

## SURVEY

# An Experimental Survey of Incremental Transfer Learning for Multicenter Collaboration

YIXING HUANG<sup>1,2,3</sup>, CHRISTOPH BERT<sup>1,2,3</sup>, AHMED GOMAA<sup>1,2,3</sup>, RAINER FIETKAU<sup>1,2,3</sup>, ANDREAS MAIER<sup>4</sup>, (Senior Member, IEEE), AND FLORIAN PUTZ<sup>1,2,3</sup>

<sup>1</sup>Department of Radiation Oncology, Universitätsklinikum Erlangen, Friedrich-Alexander-Universität Erlangen-Nürnberg, 91054 Erlangen, Germany

<sup>2</sup>Comprehensive Cancer Center Erlangen-EMN (CCC ER-EMN), 91054 Erlangen, Germany

<sup>3</sup>Bavarian Cancer Research Center (BZKF), 91052 Erlangen, Germany

<sup>4</sup>Pattern Recognition Lab, Friedrich-Alexander-Universität Erlangen-Nürnberg, 91054 Erlangen, Germany

Corresponding author: Yixing Huang (yixing.yh.huang@fau.de)

**ABSTRACT** Due to data privacy constraints, data sharing among multiple clinical centers is restricted, which impedes the development of high performance deep learning models from multicenter collaboration. Naive weight transfer methods share intermediate model weights without raw data and hence can bypass data privacy restrictions. However, performance drops are typically observed when the model is transferred from one center to the next because of the forgetting problem. Incremental transfer learning, which combines peer-to-peer federated learning and domain incremental learning, can overcome the data privacy issue and meanwhile preserve model performance by using continual learning techniques. In this work, a conventional domain/task incremental learning framework is adapted for incremental transfer learning. A survey on the efficacy of prevalent regularization-based continual learning methods for multicenter collaboration is performed. The influences of data heterogeneity, classifier head setting, network optimizer, model initialization, center order, and weight transfer type have been investigated thoroughly. Our framework is publicly accessible to the research community for further development.

**INDEX TERMS** Continual learning, multicenter collaboration, federated learning, data privacy, deep learning, peer-to-peer, domain incremental learning, incremental transfer learning.

## I. INTRODUCTION

Deep learning has played important roles in various medical applications nowadays, including image registration [1], reconstruction [2], and segmentation [3], [4] as well as disease diagnosis and treatment planning [5]. The performance of deep learning algorithms relies highly on the amount and quality of training data. Many research centers have the required technological and human resources as well as computation power. However, the limited access to data becomes an obstacle for them to develop deep learning algorithms independently. Therefore, collaboration among different centers (including research centers and hospitals) is always important. Due to data privacy and data management regulations, e.g., the EU medical device regulation [6] and the EU General Data Protection Regulation

for medical imaging data, data sharing among multiple centers is restricted, which impedes the development of high performance deep learning models from multicenter collaboration.

To overcome the data privacy issue, federated learning has been proposed [7], [8], [9], [10], which enables multiple centers to train a high performance model without sharing data. In center-to-peer federated learning (C2PFL) (Fig. 1(a)), a central server is required to coordinate training information for a global model. However, such a central server is financially expensive to build and the communication cost between the central server and multiple clients is expensive as well [9]. In IT industry, because of the extremely large number of mobile clients and the limited computational ability of mobile clients, C2PFL is preferred over peer-to-peer federated learning (P2PFL). However, for multicenter collaboration in medical fields, the number of centers is typically low and each center has sufficient computation resources.

The associate editor coordinating the review of this manuscript and approving it for publication was Ramakrishnan Srinivasan<sup>1</sup>.

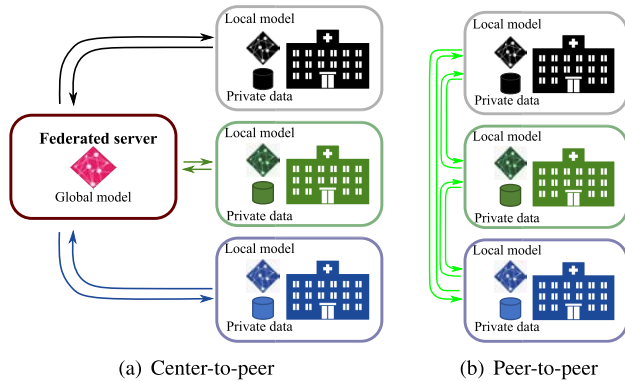


FIGURE 1. The center-to-peer (a) and peer-to-peer (b) federated learning.

Therefore, P2PFL in a decentralised mode (Fig. 1(b)) is more feasible in practice [9], [11], since a central server is no longer required. The simplest way for such P2PFL is to continually train the same model one center after another via weight transfer (either **single weight transfer (SWT)** or **cyclic weight transfer (CWT)**) [12], [13], [14].

However, in P2PFL with naive weight transfer, performance fluctuation from center to center is observed [12], [13], [14]. When a model is retrained on new datasets or tasks, deep learning suffers from the problem of catastrophic forgetting [15], [16], [17], i.e., deep learning models forget learned old knowledge catastrophically. To address the forgetting problem, continual learning has been proposed. According to application scenarios, three types of continual learning are investigated [18]: task-incremental learning [17], class-incremental learning [19], and domain-incremental learning [18]. In the task-incremental learning, a model incrementally learns a set of distinct tasks, while in the class-incremental learning a model incrementally discriminates between a growing number of classes. Domain-incremental learning addresses the same task with the same number of classes, but the distribution of training/test data varies, e.g., cardiac image segmentation on heterogeneous data with domain discrepancy [20], where forgetting caused by domain shift needs to be addressed. Domain incremental learning is very similar to P2PFL for multicenter collaboration but with differences, e.g., the necessity of domain discrepancy.

So far, many continual learning algorithms have been proposed, which are mainly categorized to three categories [17]: replay methods [21], [22], [23], architectural (parameter isolation) methods [24], [25], [26], [27], [28], [29], and regularization methods [30], [31], [32]. Replay methods select representative samples from previous datasets to preserve learned knowledge. This is feasible to overcome data storage constraints, but not feasible for multicenter collaboration since samples from other centers are not available due to data privacy. In this case, using synthetic samples from generative adversarial networks (GANs) is an alternative choice [33]. Architectural methods design dynamic network architectures or dynamic parameters for

multi-task scenarios, where certain parts of the network (e.g., certain modules [28], [29], weights [24], [25], [26], or neuron connections [27]) are responsible for certain tasks. Regularization methods use the same conventional neural networks, but with new regularization terms during training to preserve important parameters for learned knowledge, like **learning without forgetting (LWF)** [30], **elastic weight consolidation (EWC)** [31] and **synaptic intelligence (SI)** [32].

To date, comprehensive surveys for task-incremental learning [17] and class-incremental learning [19] have been conducted. However, domain-incremental learning is the least studied continual learning scenario, as stated in [18], and its application to overcome data privacy in multicenter collaboration for medical applications has not been thoroughly investigated yet. The contributions of this work mainly lie in the following aspects:

- An in-depth experimental survey on the efficacy of prevalent regularization-based continual learning methods for multicenter collaboration is performed, which can be sister surveys of [17] and [19];
- The influences of data heterogeneity, classifier head setting, network optimizer, model initialization, center order, and weight transfer type have been investigated thoroughly;
- A framework to combine P2PFL and domain-incremental learning is presented, which is publicly accessible<sup>1</sup> to the research community for further development.

