Incremental Learning for Heterogeneous Structure Segmentation in Brain Tumor MRI

Xiaofeng Liu[1](✉), Helen A. Shih[2], Fangxu Xing[1], Emiliano Santarnecchi[1], Georges El Fakhri[1], and Jonghye Woo[1]

[1] Gordon Center for Medical Imaging, Department of Radiology, Massachusetts General Hospital and Harvard Medical School, Boston, MA 02114, USA
xliu61@mgh.harvard.edu

[2] Department of Radiation Oncology, Massachusetts General Hospital and Harvard Medical School, Boston, MA 02114, USA

Abstract. Deep learning (DL) models for segmenting various anatomical structures have achieved great success via a static DL model that is trained in a single source domain. Yet, the static DL model is likely to perform poorly in a continually evolving environment, requiring appropriate model updates. In an incremental learning setting, we would expect that well-trained static models are updated, following continually evolving target domain data—e.g., additional lesions or structures of interest—collected from different sites, without catastrophic forgetting. This, however, poses challenges, due to distribution shifts, additional structures not seen during the initial model training, and the absence of training data in a source domain. To address these challenges, in this work, we seek to progressively evolve an "off-the-shelf" trained segmentation model to diverse datasets with additional anatomical categories in a unified manner. Specifically, we first propose a divergence-aware dual-flow module with balanced rigidity and plasticity branches to decouple old and new tasks, which is guided by continuous batch renormalization. Then, a complementary pseudo-label training scheme with self-entropy regularized momentum MixUp decay is developed for adaptive network optimization. We evaluated our framework on a brain tumor segmentation task with continually changing target domains—i.e., new MRI scanners/modalities with incremental structures. Our framework was able to well retain the discriminability of previously learned structures, hence enabling the realistic life-long segmentation model extension along with the widespread accumulation of big medical data.

1 Introduction

Accurate segmentation of a variety of anatomical structures is a crucial prerequisite for subsequent diagnosis or treatment [28]. While recent advances in data-driven deep learning (DL) have achieved superior segmentation performance [29], the segmentation task is often constrained by the availability of costly pixel-wise labeled training datasets. In addition, even if static DL models are trained with

H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14221, pp. 46–56, 2023.
https://doi.org/10.1007/978-3-031-43895-0_5

extraordinarily large amounts of training datasets in a supervised learning manner [29], there exists a need for a segmentor to update a trained model with new data alongside incremental anatomical structures [24].

In real-world scenarios, clinical databases are often sequentially constructed from various clinical sites with varying imaging protocols [19–21,23]. As well, labeled anatomical structures are incrementally increased with additional lesions or new structures of interest, depending on study goals or clinical needs [18,27]. Furthermore, access to previously used data for training can be restricted, due to data privacy protocols [17,18]. Therefore, efficiently utilizing heterogeneous structure-incremental (HSI) learning is highly desired for clinical practice to develop a DL model that can be generalized well for different types of input data and varying structures involved. Straightforwardly fine-tuning DL models with either new structures [30] or heterogeneous data [17] in the absence of the data used for the initial model training, unfortunately, can easily overwrite previously learned knowledge, i.e., catastrophic forgetting [14,17,30].

At present, satisfactory methods applied in the realistic HSI setting are largely unavailable. *First*, recent structure-incremental works cannot deal with domain shift. Early attempts [27] simply used exemplar data in the previous stage. [5,18,30,33] combined a trained model prediction and a new class mask as a pseudo-label. However, predictions from the old model under a domain shift are likely to be unreliable [38]. The widely used pooled feature statistics consistency [5,30] is also not applicable for heterogeneous data, since the statistics are domain-specific [2]. In addition, a few works [13,25,34] proposed to increase the capacity of networks to avoid directly overwriting parameters that are entangled with old and new knowledge. However, the solutions cannot be domain adaptive. *Second*, from the perspective of continuous domain adaptation with the consistent class label, old exemplars have been used for the application of prostate MRI segmentation [32]. While Li et al. [17] further proposed to recover the missing old stage data with an additional generative model, hallucinating realistic data, given only the trained model itself, is a highly challenging task [31] and may lead to sensitive information leakage [35]. *Third*, while, for natural image classification, Kundu et al. [16] updated the model for class-incremental unsupervised domain adaption, its class prototype is not applicable for segmentation.

