogonal projection for batch and online continual learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6783–6791, 2022. 2

[16] Zhiyuan Hu, Yunsheng Li, Jiancheng Lyu, Dashan Gao, and Nuno Vasconcelos. Dense network expansion for class incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11858–11867, 2023. 2

[17] Ahmet Iscen, Jeffrey Zhang, Svetlana Lazebnik, and Cordelia Schmid. Memory-efficient incremental learning through feature adaptation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVI 16*, pages 699–715. Springer, 2020. 2

[18] Minsoo Kang, Jaeyoo Park, and Bohyung Han. Class-incremental learning by knowledge distillation with adaptive feature consolidation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16071–16080, 2022. 2

[19] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017. 2

[20] Shyam Lal, Devikalyan Das, Kumar Alabhya, Anirudh Kanfade, Aman Kumar, and Jyoti Kini. Nucleisegnet: Robust deep learning architecture for the nuclei segmentation of liver cancer histopathology images. *Computers in Biology and Medicine*, 128:104075, 2021. 1

[21] Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE transactions on pattern analysis and machine intelligence*, 40(12):2935–2947, 2017. 2

[22] Huiwei Lin, Baoquan Zhang, Shanshan Feng, Xutao Li, and Yunming Ye. Pcr: Proxy-based contrastive replay for online class-incremental continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24246–24255, 2023. 1, 2

[23] Zihan Lin, Zilei Wang, and Yixin Zhang. Continual semantic segmentation via structure preserving and projected feature alignment. In *European Conference on Computer Vision*, pages 345–361. Springer, 2022. 2

[24] Michael McCloskey and Neal J Cohen. Catastrophic interference in connectionist networks: The sequential learning

problem. In *Psychology of learning and motivation*, pages 109–165. Elsevier, 1989. 1

[25] Umberto Michieli and Pietro Zanuttigh. Incremental learning techniques for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision workshops*, pages 0–0, 2019. 2, 6, 7

[26] Umberto Michieli and Pietro Zanuttigh. Continual semantic segmentation via repulsion-attraction of sparse and disentangled latent representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1114–1124, 2021. 2, 6, 7

[27] Shivang Naik, Scott Doyle, Shannon Agner, Anant Madabhushi, Michael Feldman, and John Tomaszewski. Automated gland and nuclei segmentation for grading of prostate and breast cancer histopathology. In *2008 5th IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pages 284–287. IEEE, 2008. 2

[28] Oleksiy Ostapenko, Mihai Puscas, Tassilo Klein, Patrick Jahnichen, and Moin Nabi. Learning to remember: A synaptic plasticity driven framework for continual learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11321–11329, 2019. 2

[29] Nobuyuki Ostu. A threshold selection method from gray-level histograms. *IEEE Trans SMC*, 9:62, 1979. 2

[30] Minh Hieu Phan, Son Lam Phung, Long Tran-Thanh, Abdesselam Bouzerdoum, et al. Class similarity weighted knowledge distillation for continual semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16866–16875, 2022. 2, 3, 6

[31] Shan E Ahmed Raza, Linda Cheung, Muhammad Shaban, Simon Graham, David Epstein, Stella Pelengaris, Michael Khan, and Nasir M Rajpoot. Micro-net: A unified model for segmentation of various objects in microscopy images. *Medical image analysis*, 52:160–173, 2019. 2

[32] Eric Upschulte, Stefan Harmeling, Katrin Amunts, and Timo Dickscheid. Uncertainty-aware contour proposal networks for cell segmentation in multi-modality high-resolution microscopy images. In *Competitions in Neural Information Processing Systems*, pages 1–12. PMLR, 2023. 1, 2

[33] Ruchika Verma, Neeraj Kumar, Abhijeet Patil, Nikhil Cherian Kurian, Swapnil Rane, Simon Graham, Quoc Dang Vu, Mieke Zwager, Shan E Ahmed Raza, Nasir Rajpoot, et al. Monusac2020: A multi-organ nuclei segmentation and classification challenge. *IEEE Transactions on Medical Imaging*, 40(12):3413–3423, 2021. 2

[34] Huisi Wu, Zhaoze Wang, Youyi Song, Lin Yang, and Jing Qin. Cross-patch dense contrastive learning for semi-supervised segmentation of cellular nuclei in histopathologic images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11666–11675, 2022. 1, 2

[35] Huisi Wu, Zhaoze Wang, Zebin Zhao, Cheng Chen, and Jing Qin. Continual nuclei segmentation via prototype-wise relation distillation and contrastive learning. *IEEE Transactions on Medical Imaging*, 2023. 1, 2, 6, 7

[36] Jia-Wen Xiao, Chang-Bin Zhang, Jiekang Feng, Xialei Liu, Joost van de Weijer, and Ming-Ming Cheng. Endpoints weight fusion for class incremental semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7204–7213, 2023. 2, 3, 6, 7

[37] Fuyong Xing and Lin Yang. Robust nucleus/cell detection and segmentation in digital pathology and microscopy images: a comprehensive review. *IEEE reviews in biomedical engineering*, 9:234–263, 2016. 1

[38] Chang-Bin Zhang, Jia-Wen Xiao, Xialei Liu, Ying-Cong Chen, and Ming-Ming Cheng. Representation compensation networks for continual semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7053–7064, 2022. 2

[39] Zekang Zhang, Guangyu Gao, Zhiyuan Fang, Jianbo Jiao, and Yunchao Wei. Mining unseen classes via regional objectness: A simple baseline for incremental segmentation. *Advances in Neural Information Processing Systems*, 35: 24340–24353, 2022. 2

[40] Bingchao Zhao, Xin Chen, Zhi Li, Zhiwen Yu, Su Yao, Lixu Yan, Yuqian Wang, Zaiyi Liu, Changhong Liang, and Chu Han. Triple u-net: Hematoxylin-aware nuclei segmentation with progressive dense feature aggregation. *Medical Image Analysis*, 65:101786, 2020. 1

[41] Danpei Zhao, Bo Yuan, and Zhenwei Shi. Inherit with distillation and evolve with contrast: Exploring class incremental semantic segmentation without exemplar memory. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 6, 7

[42] Yanning Zhou, Omer Fahri Onder, Qi Dou, Efstratios Tsougenis, Hao Chen, and Pheng-Ann Heng. Cia-net: Robust nuclei instance segmentation with contour-aware information aggregation. In *Information Processing in Medical Imaging: 26th International Conference, IPMI 2019, Hong Kong, China, June 2–7, 2019, Proceedings 26*, pages 682–693. Springer, 2019. 2

[43] Fei Zhu, Xu-Yao Zhang, Chuang Wang, Fei Yin, and Cheng-Lin Liu. Prototype augmentation and self-supervision for incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5871–5880, 2021. 2

[44] Kai Zhu, Wei Zhai, Yang Cao, Jiebo Luo, and Zheng-Jun Zha. Self-sustaining representation expansion for non-exemplar class-incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9296–9305, 2022. 2

[45] Lanyun Zhu, Tianrun Chen, Jianxiong Yin, Simon See, and Jun Liu. Continual semantic segmentation with automatic memory sample selection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3082–3092, 2023. 1, 2