## II. RELATED WORK

This work includes the topics of transfer learning, federated learning, and continual learning. Related surveys on transfer learning and federated learning include [34], [35], [36], [37], but they do not address the forgetting problem. As this work focuses on continual learning methods, the most related surveys are task-incremental learning [17], class-incremental learning [19], and online continual learning [38], where the task or class changes in [17], [19], and [38] does not consider data privacy issues, whereas the same task/class with data privacy issues is considered in this work for multicenter collaboration. The concept presented in this work is similar to P2PFL with naive weight transfer, federated continual learning, and domain incremental learning, but with differences.

P2PFL with weight transfer [12], [13], [14] has been proposed for multicenter collaboration on medical applications. Chang et al. [12] proposed distributing deep learning models via SWT or CWT for multi-institutional collaboration and the CWT method was shown to achieve comparable performance to that of centrally hosted patient data. Sheller et al. [13], [14] have applied the two weight transfer methods to glioma segmentation and confirmed that CWT is superior to SWT. In their study, the performance fluctuation is observed which is caused by forgetting [15] when the model is transferred

<sup>1</sup><https://github.com/YixingHuang/ITLsurvey>

from one institution to another. However, the forgetting problem has not been addressed.

Recently, federated continual learning [39], [40], [41], [42] becomes an emerging concept. Nevertheless, these contributions were all proposed for C2PFL with different focuses, and hence are different from our peer-to-peer scenarios. Yao and Sun [39] proposed to estimate importance weights for the global model based on EWC and use such importance weights to constrain local training in clients to address the data heterogeneity problem. Yoon et al. [40] proposed to decompose model parameters into global parameters, local base parameters and task-adaptive parameters. Hence the global server can aggregate and redistribute information from different clients selectively, where continual learning is applied in local clients to reduce the communication cost. Usmanova et al. [41] proposed a federated learning without forgetting (FLwF) method for class-incremental scenarios [19], which uses the central server as the second teacher to guide a student client model to deal with catastrophic forgetting, along with the first teacher trained locally in each client. Casado et al. [42] addressed the concept drift problem in federated learning, where data feature distribution varies over time at each client. A simple rehearsal method which stores old memories at each local client is proposed for drift adaptation before global aggregation.

Domain adaptation is a common technique to address the performance drop in transfer learning [34]. However, typical domain adaptation techniques [43] aim to improve the model performance in the target domain while the performance on the source domain is no longer important. However, in multicenter collaboration, a global model with good performance on the test datasets from all centers is pursued. Domain incremental learning [18] aims to address the domain discrepancy problem and preserve the model performance on data from different domains using continual learning techniques. Its sequential training scheme fits well to P2PFL. Many domain incremental learning algorithms have been integrated into medical applications [20], [44], [45], [46], [47], [48], [49]. For example, Srivastava et al. proposed a vector quantization memory buffer to efficiently store and replay hidden representations of data from different domains [48]. Li et al. proposed a domain incremental method with style-oriented replay and domain sensitive feature whitening for cardiac image segmentation [20]. But the domain identity information is unknown for the training and test data in their setting, which is different for P2PFL in multicenter collaboration. Thakur et al. [49] proposed a replay-based self-aware stochastic gradient descent (SGD) method for domain shifts over time in a fixed center instead of multicenters. You et al. used an encoder to learn domain discrepancy between the current dataset and old exemplars in an **incremental transfer learning (ITL)** setting [45], but the old exemplars typically are challenging to obtain in multicenter collaboration because of data privacy regulations. In our previous preliminary work [47], SI was applied to address the forgetting problem in multicenter collaboration

for brain metastases segmentation, where many questions remain unresolved, which lead to this in-depth survey.

### III. PROBLEM AND LEARNING FRAMEWORK

#### A. PROBLEM DEFINITION

Conventional continual learning topics typically focus on addressing the task-incremental learning [17], and class-incremental learning [19] problems. P2PFL for multicenter collaboration is very similar to domain incremental learning [18] but without the necessity of domain discrepancy. For multicenter collaboration in this work, learning for a single task with a fixed number of classes but incremental data from multiple centers is addressed. During the training and test stages, the training and test data is located locally in each center and hence the center identity information is known. The deep learning model iteratively learns from a sequence of datasets  $\{\mathcal{C}^{(1)}, \mathcal{C}^{(2)}, \dots, \mathcal{C}^{(N)}\}$  where  $\mathcal{C}^{(\mu)}$  is a labeled dataset from the  $\mu$ -th center,  $\mathcal{C}^{(\mu)} = \{\mathbf{x}_i^{(\mu)}, \mathbf{y}_i^{(\mu)}\}_{i=1}^{N_\mu}$ , which consists of  $N_\mu$  pairs of instances  $\mathbf{x}_i^{(\mu)}$  and their corresponding labels  $\mathbf{y}_i^{(\mu)}$ . Considering data privacy constraints, center  $\mu$  has the access to its own data  $\mathcal{C}^{(\mu)}$  for training a shared model and the datasets from other centers are all inaccessible. Given  $\mathcal{C}^{(\mu)}$  and a model trained in a previous center, the primary learning objective at the  $\mu$ -th center is to minimize the task-specific loss  $\mathcal{L}_{\text{task}}$ ,

$$\arg \min_{\theta^{(\mu)}} \mathcal{L}_{\text{task}} \left( \theta^{(\mu)}; \theta^{(\mu-1)}, \mathcal{C}^{(\mu)} \right), \quad (1)$$

where  $\theta^{(\mu)}$  is the parameter set of the model at the  $\mu$ -th center, while the model performance on test data in previous centers is desired to be preserved. Due to the incremental weight transfer nature of our problem setting, we call it **incremental transfer learning (ITL)** named after [45], but without the use of old exemplars because of data privacy.

The problem definition of our ITL is very similar to domain-incremental learning [18] but with the following differences:

- In domain-incremental learning datasets must be from different domains, while in multicenter collaboration, datasets from different centers can be from the same domain, because the data can be collected from the same medical imaging devices with the same preprocessing pipeline.
- Mixing representative sample data together with new training data is possible for general domain-incremental learning, while it is impossible in ITL for multicenter collaboration because of data privacy. However, the model can revisit each center via CWT for training and test in ITL.
- Domain-incremental learning typically aims at domain-specific (center-specific in this work) output instead of a global output (Fig. 2 in [18]). Therefore, multi-head setting is typically used for optimal output at each domain/center. However, for multicenter

collaboration, a global model exhibiting generalizability across datasets from all the centers is pursued.

- For domain incremental learning, the source information (domain/center) of the training or test data might be unknown [20], while it is clearly known for multicenter collaboration.

The problem definition of our ITL is different from transfer learning [34], [35] as well:

- The focus of transfer learning [34], [35] is on optimizing the model's performance on the target center, often without consideration for the performance on the original source center. This approach is effective in scenarios where the target center has limited data but can benefit from knowledge gained from a larger, related dataset.
- ITL extends the principles of transfer learning by incorporating continual learning techniques to address the forgetting problem. In this framework, the model is iteratively updated with new data from multiple centers, ensuring that it retains its performance across all centers, including both source and target centers.

## B. REGULARIZATION METHODS FOR CONTINUAL LEARNING

As direct optimization based on the primary task-specific loss (Eqn. (1)) alone has the risk of forgetting, continual learning techniques are applied. For continual learning, architectural methods are not our choice since they are mainly designed for multi-task scenarios instead of the single-task scenario in multicenter collaboration. Replay methods require samples from other centers, which is typically not easily feasible due to data privacy constraints. Generative replay methods using synthetic samples or extracted representative features is a promising approach. Because of the above reasons, regularization-based continual learning methods naturally fit well in multicenter collaboration scenarios. Due to the limited space, it is challenging to cover all the regularization methods and hence the following prevalent regularization methods are investigated: LWF [30], EWC [31], SI [32], **memory aware synapse (MAS)** [50], **encoder based lifelong learning (EBLL)** [51], and **incremental moment matching (IMM)** [52]. The investigation on such methods can already provide important hints for further research in this field. For these regularization methods, the overall loss function  $L_\mu$  for center  $\mu$  can be represented as follows in general,

$$\mathcal{L}^{(\mu)} = \mathcal{L}_{\text{task}} + \lambda^{(\mu)} \phi(\boldsymbol{\theta}^{(\mu-1)}, \boldsymbol{\theta}), \quad (2)$$

where  $\boldsymbol{\theta}$  is the current parameter set to optimize during training, while  $\boldsymbol{\theta}^{\mu-1}$  is the parameter set transferred from the previous center,  $\mu$  is the index for the current center, and  $\lambda$  is a regularization parameter to control the amount of the regularization term  $\phi$ .