In this work, we propose a unified HSI segmentor evolving framework with a divergence-aware decoupled dual-flow (D^{3}F) module, which is adaptively optimized via HSI pseudo-label distillation using a momentum MixUp decay (MMD) scheme. To explicitly avoid the overwriting of previously learned parameters, our D^{3}F follows a “divide-and-conquer” strategy to balance the old and new tasks with a fixed rigidity branch and a compensated learnable plasticity branch, which is guided by our novel divergence-aware continuous batch renormalization (cBRN). The complementary knowledge can be flexibly integrated with the model re-parameterization [4]. Our additional parameters are constant in training, and 0 in testing. Then, the flexible D^{3}F module is trained following the knowledge distillation with novel HSI pseudo-labels. Specifically, inspired by the self-knowledge distillation [15] and self-training [38] that utilize the previous prediction for better generalization, we adaptively construct the HSI pseudo-label with an MMD scheme to smoothly adjust the contribution of potential noisy old model predictions on heterogeneous data and progressively learned new model predictions along with the training. In addition, unsupervised self-entropy minimization is added to further enhance performance.

Our main contributions can be summarized as follow:

- To our knowledge, this is the first attempt at realistic HSI segmentation with both incremental structures of interest and diverse domains.
- We propose a divergence-aware decoupled dual-flow module guided by our novel continuous batch renormalization (cBRN) for alleviating the catastrophic forgetting under domain shift scenarios.
- The adaptively constructed HSI pseudo-label with self-training is developed for efficient HSI knowledge distillation.

We evaluated our framework on anatomical structure segmentation tasks from different types of MRI data collected from multiple sites. Our HSI scheme demonstrated superior performance in segmenting all structures with diverse data distributions, surpassing conventional class-incremental methods without considering data shift, by a large margin.

2 Methodology

For the segmentation model under incremental structures of interest and domain shift scenarios, we are given an off-the-shelf segmentor $f_{\theta^0}: \mathcal{X}^0 \rightarrow \mathcal{Y}^0$ parameterized with θ^0, which has been trained with the data $\{x_n^0, y_n^0\}_{n=1}^{N^0}$ in an initial source domain $\mathcal{D}^0 = \{\mathcal{X}^0, \mathcal{Y}^0\}$, where $x_n^0 \in \mathbb{R}^{H\times W}$ and $y_n^0 \in \mathbb{R}^{H\times W}$ are the paired image slice and its segmentation mask with the height of H and width of W, respectively. There are T consecutive evolving stages with heterogeneous target domains $\mathcal{D}^t = \{\mathcal{X}^t, \mathcal{S}^t\}_{t=1}^{T}$, each with the paired slice set $\{x_n^t\}_{n=1}^{N^t} \in \mathcal{X}^t$ and the current stage label set $\{s_n^t\}_{n=1}^{N^t} \in \mathcal{S}^t$, where $x_n^t, s_n^t \in \mathbb{R}^{H\times W}$. Due to heterogeneous domain shifts, $\mathcal{X}^t$ from different sites or modalities follows diverse distributions across all T stages. Due to incremental anatomical structures, the overall label space, across the previous t stages, $\mathcal{Y}^t$ is expanded from $\mathcal{Y}^{t-1}$ with the additional annotated structures $\mathcal{S}^t$ in stage t, i.e., $\mathcal{Y}^t = \mathcal{Y}^{t-1} \cup \mathcal{S}^t = \mathcal{Y}^0 \cup \mathcal{S}^1 \cdots \cup \mathcal{S}^t$. We are targeting to learn $f_{\theta^T}: \{\mathcal{X}^t\}_{t=1}^{T} \rightarrow \mathcal{Y}^T$ that performs well on all $\{\mathcal{X}^t\}_{t=1}^{T}$ for delineating all of the structures $\mathcal{Y}^T$ seen in T stages.

2.1 cBRN Guided Divergence-Aware Decoupled Dual-Flow

To alleviate the forgetting through parameter overwriting, caused by both new structures and data shift, we propose a D^{3}F module for flexible decoupling and integration of old and new knowledge.

Specifically, we duplicate the convolution in each layer initialized with the previous model $f_{\theta^{t-1}}$ to form two branches as in [13,25,34]. The first *rigidity* branch $f_{\theta^t}^r$ is fixed at the stage t to keep the old knowledge we have learned. In contrast, the extended *plasticity* branch $f_{\theta^t}^p$ is expected to be adaptively updated