EWC [31], SI [32], and MAS [50] share the same idea of penalizing the change of important network parameters, but they have different ways to compute the importance

weights  $\boldsymbol{\Omega}^\mu$ : EWC computes the importance weights after model training using the Fisher information matrix; SI computes the importance weights during training by tracking the magnitudes of parameter changes; MAS computes the importance weights after training by measuring the change of the model output with respect to the parameter changes. The overall loss function  $L_\mu$  of the  $\mu$ -th center for these three methods can be represented as follows in general,

$$\mathcal{L}^{(\mu)} = \mathcal{L}_{\text{task}} + \lambda^{(\mu)} \sum_k \boldsymbol{\Omega}_k^{(\mu-1)} \left( \boldsymbol{\theta}_k^{(\mu-1)} - \boldsymbol{\theta}_k \right)^2, \quad (3)$$

where the right term penalizes the change of important network parameters trained in previous centers.  $\lambda$  is the relaxation parameter to trade off old versus new knowledge and  $k$  is the index for a certain parameter.

LWF [30] controls forgetting by imposing network output stability. In other words, for certain samples, the model after training with local data should get similar predictions to those before training. The objective function for LWF is as follows,

$$\begin{aligned} \mathcal{L}_{\text{LWF}}^{(\mu)} = & \mathcal{L}_{\text{task}} \\ & + \lambda^{(\mu)} \mathcal{L}_{\text{KDL}} \left( \mathcal{M}(\mathcal{C}^{(\mu)}, \boldsymbol{\theta}^{(\mu-1)}), \mathcal{M}(\mathcal{C}^{(\mu)}, \boldsymbol{\theta}) \right), \end{aligned} \quad (4)$$

where  $\mathcal{L}_{\text{KDL}}$  is the knowledge distillation loss (KDL) [53] and  $\mathcal{M}$  is the network model. In ITL, samples from the previous center typically is not available. Therefore, the training dataset of  $\mathcal{C}^{(\mu)}$  is used as the sample set for computing the KDL.

EBLL [51] extends LWF by using an autoencoder to capture important task-specific features in a low dimensional manifold. The main network  $\mathcal{M}$  is divided to two parts: feature extractor  $\mathcal{F}$  and task/center-specific classifier  $\mathcal{T}$ . An autoencoder is trained to project features extracted by  $\mathcal{F}$  onto a lower dimensional manifold and reconstruct them back for each task. With the autoencoder trained from the previous center, the following objective function is used for training the main network in the  $\mu$ -th center,

$$\begin{aligned} \mathcal{L}^{(\mu)} = & \mathcal{L}_{\text{LWF}}^{(\mu)} + \alpha/2 \|\mathcal{E}^{(\mu-1)}(\mathcal{F}(\mathbf{x}^{(\mu)}, \boldsymbol{\theta})) \\ & - \mathcal{E}^{(\mu-1)}(\mathcal{F}(\mathbf{x}^{(\mu)}, \boldsymbol{\theta}^{(\mu-1)}))\|_2^2 \end{aligned} \quad (5)$$

where  $\mathcal{E}^{(\mu-1)}$  is the encoder part of the autoencoder trained in the previous center and  $\alpha$  is a parameter for the additional constraint on the manifold space.

IMM [52] merges models trained from each center to obtain an improved model. Such model merging relies on a flat or convex loss surface, which is obtained by weight transfer, L2 transfer and drop transfer. The L2 transfer uses an  $\ell_2$  regularization to get a smooth loss surface between  $\boldsymbol{\theta}^{\mu-1}$  and  $\boldsymbol{\theta}^\mu$ ,

$$\mathcal{L}^{(\mu)} = \mathcal{L}_{\text{task}} + \beta \left( \boldsymbol{\theta}^{(\mu-1)} - \boldsymbol{\theta} \right)^2, \quad (6)$$

where  $\beta$  is a small value (e.g. 0.001). After training in all the centers, the models are merged as the following,

$$\hat{\theta}_k^{(\mu)} = \sum_{v=1}^{\mu} \alpha^{(v)} \theta_k^{(v)}, \quad (7)$$

where  $\alpha^v$  sums up to 1 for all centers. Two ways for model merging have been proposed in [52]: mean-IMM and mode-IMM. Mean-IMM minimizes the KL-divergence between a model trained from an individual task (center) and the model trained from all tasks (centers), and  $\alpha^v$  simply equals  $1/N$ , where  $N$  is the total number of centers. Mode-IMM merges model parameters using a Laplacian approximation to estimate modes (similar to local minima), and  $\alpha^v$  is determined by parameter diagonal covariance matrices. One way to compute  $\alpha^v$  is to use the inverse of a Fisher information matrix as an indicator of each parameter's importance  $\Omega_k^\mu$  like the EWC method [31],

$$\alpha^{(\mu)} = \Omega_k^{(\mu)} / \sum_{v=1}^{\mu} \Omega_k^{(v)}. \quad (8)$$

### C. INCREMENTAL TRANSFER LEARNING FRAMEWORK

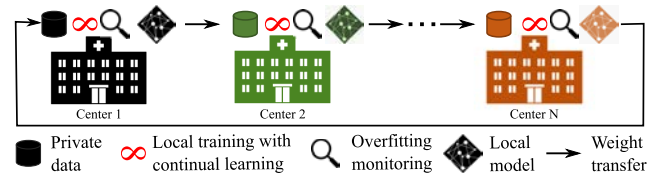
The multicenter collaboration scenario is different from the typical continual learning scenarios [18]. Hence, the direct application of existing continual learning frameworks is suboptimal for multicenter collaboration. Therefore, the following components are suggested for our ITL framework.

#### 1) SINGLE-HEAD INSTEAD OF MULTI-HEAD SETTING

For task-incremental learning and class-incremental learning, the multi-head setting imposes a separate classifier for each task to achieve best adaptability. Especially, because of the different number of classes, the old classifiers cannot be used for new tasks. For domain-incremental learning, although the same task is optimized, the multi-head setting is typically used for optimal output at each context/center [18]. However, when a trained classifier  $\mathcal{T}$  is replaced by a new classifier  $\mathcal{T}'$  for training in the current center, the model parameters in the feature extractor  $\mathcal{F}$  will also be updated to a new one  $\mathcal{F}'$ . **Although the combination of  $\mathcal{F}'$  and  $\mathcal{T}'$  is optimal for the current center, feature mismatch will occur between the updated feature extractor  $\mathcal{F}'$  and the classifier for the previous center  $\mathcal{T}$ .** Therefore, catastrophic forgetting occurs. To avoid such mismatch, the single-head setting is used in our proposed ITL framework so that the feature extractor and the classifier are always optimized jointly.

#### 2) RELOAD OPTIMIZERS

For task-incremental continual learning, maximum plasticity is achieved if the model is trained solely on the dataset for this certain task. Therefore, a learning rate grid search (LRGS) step [17], [19] is commonly used to achieve the maximum plasticity for new tasks and a new training with the obtained learning rate starts after weight transfer. However, for multicenter collaboration, due to the limited amount of



**FIGURE 2. The incremental transfer learning (ITL) framework adds continual learning techniques to reduce forgetting in naive transfer learning for multicenter collaboration. An initial model is trained in the first center. Afterwards the weights of a trained model and its optimizer at each center are transferred from Center 1 to Center N successively or cyclically for continual training.**

data at each center, training solely on the local data will not lead to maximum plasticity. Therefore, such a LRGS step is not necessary. In fact, a sudden change of learning rate will lead to performance fluctuations for other centers. Therefore, a smooth learning rate transition from center to center is preferred. For this purpose, a simple solution is simply reloading optimizers from the previous center for the training in the current center, while LRGS is used only in the first center.

#### 3) USE ADAPTIVE OPTIMIZERS

So far, all the above continual learning methods use SGD for optimization [17], [19], [30], [31], [32], [50], [51], [52]. SGD achieves good performance in tasks. However, the use of SGD is known to be difficult for initial training and needs a proper learning rate scheduler for model convergence. Adaptive optimizers like adaptive moment estimation (Adam) automatically adjust the learning weight for each model parameter. Hence, they will relieve the choice of learning rate schedulers. To integrate the regularization-based continual learning techniques, the adaptive optimizers need to be modified. An example of using Adam for SI is presented in Appendix VIII-B.

#### 4) OVERFITTING MONITOR

To avoid overfitting to local data, the learning rate is decayed if no performance gain is achieved after every  $E_{\text{val}}$  epochs at each center. Early stopping is used if no performance gain is achieved after  $E_{\text{stop}}$  epochs. The model with the best validation loss is shared to the next center. Such monitoring for overfitting reduces the overfitting problem, although overfitting typically still exists.

#### 5) CYCLIC WEIGHT TRANSFER

For continual learning, a common setting is that the datasets from previous tasks are no longer accessible. Therefore, SWT instead of CWT is applied. For ITL, datasets from different centers can be re-accessed via CWT and CWT has been reported to achieve better performance than SWT [12], [13], [14]. Therefore, CWT is integrated with continual learning techniques in our ITL framework, as displayed in Fig. 2.

With the above ITL framework, the efficacy of different regularization-based continual learning methods is investigated on different applications.

## IV. EXPERIMENTS ON TINY IMAGENET DATASET

### A. EXPERIMENTAL SETUP

#### 1) DATASET

As a proof-of-concept investigation, image classification experiments on the Tiny ImageNet dataset [54] in the multicenter collaboration scenario are conducted. The image size is  $64 \times 64$ . 10 classes among the 200 classes from the Tiny ImageNet dataset are chosen. Each class contains 500 samples subdivided into training (80%) and validation (20%), and 50 samples for test. These images are equally split into 5 centers, i.e., each center 80 images per class (in total 800 images) for training, 20 images per class (in total 200 images) for validation and 10 images per class (in total 100 images) for test. As the original Tiny ImageNet images are randomly split into each center, the data distribution among all the 5 centers can be assumed to be **independent and identically distributed (IID)**.

In practice, data from different centers might be heterogeneous. To simulate non-IID data, Gaussian noise with a zero mean and a standard deviation of 25 is added to the images in one center. The noise is first added to the fifth center and then the datasets from the third and fifth centers are swapped to investigate the influence of training order.

#### 2) NETWORK

The small VGG-9 network from [17] is used, which has 6 convolutional layers (first 4 layers with 64 filters and the remaining 2 layers with 128 filters), 4 max-pooling layers and 3 fully connected layers (128 classifier dimensions). The task-specific loss  $\mathcal{L}_{\text{task}}$  is the cross-entropy loss.

For SWT, all the models are trained for 50 epochs at each center. For CWT, a weight transfer frequency of  $E_{\text{transfer}} = 10$  is investigated with a total number of 5 iterations. For overfitting monitor, the learning rate is decayed into half after every  $E_{\text{val}} = 5$  epochs and early stopping is executed after  $E_{\text{stop}} = 20$  epochs. The joint model is trained from mixed data for 50 epochs. A constant  $\lambda$  value of 1 for LWF, EWC, SI and MAS is used. For IMM, the  $\lambda$  value for the L2-weight transfer is 0.01. The batch size is set to 100. The Adam optimizer as a representative of adaptive optimizers is investigated in comparison to the widely-used SGD optimizer in continual learning. For SGD, the initial learning rate is set to 0.1 after LRGS. An exponentially decaying learning rate scheduler is set for the SGD optimizer,  $r = 0.1 * 0.8^{e/5}$ , where  $e$  is the accumulated epoch number. For Adam, the learning rate is fixed to 0.001 after LRGS.

#### 3) EVALUATION METRICS

##### a: ACCURACY

For the test data in the  $\mu$ -th center, an accuracy sequence  $\mathbf{a}^{(\mu)} = a_1^{(\mu)}, a_2^{(\mu)}, \dots, a_{N_{\text{total}}}^{(\mu)}$  is computed for a sequence of trained models, where  $N_{\text{total}}$  is the total number of center-wise training. The average accuracy  $a_{\text{mean}}$  of the final model

evaluated in all the centers is reported for each method,

$$a_{\text{mean}} = \sum_{\mu=1}^N a_{N_{\text{total}}}^{(\mu)} / N. \quad (9)$$

With the single-head setting, forgetting is no longer catastrophic. Therefore, the forgetting metric in [17] is not reported in this work.

##### b: MONOTONICITY

As a monotonic increase of accuracy is desired, a monotonicity sequence is computed as well,  $\mathbf{m}^{(\mu)} = m_2^{(\mu)}, m_3^{(\mu)}, \dots, m_{N_{\text{total}}}^{(\mu)}$ , where  $m_i^{(\mu)} = 1$  if  $a_i^{(\mu)} \geq a_{i-1}^{(\mu)}$  else 0. For each method, the average monotonicity of all the models is,

$$m_{\text{mean}} = \sum_{\mu=1}^N \sum_{i=2}^{N_{\text{total}}} m_i^{(\mu)} / (N \cdot (N_{\text{total}} - 1)), \quad (10)$$

where  $m_{\text{mean}}$  is in the range of  $[0, 1]$ .  $m_{\text{mean}} = 1$  stands for fully monotonic increase and  $m_{\text{mean}} = 0$  stands for fully monotonic decrease, while  $m_{\text{mean}} = 0.5$  stands for 50% of weight transfers leading to accuracy increase (the most oscillation situation).

##### c: SIGNIFICANCE

Due to the stochastic nature of model training, each method is repeated 30 times with different random seeds. The mean accuracy, standard deviation of accuracy, and mean monotonicity are displayed. In addition, the statistical significance ( $p < 0.05$ ) of each method compared with FT is reported: “Yes+” indicates whether a method is superior to FT with significance; “Yes-” indicates whether a method is inferior to FT with significance; otherwise, “No” is displayed to indicate no statistical significance.

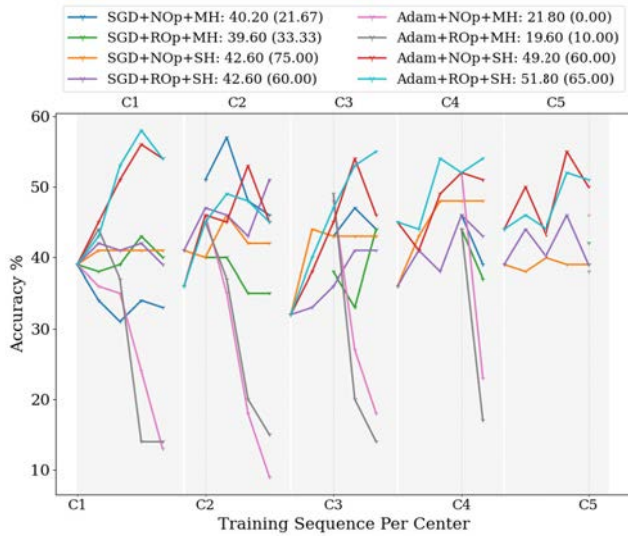
## B. RESULTS OF IID DATA

### 1) IMPROVE THE BASELINE CONFIGURATION FOR NAIVE SWT

We first investigate the performance of naive SWT, i.e. **fine-tuning (FT)** after weight transfer, on IID data using different configurations. When the network uses the multi-head setting, since no classifier heads are available for the subsequent centers, the trained model is only shared to the previous centers for evaluation to track the performance variation, as displayed in Fig. 3. When the single-head setting is used, each trained model is shared to all the centers for evaluation.