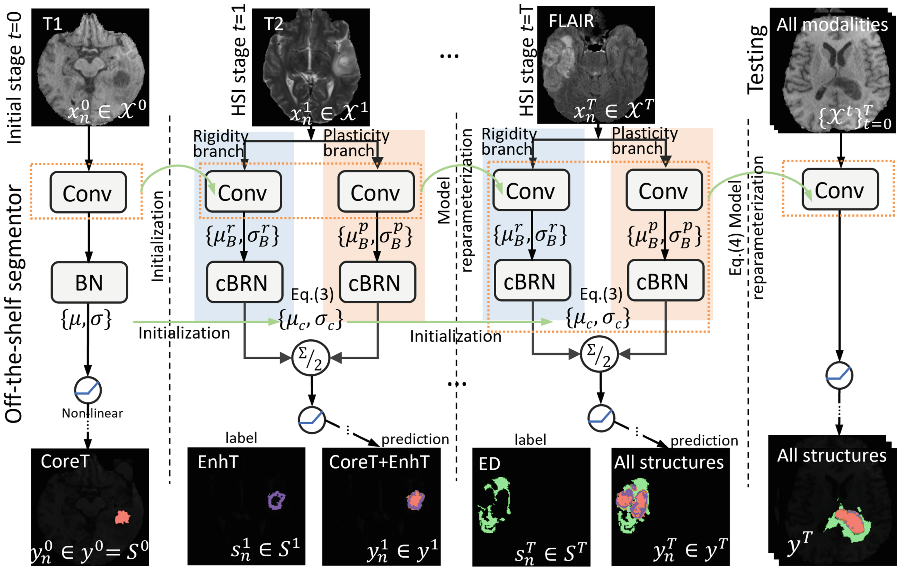

Fig. 1. Illustration of one layer in our proposed divergence-aware decoupled dual-flow module guided with cBRN for our cross-MR-modality HSI task, i.e., subject-independent (CoreT with T1) → (EnhT with T2) → (ED with FLAIR). Notably, we do not require the dual-flow or cBRN, for the initial segmentor.

to learn the new task in $\mathcal{D}^t$. At the end of current training stage t, we can flexibly integrate the convolutions in two branches, i.e., $\{W_t^r, b_t^r\}$ and $\{W_t^p, b_t^p\}$ to $\{W_{t+1}^r = \frac{W_t^r+W_t^p}{2}, b_{t+1}^r = \frac{b_t^r+b_t^p}{2}\}$ with the model re-parameterization [4]. In fact, the dual-flow model can be regarded as an implicit ensemble scheme [9] to integrate multiple sub-modules with a different focus. In addition, as demonstrated in [6], the fixed modules will regularize the learnable modules to act as the fixed one. Thus, the plasticity modules can also be implicitly encouraged to keep the previous knowledge along with its HSI learning.

However, under the domain shift, it can be sub-optimal to directly average the parameters, since $f_{\theta^t}^r$ may not perform well to predict $\mathcal{Y}^{t-1}$ on $\mathcal{X}^t$. It has been demonstrated that batch statistics adaptation plays an important role in domain generalizable model training [22]. Therefore, we propose a continual batch renormalization (cBRN) to mitigate the feature statistics divergence between each training batch at a specific stage and the life-long global data distribution.

Of note, as a default block in the modern convolutional neural networks (CNN) [8,37], batch normalization (BN) [11] normalizes the input feature of each CNN channel $z \in \mathbb{R}^{H_c\times W_c}$ with its batch-wise statistics, e.g., mean μ_B and standard deviation σ_B, and learnable scaling and shifting factors $\{\gamma, \beta\}$ as $\tilde{z}_i = \frac{z_i-\mu_B}{\sigma_B}\cdot\gamma+\beta$, where i indexes the spatial position in $\mathbb{R}^{H_c\times W_c}$. BN assumes that the same mini-batch training and testing distribution [10], which does not

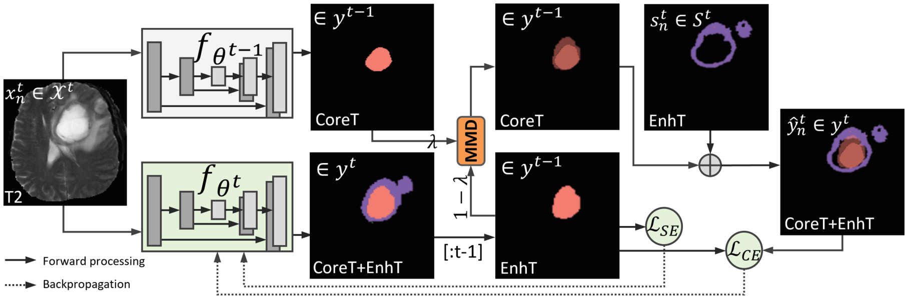

Fig. 2. Illustration of the proposed HSI pseudo-label distillation with MMD

hold in HSI. Simply enforcing the same statistics across domains as [5,30,33] can weaken the model expressiveness [36].