The basic configuration for domain-incremental learning scenarios is SGD + new optimizer (NOp) + multi-head (MH). In this setting, apparent forgetting is observed as the accuracy curves drop in multiple sites. With the SGD optimizer, both reloading optimizer (ROp) and single-head (SH) have a minor impact on the overall accuracies, all staying around 40%.

With the Adam optimizer, the multi-head setting leads to catastrophic forgetting of the trained models, since FT



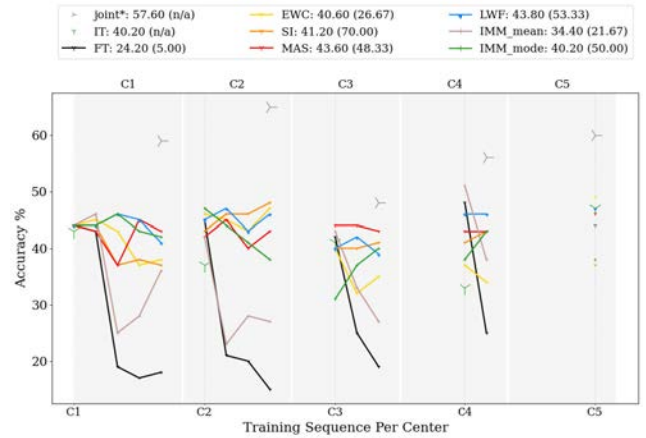
**FIGURE 3.** The naive SWT (i.e., FT) performance with different configurations on IID data. The configurations include SGD vs. Adam, new optimizer (NOp) vs. reload optimizer (ROp), and multi-head (MH) vs. single-head (SH). The average accuracy (monotonicity) for each method is reported in the legend. The implementation and performance visualization are adapted from [17].

only has low accuracies of 21.80% and 19.60% for NOp and ROp, respectively. The forgetting problem is more severe for Adam than SGD with the multi-head setting, because Adam allows a larger learning rate for training. When the single-head (SH) setting is used, FT achieves high accuracies of 49.20% and 51.80% for NOp and ROp, respectively. Since the configuration of Adam + ROp + SH achieves the best accuracy among different configurations (also better than all the configurations with SGD), it is used as the baseline configuration to investigate the influence of continual learning regularization methods.

In our experiments, when the SGD optimizer is replaced by Adam, the accuracies of **joint** training (all the data from different centers is mixed together) and **independent training (IT)** (each center trains a model independently from its own data) are improved in general as well. This confirms the benefit of using an adaptive optimizer.

## 2) CONTINUAL LEARNING REGULARIZATION AVOIDS CATASTROPHIC FORGETTING IN THE MULTI-HEAD SETTING

When the typical configuration for continual learning is used for ITL but with the Adam optimizer (i.e., Adam + NOp + MH), the accuracy plots of different methods are displayed in Fig. 4. The accuracy of FT for Center 1 stays around 41% after training in Center 2. However, it decreases drastically to below 20% after the subsequent trainings. Its accuracies for other centers drop drastically as well after training in the next centers. This is caused by the large change in the feature extractor ( $\mathcal{F}$ ) parameters after training in different centers, which leads to the mismatch of the feature extractor  $\mathcal{F}$  and the center-specific classifier  $\mathcal{T}$ .



**FIGURE 4.** The SWT results of continual learning regularization methods on the Tiny ImageNet data with the setting: Adam + new optimizer (NOp) + multi-head (MH).

Instead, all the continual learning regularization methods, except IMM-mean, achieve relatively stable accuracies. Although the accuracy of IMM-mean for Center 1 drops together with that of FT first, later it turns to rise back. It is worth mentioning that SI and FT have the same implementation in our work except that  $\lambda$  is set to zero for FT in Eqn. (2). Fig. 4 demonstrates that all the continual learning regularization methods are effective to reduce the catastrophic forgetting problem for single-task multicenter collaboration with the multi-head setting.

## 3) CONTINUAL LEARNING REGULARIZATION METHODS IMPROVE MONOTONICITY BUT NOT ACCURACY FOR IID DATA

The performances of different methods with SWT for IID data are displayed in Tab. 1. With all random initializations in the first center (for 30 repeats, all the models trained in the first center use different random seeds), the average accuracy of FT is 49.49%, while those of joint and IT are 57.30% and 39.26%, respectively. Among all the continual learning regularization methods, EWC, SI and MAS have similar or even worse monotonicity, but LWF, EBLL, IMM-mean and IMM-mode have higher monotonicity than FT. Nevertheless, all the continual learning regularization methods cannot improve the accuracy, and MAS, EBLL and IMM-mode even have significantly worse accuracies. The same conclusion is observed from the CWT results for IID data displayed in Tab. 2, as all the continual learning regularization methods are not significantly better than FT when the initial model uses 10 epochs, although the monotonicity of LWF, EBLL, IMM-mean and IMM-mode is better.

## 4) INITIALIZATION MODEL MATTERS FOR SWT

The 30 models trained in the first center have an average accuracy of 40.90%. Among them, a low accuracy model (35.0%) and a high accuracy model (45.2%) are chosen as a fixed initial model for subsequent SWT training.

In general, all methods achieve higher accuracies when a high-accuracy model is used as initialization than a low-accuracy model. Especially, with non-IID data (noisy data in Center 5) as displayed in Tab. 3, LWF, EBLL, IMM-mean and IMM-mode are significantly superior to FT when a high-accuracy initialization is used, while EBLL and IMM-mean are significantly worse than FT when a low-accuracy initialization is used. For LWF and EBLL, the model trained in the first center is used as a teacher model for the training in the second center. Therefore, a high accuracy is beneficial to improve the accuracies of subsequent models. For IMM-mean and IMM-mode, the model weights trained in the subsequent centers are all constrained (by the L2 transfer) to be close to the weights in the first center and the model trained in the first center always contributes to the subsequent merged models. Therefore, a high performance initialization is crucial for the performance of the final merged model.

#### 5) NON-CONVERGENT MODEL IS PREFERRED FOR INITIALIZATION IN CWT

The CWT results on IID data are displayed in Tab. 2. When the initial model is trained with 10 epochs in Center 1, FT achieves an average accuracy of 56.18%, which is slightly lower than (0.82% with significance) but very close to that of joint data. All other methods except MAS also achieve very high accuracies. In comparison, when the initial model is trained with 50 epochs in Center 1 (models typically converge before 50 epochs), most methods (except IMM-mean and IMM-mode without significance) achieve a significantly lower average accuracy, irrespective if a high-accuracy initial model or a low-accuracy initial model is used, on both IID and non-IID data. Please see the CWT results with convergent models as initializations in Tab. 10 and Tab. 11 in the appendix. A convergent model in the first center very likely reaches a local minimum, which has high energy barriers to reach other local minima or the global minimum. Hence, it is more difficult to be optimized to reach a better minimum in the subsequent training than a non-convergent model. Therefore, a non-convergent model is preferred for the initialization in CWT, which is in a stark contrast to the observation that SWT prefers a high-performance initial model.