The recent BRN [10] proposes to rectify the data shift between each batch and the dataset by using the moving average μ and σ along with the training:

$$\mu = (1-\eta)\cdot\mu + \eta\cdot\mu_B, \quad \sigma = (1-\eta)\cdot\sigma + \eta\cdot\sigma_B, \tag{1}$$

where $\eta \in [0,1]$ is applied to balance the global statistics and the current batch. In addition, $\gamma = \frac{\sigma_B}{\sigma}$ and $\beta = \frac{\mu_B - \mu}{\sigma}$ are used in both training and testing. Therefore, BRN renormalizes $\tilde{z}_i = \frac{z_i - \mu}{\sigma}$ to highlight the dependency on the global statistics $\{\mu, \sigma\}$ in training for a more generalizable model, while limited to the static learning.

In this work, we further explore the potential of BRN in the continuously evolving HSI task to be general for all of domains involved. Specifically, we extend BRN to cBRN across multiple consecutive stages by updating $\{\mu_c, \sigma_c\}$ along with all stages of training, which is transferred as shown in Fig. 1. The conventional BN also inherits $\{\mu, \sigma\}$ for testing, while not being used in training [11]. At the stage t, μ_c and σ_c are succeeded from $t-1$ stage, and are updated with the current batch-wise $\{\mu_B^r, \sigma_B^r\}$ and $\{\mu_B^p, \sigma_B^p\}$ in rigidity and plasticity branches:

$$\mu_c = (1-\eta)\cdot\mu_c + \eta\cdot\frac{1}{2}\{\mu_B^r + \mu_B^p\}, \quad \sigma_c = (1-\eta)\cdot\sigma_c + \eta\cdot\frac{1}{2}\{\sigma_B^r + \sigma_B^p\}. \tag{2}$$

For testing, the two branches in final model f_{θ^T} can be merged for the lightweight implementation:

$$\tilde{z} = \frac{W_T^r z + b_T^r + \mu_c}{2\sigma_c} + \frac{W_T^p z + b_T^p + \mu_c}{2\sigma_c} = \frac{W_T^r + W_T^p}{2\sigma_c} z + \frac{b_T^r + b_T^p + 2\mu_c}{2\sigma_c} = \hat{W}z + \hat{b}. \tag{3}$$

Therefore, f_θ^T does not introduce additional parameters for deployment (Fig. 2).

2.2 HSI Pseudo-label Distillation with Momentum MixUp Decay

The training of our developed f_{θ^t} with D^{3}F is supervised with the previous model $f_{\theta^{t-1}}$ and current stage data $\{x_n^t, s_n^t\}_{n=1}^{N^t}$. In conventional class incremental learning, the knowledge distillation [31] is widely used to construct the combined label $y_n^t \in \mathbb{R}^{H\times W}$ by adding s_n^t and the prediction of $f_{\theta^{t-1}}(x_n^t)$. Then, f_{θ^t} can be optimized by the training pairs of $\{x_n^t, y_n^t\}_{n=1}^{N^t}$. However, with heterogeneous data in different stages, $f_{\theta^{t-1}}(x_n^t)$ can be highly unreliable. Simply using it as ground truth cannot guide the correct knowledge transfer.

In this work, we construct a complementary pseudo-label $\hat{y}_n^t \in \mathbb{R}^{H\times W}$ with a MixUp decay scheme to adaptively exploit the knowledge in the old segmentor for the progressively learned new segmentor. In the initial training epochs, $f_{\theta^{t-1}}$ could be a more reliable supervision signal, while we would expect f_{θ^t} can learn to perform better on predicting $\mathcal{Y}^{t-1}$. Of note, even with the rigidity branch, the integrated network can be largely distracted by the plasticity branch in the initial epochs. Therefore, we propose to dynamically adjust their importance in constructing pseudo-label along with the training progress. Specifically, we MixUp the predictions of $f_{\theta^{t-1}}$ and f_{θ^t} w.r.t. $\mathcal{Y}^{t-1}$, i.e., $f_{\theta^t}(\cdot)[: t-1]$, and control their pixel-wise proportion for the pseudo-label $\hat{y}_n^t$ with MMD:

$$\hat{y}_{n:i}^t = \{\lambda f_{\theta^{t-1}}(x_{n:i}^t) + (1-\lambda) f_{\theta^t}(x_{n:i}^t)[: t-1]\} \cup s_{n:i}^t, \;\; \lambda = \lambda^0 \exp(-I), \tag{4}$$

where i indexes each pixel, and λ is the adaptation momentum factor with the exponential decay of iteration I. λ^0 is the initial weight of $f_{\theta^{t-1}}(x_{n:i}^t)$, which is empirically set to 1 to constrain $\lambda \in (0, 1]$. Therefore, the weight of old model prediction can be smoothly decreased along with the training, and $f_{\theta^t}(x_{n:i}^t)$ gradually represents the target data for the old classes in $[: t-1]$. Of note, we have ground-truth of new structure $s_{n:i}^t$ under HSI scenarios [5,18,30,33]. We calculate the cross-entropy loss $\mathcal{L}_{CE}$ with the pseudo-label $\hat{y}_{n:i}^t$ as self-training [15,38].

In addition to the old knowledge inherited in $f_{\theta^{t-1}}$, we propose to explore unsupervised learning protocols to stabilize the initial training. We adopt the widely used self-entropy (SE) minimization [7] as a simple add-on training objective. Specifically, we have the slice-level segmentation SE, which is the averaged entropy of the pixel-wise softmax prediction as $\mathcal{L}_{SE} = \mathbb{E}_i\{-f_{\theta^t}(x_{n:i}^t)\log f_{\theta^t}(x_{n:i}^t)\}$. In training, the overall optimization loss is formulated as follows:

$$\mathcal{L} = \mathcal{L}_{CE}(\hat{y}_{n:i}^t, f_{\theta^t}(x_{n:i}^t)) + \alpha \mathcal{L}_{SE}(f_{\theta^t}(x_{n:i}^t)), \quad \alpha = \frac{I_{max} - I}{I_{max}} \alpha^0, \tag{5}$$

where α is used to balance our HSI distillation and SE minimization terms, and I_{max} is the scheduled iteration. Of note, strictly minimizing the SE can result in a trivial solution of always predicting a one-hot distribution [7], and a linear decreasing of α is usually applied, where λ^0 and α^0 are reset in each stage.

3 Experiments and Results

We carried out two evaluation settings using the BraTS2018 database [1], including cross-subset (relatively small domain shift) and cross-modality (relatively large domain shift) tasks. The BraTS2018 database is a continually evolving database [1] with a total of 285 glioblastoma or low-grade gliomas subjects,

Table 1. Numerical comparisons and ablation studies of the cross-subset brain tumor HSI segmentation task

Method	Data shift consideration	Dice similarity coefficient (DSC) [%] ↑				Hausdorff distance (HD)[mm] ↓			
		Mean	CoreT	EnhT	ED	Mean	CoreT	EnhT	ED
PLOP [5]	×	59.83 ± 0.131	45.50	57.39	76.59	19.2 ± 0.14	22.0	19.8	15.9
MargExcIL [18]	×	60.49 ± 0.127	48.37	56.28	76.81	18.9 ± 0.11	21.4	19.8	15.5
UCD [30]	×	61.84 ± 0.129	49.23	58.81	77.48	19.0 ± 0.15	21.8	19.4	15.7
HSI-MMD	√	66.87 ± 0.126	59.42	61.26	79.93	16.8 ± 0.13	18.5	17.8	14.2
HSI-D^{3}F	√	67.18 ± 0.118	60.18	63.09	78.26	16.7 ± 0.14	18.0	17.5	14.5
HSI-cBRN	√	68.07 ± 0.121	61.52	63.45	79.25	16.3 ± 0.14	17.8	17.3	13.8
HSI	√	**69.44 ± 0.119**	**63.79**	**64.71**	**79.81**	**15.7 ± 0.12**	**16.7**	**16.9**	**13.6**
Joint Static	√(upper bound)	73.98 ± 0.117	71.14	68.35	82.46	15.0 ± 0.13	15.7	16.2	13.2

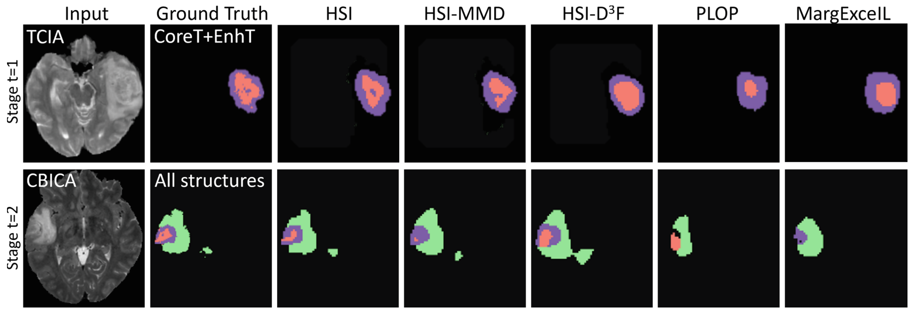

Fig. 3. Segmentation examples in $t = 1$ and $t = 2$ in the cross-subset brain tumor HSI segmentation task.