### C. RESULTS OF NON-IID DATA

#### 1) LWF, EBLL AND IMM IMPROVE BOTH ACCURACY AND MONOTONICITY FOR NON-IID DATA

The SWT results for non-IID data with noisy data in Center 5 are displayed in Tab. 3. For SWT, the initial model performance matters as described above. When the performances of initial models are randomized, LWF, IMM-mean and IMM-mode achieve slightly better average accuracies than FT. However, when a high-accuracy initial model is used for all the 30 repeats, LWF, EBLL, IMM-mean and IMM-mode all achieve significantly better accuracies than FT, in addition to better monotonicity. The plots of

representative accuracy curves for different methods with the high-accuracy initialization are displayed in Fig. 5. The curves of FT in Centers 1-4 have obvious accuracy drops after training in Center 5 because of data heterogeneity in Center 5. In contrast, the curves of LWF, EBLL, IMM-mean and IMM-mode in Centers 1-4 have lower magnitudes of accuracy drops after training in Center 5.

The benefit in accuracy improvement of LWF, EBLL, IMM-mean and IMM-mode is more evident in the CWT results for non-IID data (noisy data in Center 5), as displayed in Tab. 5. The average accuracy of FT is 50.20%, which is 5.42% lower than that of the joint method. In contrast, the average accuracies of LWF, EBLL, IMM-mean and IMM-mode are all very close to that of joint training, with a gap of around 1%. The representative accuracy curves of FT, LWF, EBLL and IMM are displayed in Fig. 6, where both FT and EWC have large and frequent oscillations, while LWF, EBLL, IMM-mean and IMM-mode have much lower oscillation magnitudes.

When the noisy data is located in Center 5, the final model tends to overfit to such noisy data and hence has decreased performance on the test data from other centers. This can be observed clearly from Fig. 5 and Fig. 6, where the accuracy of FT at each center typically drops after the model is trained in Center 5. In contrast, LWF and EBLL use teacher-student learning to avoid large deviation from the previous models, while IMM-mean and IMM-mode use model merging to avoid such drastic performance drops. All these methods (LWF, EBLL, IMM-mean and IMM-mode) are able to improve the accuracy smoothly in general after training in Center 5.

#### 2) IMPORTANCE-BASED METHODS (EWC, SI AND MAS) HAVE NO SIGNIFICANT BENEFITS

The network parameter importance based methods EWC, SI and MAS could not get significantly better performance than FT in our experiments for both IID data (Tab. 1 and Tab. 2) and non-IID data (Tab. 3 and Tab. 5). These methods have the plasticity-stability dilemma [55]: stability is gained with a larger  $\lambda$  value in Eqn. (3), while plasticity is gained with a lower  $\lambda$  value. In this work, different values for  $\lambda$  are checked. For example, for the SWT performance of SI on non-IID data (noisy data in Center 5), SI achieves average accuracies of 48.0%, 46.97%, 46.08% and 44.72% for  $\lambda = 0.1, 1, 10$  and  $100$ , respectively. With a larger  $\lambda$  value, a lower average accuracy is obtained but with higher stability, and SI with a low  $\lambda$  value of 0.1 has similar performance to FT. For multi-task continual learning scenarios, such importance-based methods constrain important parameter weights in the feature extractor with less change to avoid catastrophic forgetting. For single-task continual learning scenarios in this work, an alternative definition that allows more freedom for important parameter weights but constrains unimportant ones is also investigated with the following



**TABLE 1. The average performances of different methods with SWT for IID Tiny ImageNet data (%) (30 repeats).**

Initialization	Metric	joint	IT	FT	EWC	SI	MAS	LWF	EBLL	IMM-mean	IMM-mode
Random (40.90)	accuracy	57.30	39.26	49.49	49.62	49.12	44.25	47.96	46.36	48.12	47.35
	standard deviation	1.54	2.26	2.59	2.45	2.42	2.05	2.15	1.64	2.40	2.13
	monotonicity	n.a.	n.a.	71.03	70.69	68.28	65.86	<b>78.79</b>	<b>77.93</b>	<b>80.17</b>	<b>74.31</b>
	significance	n.a.	n.a.	n.a.	No	No	Yes-	No	Yes-	No	Yes-
Low (35.0)	accuracy	n.a.	n.a.	49.07	49.1	48.84	42.97	49.19	45.65	46.63	48.14
	standard deviation	n.a.	n.a.	1.94	1.62	1.43	0.88	1.42	2.13	1.53	1.52
	monotonicity	n.a.	n.a.	78.83	81.5	80.67	78.0	<b>86.83</b>	<b>85.33</b>	<b>88.0</b>	<b>86.5</b>
	significance	n.a.	n.a.	n.a.	No	No	Yes-	No	Yes-	Yes-	No
High (45.2)	accuracy	n.a.	n.a.	51.44	51.71	51.19	47.53	50.43	48.73	51.1	49.07
	standard deviation	n.a.	n.a.	1.86	1.73	1.55	1.26	1.12	1.59	0.89	1.22
	monotonicity	n.a.	n.a.	67.0	68.83	67.5	63.33	<b>74.5</b>	<b>73.5</b>	<b>73.0</b>	67.0
	significance	n.a.	n.a.	n.a.	No	No	Yes-	No	Yes-	No	Yes-

**TABLE 2. The average performances of different methods with CWT for IID Tiny ImageNet data (%) (30 repeats).**

Initialization	Metric	joint	IT	FT	EWC	SI	MAS	LWF	EBLL	IMM-mean	IMM-mode
Random (27.25) (10 epochs)	accuracy	57.30	39.26	<b>56.18</b>	<b>56.03</b>	<b>55.47</b>	<b>50.82</b>	<b>56.33</b>	<b>56.62</b>	<b>55.34</b>	<b>55.23</b>
	standard deviation	1.54	2.26	1.54	1.72	1.22	3.56	1.89	2.07	2.44	2.34
	monotonicity	n.a.	n.a.	63.19	63.0	63.86	64.67	<b>71.73</b>	<b>72.63</b>	<b>75.10</b>	<b>74.30</b>
	significance	n.a.	n.a.	n.a.	No	No	Yes-	No	No	No	No

**TABLE 3. The average performances of different methods with SWT for non-IID Tiny ImageNet data (noisy data in Center 5) (%) (30 repeats).**

Initialization	Metric	joint	IT	FT	EWC	SI	MAS	LWF	EBLL	IMM-mean	IMM-mode
Random (40.75)	accuracy	55.62	39.12	46.51	46.75	46.13	42.47	<b>48.21</b>	46.97	47.84	47.78
	standard deviation	1.42	1.94	2.26	1.41	1.88	2.26	1.83	1.99	1.98	2.02
	monotonicity	n.a.	n.a.	61.17	62.67	61.83	57.83	<b>77.5</b>	<b>74.33</b>	<b>74.83</b>	<b>75.17</b>
	significance	n.a.	n.a.	n.a.	No	No	Yes-	<b>Yes+</b>	No	No	No
Low (34.4)	accuracy	n.a.	n.a.	47.07	46.26	46.03	38.52	47.52	44.53	44.82	46.11
	standard deviation	n.a.	n.a.	1.73	1.92	1.58	1.06	1.58	1.99	2.19	2.01
	monotonicity	n.a.	n.a.	74.17	74.33	73.0	63.0	<b>84.17</b>	<b>80.83</b>	<b>88.17</b>	<b>85.33</b>
	significance	n.a.	n.a.	n.a.	No	No	Yes-	No	Yes-	Yes-	No
High (43.8)	accuracy	n.a.	n.a.	47.11	46.67	46.97	45.69	<b>49.91</b>	<b>48.55</b>	<b>50.67</b>	<b>48.24</b>
	standard deviation	n.a.	n.a.	1.57	1.93	1.44	1.46	1.16	1.91	1.08	1.53
	monotonicity	n.a.	n.a.	54.83	55.67	56.83	56.0	<b>72.0</b>	<b>71.17</b>	<b>71.5</b>	<b>70.17</b>
	significance	n.a.	n.a.	n.a.	No	No	Yes-	<b>Yes+</b>	<b>Yes+</b>	<b>Yes+</b>	<b>Yes+</b>
High (43.8) 4-center training	accuracy	n.a.	n.a.	46.89	46.99	47.53	46.51	47.58	47.26	<b>49.25</b>	<b>48.11</b>
	standard deviation	n.a.	n.a.	1.60	1.45	1.87	1.39	1.28	1.50	1.24	1.66
	monotonicity	n.a.	n.a.	58.0	61.56	63.78	61.33	67.78	69.78	70.44	71.33
	significance	n.a.	n.a.	n.a.	No	No	No	No	No	<b>Yes+</b>	<b>Yes+</b>