comprising three consecutive subsets, i.e., 30 subjects from BraTS2013 [26], 167 subjects from TCIA [3], and 88 subjects from CBICA [1]. Notably, these three subsets were collected from different clinical sites, vendors, or populations [1]. Each subject has T1, T1ce, T2, and FLAIR MRI volumes with voxel-wise labels for the tumor core (CoreT), the enhancing tumor (EnhT), and the edema (ED).

We incrementally learned CoreT, EnhT, and ED structures throughout three consecutive stages, each following different data distributions. We used subject-independent 7/1/2 split for training, validation, and testing. For a fair comparison, we adopted the ResNet-based 2D nnU-Net backbone with BN as in [12] for all of the methods and all stages used in this work.

3.1 Cross-Subset Structure Incremental Evolving

In our cross-subset setting, three structures were sequentially learned across three stages: (CoreT with BraTS2013) → (EnhT with TCIA) → (ED with CBICA). Of note, we used a CoreT segmentator trained with BraTS2013 as our off-the-shelf segmentor in $t = 0$. Testing involved all subsets and anatomical structures. We compared our framework with the three typical structure-incremental (SI-only) segmentation methods, e.g., PLOP [5], MargExcIL [18],

Table 2. Numerical comparisons and ablation studies of the cross-modality brain tumor HSI segmentation task

Method	Data shift consideration	Dice similarity coefficient (DSC) [%] ↑				Hausdorff distance (HD)[mm] ↓			
		Mean	CoreT	EnhT	ED	Mean	CoreT	EnhT	ED
PLOP [5]	×	39.58 ± 0.231	13.84	38.93	65.98	30.7 ± 0.26	48.1	25.4	18.7
MargExcIL [18]	×	42.84 ± 0.189	19.56	41.56	67.40	29.1 ± 0.28	46.7	22.1	18.6
UCD [30]	×	44.67 ± 0.214	21.39	45.28	67.35	29.4 ± 0.32	46.2	23.6	18.4
HSI-MMD	√	59.81 ± 0.207	51.63	53.82	73.97	19.4 ± 0.26	21.6	20.5	16.2
HSI-D^{3}F	√	60.81 ± 0.195	53.87	55.42	73.15	19.2 ± 0.21	21.4	19.9	16.2
HSI-cBRN	√	61.87 ± 0.180	54.90	56.62	74.08	18.5 ± 0.25	20.1	19.5	16.0
HSI	√	**64.15 ± 0.205**	**58.11**	**59.51**	**74.83**	**17.7 ± 0.29**	**18.9**	**18.6**	**15.8**
Joint Static	√ (upper bound)	70.64 ± 0.184	67.48	65.75	78.68	16.7 ± 0.26	17.2	17.8	15.1

and UCD [30], which cannot address the heterogeneous data across stages. As tabulated in Table 1, PLOP [5] with additional feature statistic constraints has lower performance than MargExcIL [18], since the feature statistic consistency was not held in HSI scenarios. Of note, the domain-incremental methods [17,32] cannot handle the changing output space. Our proposed HSI framework outperformed SI-only methods [5,18,30] with respect to both DSC and HD, by a large margin. For the anatomical structure CoreT learned in $t = 0$, the difference between our HSI and these SI-only methods was larger than 10% DSC, which indicates the data shift related forgetting lead to a more severe performance drop in the early stages. We set $\eta = 0.01$ and $\alpha^0 = 10$ according to the sensitivity study in the supplementary material.

For the ablation study, we denote HSI-D^{3}F as our HSI without the D^{3}F module, simply fine-tuning the model parameters. HSI-cBRN used dual-flow to avoid direct overwriting, while the model was not guided by cBRN for more generalized prediction on heterogeneous data. As shown in Table 1, both the dual-flow and cBRN improve the performance. Notably, the dual-flow model with flexible re-parameterization was able to alleviate the overwriting, while our cBRN was developed to deal with heterogeneous data. In addition, HSI-MMD indicates our HSI without the momentum MixUp decay in pseudo-label construction, i.e., simply regarding the prediction of $f_{\theta^{t-1}}(x^t)$ is ground truth for $\mathcal{Y}^{t-1}$. However, $f_{\theta^{t-1}}(x^t)$ can be quite noisy, due to the low quantification performance of early stage structures, which can be aggravated in the case of the long-term evolving scenario. Of note, the pseudo-label construction is necessary as in [5,18,30]. We also provide the qualitative comparison with SI-only methods and ablation studies in Fig. 3.