**TABLE 4. The average performances of different methods with SWT for non-IID Tiny ImageNet data (noisy data in Center 3) (%) (30 repeats).**

Initialization	Metric	joint	IT	FT	EWC	SI	MAS	LWF	EBLL	IMM-mean	IMM-mode
Random (39.63)	accuracy	55.55	38.86	47.75	48.15	47.54	42.85	46.99	45.47	46.43	46.89
	standard deviation	1.53	1.94	2.0	1.75	1.68	1.84	2.13	2.34	2.21	2.36
	monotonicity	n.a.	n.a.	67.17	70.0	63.0	60.17	<b>74.33</b>	<b>76.67</b>	<b>76.5</b>	<b>76.33</b>
	significance	n.a.	n.a.	n.a.	No	No	Yes-	No	Yes-	No	No
Low (34.2)	accuracy	n.a.	n.a.	47.43	47.23	46.35	40.99	47.76	44.03	44.97	46.47
	standard deviation	n.a.	n.a.	1.71	1.84	1.46	0.97	1.41	2.11	1.41	1.48
	monotonicity	n.a.	n.a.	78.67	75.5	75.67	68.5	<b>84.5</b>	<b>80.33</b>	<b>86.33</b>	<b>84.67</b>
	significance	n.a.	n.a.	n.a.	No	No	Yes-	No	Yes-	Yes-	No
High (44.8)	accuracy	n.a.	n.a.	48.0	48.31	48.11	44.9	48.49	46.56	46.6	45.04
	standard deviation	n.a.	n.a.	1.86	1.89	1.65	1.98	1.65	2.27	3.74	4.38
	monotonicity	n.a.	n.a.	60.5	61.33	60.33	57.5	<b>72.67</b>	<b>67.67</b>	<b>69.33</b>	<b>68.5</b>
	significance	n.a.	n.a.	n.a.	No	No	Yes-	No	No	No	Yes-

modified objective function,

$$\mathcal{L}^{(\mu)} = \mathcal{L}_{\text{task}} + \lambda^{(\mu)} \sum_k \frac{1}{1 + \Omega_k^{(\mu-1)}} \left( \theta_k^{(\mu-1)} - \theta_k \right)^2.$$

However, with the modification, EWC, SI and MAS still do not outperform FT. All in all, our experiments demonstrate that importance-based methods (EWC, SI and MAS)

have no significant benefits for the single-task multicenter collaboration scenarios in our ITL framework.

### 3) THE ORDER OF CENTERS MATTERS FOR TRAINING WITH NON-IID DATA

When the noisy data is located in Center 5, a significant benefit in accuracy improvement for LWF, EBLL and IMM

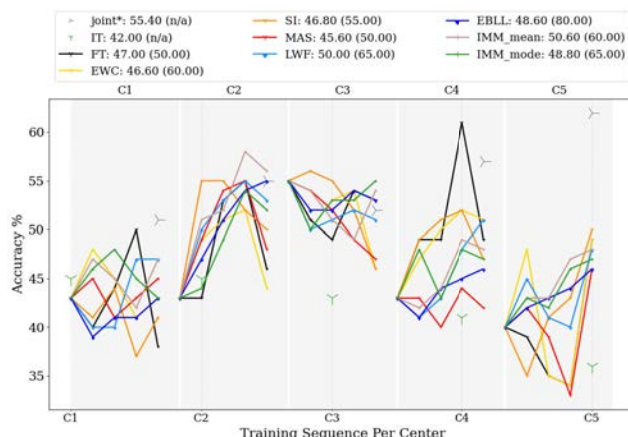
**TABLE 5.** The average performances of different methods with CWT for Non-IID Tiny ImageNet data with random 10-epoch initializations (%)(30 repeats).

Noisy center	Metric	joint	IT	FT	EWC	SI	MAS	LWF	EBLL	IMM-mean	IMM-mode
Center 5	accuracy	55.62	39.12	50.20	50.73	49.54	40.05	<b>54.67</b>	<b>54.90</b>	<b>54.17</b>	<b>54.42</b>
	standard deviation	1.42	1.94	1.62	1.75	1.77	3.58	2.07	1.79	1.79	1.80
	monotonicity	n.a.	n.a.	59.80	58.94	59.92	59.97	<b>67.83</b>	<b>70.08</b>	<b>73.33</b>	<b>71.50</b>
	significance	n.a.	n.a.	n.a.	No	No	Yes-	<b>Yes+</b>	<b>Yes+</b>	<b>Yes+</b>	<b>Yes+</b>
Center 3	accuracy	55.55	38.86	52.51	52.17	51.15	42.23	53.61	53.59	52.72	52.67
	standard deviation	1.53	1.94	1.77	1.98	1.77	4.14	2.10	1.85	1.87	1.83
	monotonicity	n.a.	n.a.	60.94	60.36	59.75	60.17	<b>67.42</b>	<b>67.58</b>	<b>72.61</b>	<b>71.22</b>
	significance	n.a.	n.a.	n.a.	No	Yes-	Yes-	No	No	No	No
Center 4 and 5	accuracy	54.82	37.79	50.51	50.19	49.58	39.65	<b>52.63</b>	<b>52.78</b>	51.67	51.25
	standard deviation	1.58	1.90	1.98	1.83	1.52	3.34	2.13	1.85	1.87	1.83
	monotonicity	n.a.	n.a.	59.61	59.14	59.0	58.53	<b>65.39</b>	<b>67.61</b>	<b>73.19</b>	<b>71.06</b>
	significance	n.a.	n.a.	n.a.	No	No	Yes-	<b>Yes+</b>	<b>Yes+</b>	No	No

is observed. However, when the noisy data is located in Center 3, all the continual learning regularization methods are not better than FT, with some (e.g., IMM-mode for SWT with a high-accuracy initialization) even being significantly worse than FT, as displayed in Tab. 4 and Tab. 5. It is worth noting that the accuracies of all the methods with noisy data in Center 3 are lower than those obtained by LWF, EBLL and IMM when the noisy data is in Center 5. This indicates that the order of centers matters for training with non-IID data. When two centers (Center 4 and Center 5) have noisy data, LWF and EBLL can still achieve significantly better accuracies than FT, as demonstrated by the last row of Tab. 5. Although the 95% confidence intervals of IMM-mean and IMM-mode overlap with that of FT, their average accuracies are higher than that of FT. It is suggested by our experiments that superior accuracy can be obtained by certain continual learning regularization (LWF, EBLL and IMM) when the centers with heterogeneous data are placed at the end of the training sequence; otherwise no improvement from continual learning regularization is achieved. The fundamental reason is similar to the reliance on initialization: if the center with heterogeneous data is located in the beginning or middle of the training sequence, a suboptimal/inferior model for all the test data is obtained at that center, which influences all the subsequent trainings.