3.2 Cross-Modality Structure Incremental Evolving

In our cross-modality setting, three structures were sequentially learned across three stages: (CoreT with T1) → (EnhT with T2) → (ED with T2 FLAIR). Of note, we used the CoreT segmentator trained with T1 modality as our off-the-shelf segmentor in $t = 0$. Testing involved all MRI modalities and all structures. With the hyperparameter validation, we empirically set $\eta = 0.01$ and $\alpha^0 = 10$.

In Table 2, we provide quantitative evaluation results. We can see that our HSI framework outperformed SI-only methods [5,18,30] consistently. The improvement can be even larger, compared with the cross-subset task, since we have much more diverse input data in the cross-modality setting. Catastrophic forgetting can be severe, when we use SI-only method for predicting early stage structures, e.g., CoreT. We also provide the ablation study with respect to D^{3}F, cBRN, and MMD in Table 2. The inferior performance of HSI-D^{3}F/cBRN/MMD demonstrates the effectiveness of these modules for mitigating domain shifts.

4 Conclusion

This work proposed an HSI framework under a clinically meaningful scenario, in which clinical databases are sequentially constructed from different sites/imaging protocols with new labels. To alleviate the catastrophic forgetting alongside continuously varying structures and data shifts, our HSI resorted to a D^{3}F module for learning and integrating old and new knowledge nimbly. In doing so, we were able to achieve divergence awareness with our cBRN-guided model adaptation for all the data involved. Our framework was optimized with a self-entropy regularized HSI pseudo-label distillation scheme with MMD to efficiently utilize the previous model in different types of MRI data. Our framework demonstrated superior segmentation performance in learning new anatomical structures from cross-subset/modality MRI data. It was experimentally shown that a large improvement in learning anatomic structures was observed.

Acknowledgements. This work is supported by NIH R01DC018511, R01DE027989, and P41EB022544. The authors would like to thank Dr. Jonghyun Choi for his valuable insights and helpful discussions.

References

1. Bakas, S., et al.: Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the brats challenge. arXiv:1811.02629 (2018)
2. Chang, W.G., You, T., Seo, S., Kwak, S., Han, B.: Domain-specific batch normalization for unsupervised domain adaptation. In: CVPR, pp. 7354–7362 (2019)
3. Clark, K., et al.: The cancer imaging archive (TCIA): maintaining and operating a public information repository. J. Digit. Imaging **26**(6), 1045–1057 (2013)
4. Ding, X., Zhang, X., Ma, N., Han, J., Ding, G., Sun, J.: RepVGG: making VGG-style convnets great again. In: CVPR, pp. 13733–13742 (2021)
5. Douillard, A., Chen, Y., Dapogny, A., Cord, M.: PLOP: learning without forgetting for continual semantic segmentation. In: CVPR, pp. 4040–4050 (2021)
6. Fu, S., Li, Z., Liu, Z., Yang, X.: Interactive knowledge distillation for image classification. Neurocomputing **449**, 411–421 (2021)

7. Grandvalet, Y., Bengio, Y.: Semi-supervised learning by entropy minimization. In: NeurIPS (2005)
8. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
9. Huang, G., Sun, Yu., Liu, Z., Sedra, D., Weinberger, K.Q.: Deep networks with stochastic depth. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9908, pp. 646–661. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46493-0_39
10. Ioffe, S.: Batch renormalization: towards reducing minibatch dependence in batch-normalized models. In: NeurIPS, vol. 30 (2017)
11. Ioffe, S., Szegedy, C.: Batch normalization: accelerating deep network training by reducing internal covariate shift. In: ICML, pp. 448–456. PMLR (2015)
12. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nat. Methods **18**(2), 203–211 (2021)
13. Kanakis, M., Bruggemann, D., Saha, S., Georgoulis, S., Obukhov, A., Van Gool, L.: Reparameterizing convolutions for incremental multi-task learning without task interference. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12365, pp. 689–707. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58565-5_41
14. Kim, D., Bae, J., Jo, Y., Choi, J.: Incremental learning with maximum entropy regularization: Rethinking forgetting and intransigence. arXiv:1902.00829 (2019)
15. Kim, K., Ji, B., Yoon, D., Hwang, S.: Self-knowledge distillation: a simple way for better generalization. arXiv:2006.12000 (2020)
16. Kundu, J.N., Venkatesh, R.M., Venkat, N., Revanur, A., Babu, R.V.: Class-incremental domain adaptation. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12358, pp. 53–69. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58601-0_4
17. Li, K., Yu, L., Heng, P.A.: Domain-incremental cardiac image segmentation with style-oriented replay and domain-sensitive feature whitening. TMI **42**(3), 570–581 (2022)
18. Liu, P., et al.: Learning incrementally to segment multiple organs in a CT image. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) MICCAI 2022. LNCS, vol. 13434, pp. 714–724. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16440-8_68
19. Liu, X., et al.: Attentive continuous generative self-training for unsupervised domain adaptive medical image translation. Med. Image Anal. (2023)
20. Liu, X., Xing, F., El Fakhri, G., Woo, J.: Memory consistent unsupervised off-the-shelf model adaptation for source-relaxed medical image segmentation. Med. Image Anal. **83**, 102641 (2023)
21. Liu, X., et al.: Act: Semi-supervised domain-adaptive medical image segmentation with asymmetric co-training. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) MICCAI 2022. LNCS, vol. 13435, pp. 66–76. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16443-9_7

22. Liu, X., Xing, F., Yang, C., El Fakhri, G., Woo, J.: Adapting off-the-shelf source segmenter for target medical image segmentation. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12902, pp. 549–559. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87196-3_51
23. Liu, X., et al.: Subtype-aware dynamic unsupervised domain adaptation. IEEE TNNLS (2022)
24. Liu, X., et al.: Deep unsupervised domain adaptation: a review of recent advances and perspectives. APSIPA Trans. Signal Inf. Process. **11**(1) (2022)
25. Liu, Y., Schiele, B., Sun, Q.: Adaptive aggregation networks for class-incremental learning. In: CVPR, pp. 2544–2553 (2021)
26. Menze, B.H., et al.: The multimodal brain tumor image segmentation benchmark (BRATS). TMI **34**(10), 1993–2024 (2014)
27. Ozdemir, F., Fuernstahl, P., Goksel, O.: Learn the new, keep the old: extending pretrained models with new anatomy and images. In: Frangi, A.F., Schnabel, J.A., Davatzikos, C., Alberola-López, C., Fichtinger, G. (eds.) MICCAI 2018. LNCS, vol. 11073, pp. 361–369. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00937-3_42
28. Shusharina, N., Söderberg, J., Edmunds, D., Löfman, F., Shih, H., Bortfeld, T.: Automated delineation of the clinical target volume using anatomically constrained 3D expansion of the gross tumor volume. Radiot. Oncol. **146**, 37–43 (2020)
29. Tajbakhsh, N., Jeyaseelan, L., Li, Q., Chiang, J.N., Wu, Z., Ding, X.: Embracing imperfect datasets: a review of deep learning solutions for medical image segmentation. Med. Image Anal. **63**, 101693 (2020)
30. Yang, G., et al.: Uncertainty-aware contrastive distillation for incremental semantic segmentation. IEEE Trans. Pattern Anal. Mach. Intell. **45**(2), 2567–2581 (2022)
31. Yin, H., et al.: Dreaming to distill: Data-free knowledge transfer via deepinversion. In: CVPR, pp. 8715–8724 (2020)
32. You, C., et al.: Incremental learning meets transfer learning: application to multi-site prostate MRI segmentation. arXiv:2206.01369 (2022)
33. Yu, L., Liu, X., Van de Weijer, J.: Self-training for class-incremental semantic segmentation. IEEE Trans. Neural Netw. Learn. Syst. (2022)
34. Zhang, C.B., Xiao, J.W., Liu, X., Chen, Y.C., Cheng, M.M.: Representation compensation networks for continual semantic segmentation. In: CVPR (2022)
35. Zhang, H., Zhang, Y., Jia, K., Zhang, L.: Unsupervised domain adaptation of black-box source models. arXiv:2101.02839 (2021)
36. Zhang, J., Qi, L., Shi, Y., Gao, Y.: Generalizable semantic segmentation via model-agnostic learning and target-specific normalization. arXiv:2003.12296 (2020)
37. Zhou, X.Y., Yang, G.Z.: Normalization in training u-net for 2-D biomedical semantic segmentation. IEEE Robot. Autom. Lett. **4**(2), 1792–1799 (2019)
38. Zou, Y., Yu, Z., Liu, X., Kumar, B., Wang, J.: Confidence regularized self-training. In: ICCV, pp. 5982–5991 (2019)