4) THE HETEROGENEOUS DATA IS STILL BENEFICIAL TO IMPROVE PERFORMANCE

When Center 5 contains noisy data, training in Center 5 causes accuracy drops for FT. Therefore, it is interesting to know whether training on four centers without Center 5 will lead to superior performance for the overall test in all the five centers. The last row of Tab. 3 indicates that FT achieves slightly lower accuracy (four centers: 46.89% vs. five centers: 47.11% ) without significance. The same is observed for IMM-mode (48.11% vs. 48.24%). However, LWF (47.58% vs. 49.91%), EBLL (47.26% vs. 48.55%) and IMM-mean (49.25% vs 50.67%) all obtain **significantly** lower accuracies when only data from four centers is used for training. Therefore, the heterogeneous data in Center 5 is still beneficial to improve performance for LWF, EBLL and IMM-mean. Note that the significance here is computed



**FIGURE 5.** The SWT results of continual learning regularization methods in ITL on Tiny ImageNet data, where the data in Center 5 contains noise.

between four centers and 5 centers, while the significance in Tab. 3 is computed between each method and FT.

V. EXPERIMENTS ON RETINAL FUNDUS DATA FOR DIABETIC RETINOPATHY CLASSIFICATION

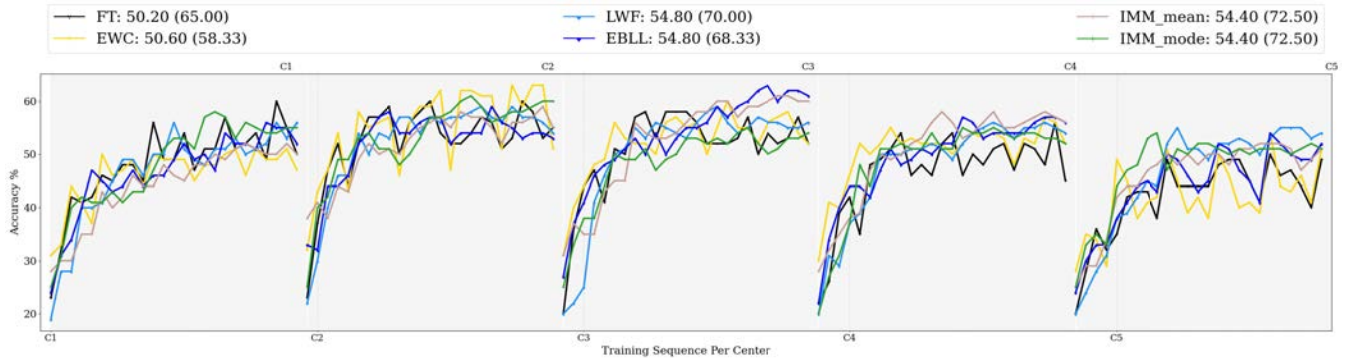
To demonstrate the efficacy of the continual learning methods on real medical data, all the methods are applied to retinal fundus images for diabetic retinopathy classification and T1 contrast enhanced MRI images for glioma segmentation (Section VI). In this diabetic retinopathy classification experiment, the data is split based on imaging devices where data heterogeneity is clearly observed, which is a realistic data split for multicenter collaboration.

A. EXPERIMENTAL SETUP

1) DATASET

The retinal fundus images used in this work are from two sources: the retinal fundus multi-disease image dataset (RFMiD) [56] and the Kaggle diabetic retinopathy detection dataset (KDRDD).<sup>2</sup> Such images are acquired from different imaging systems with different cameras and hence have different sizes. All the images are cropped and resized to

<sup>2</sup><https://www.kaggle.com/competitions/diabetic-retinopathy-detection/data>



**FIGURE 6.** The representative CWT results of continual learning regularization methods on the Tiny ImageNet data (with different initial models), where the data in Center 5 contains noise. LWF, EBLL, IMM-mean, and IMM-mode demonstrate a lower degree of oscillations than FT and EWC. The joint model has an average accuracy of 55.62%. The curves of SI and MAS are similar to that of EWC and hence are omitted for better visualization.

**TABLE 6.** The details of retinal fundus images in the five centers.

Center	Center 1	Center 2	Center 3	Center 4	Center 5
Data source	RFMiD	RFMiD	KDRDD	KDRDD	KDRDD
Original size	1424 × 2144	2848 × 4288	2848 × 4288	2000 × 3008	3264 × 4928
Total number	852	254	155	398	720
Number of negative	333	142	122	326	512
Number of positive	519	112	33	72	208

**TABLE 7.** The average performances of different methods with SWT for retinal data (%) (30 repeats).

Initialization	Metric	joint	IT	FT	EWC	SI	MAS	LWF	EBLL	IMM-mean	IMM-mode
Random (53.12)	accuracy	79.69	44.69 (64.86)	57.21	54.57	56.87	55.30	<b>61.09</b>	<b>59.85</b>	<b>61.22</b>	<b>63.76</b>
	standard deviation	4.28	3.25 (5.45)	10.02	8.23	9.46	9.79	8.42	7.33	9.86	10.21
	monotonicity	n.a.	n.a.	55.50	53.0	53.67	53.17	<b>63.83</b>	<b>67.83</b>	<b>63.0</b>	<b>63.67</b>
	significance	n.a.	n.a.	n.a.	No	No	No	No	No	No	No
High (65.29)	accuracy	n.a.	n.a.	61.37	63.12	63.78	<b>66.06</b>	<b>71.11</b>	<b>69.58</b>	<b>68.83</b>	<b>70.27</b>
	standard deviation	n.a.	n.a.	5.71	5.34	5.22	3.45	4.23	3.13	2.67	2.38
	monotonicity	n.a.	n.a.	47.50	47.33	48.17	49.83	<b>65.50</b>	<b>64.67</b>	<b>58.33</b>	<b>58.50</b>
	significance	n.a.	n.a.	n.a.	No	No	Yes+	Yes+	Yes+	Yes+	Yes+

**TABLE 8.** The average performances of different methods with CWT for retinal data (%) (30 repeats).

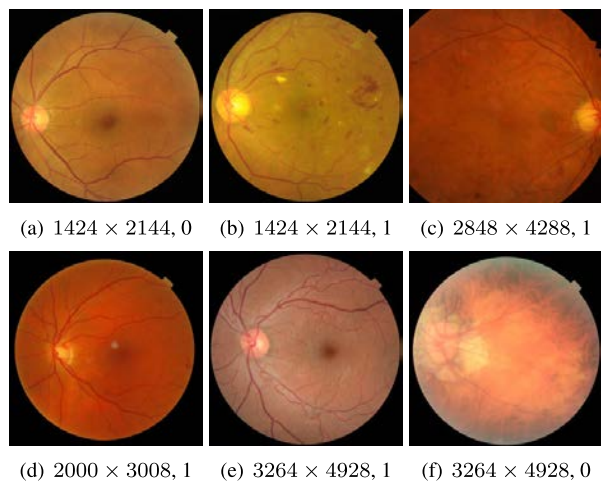
Initialization	Metric	joint	IT	FT	EWC	SI	MAS	LWF	EBLL	IMM-mean	IMM-mode
Random (54.25)	accuracy	79.69	44.69 (64.86)	72.45	73.26	70.69	71.90	<b>74.35</b>	<b>74.22</b>	<b>74.53</b>	<b>74.96</b>
	standard deviation	4.28	3.25 (5.45)	2.53	3.61	2.89	3.10	2.36	2.82	3.02	2.53
(15 epochs)	monotonicity	n.a.	n.a.	52.17	52.98	50.48	58.26	<b>63.61</b>	<b>83.42</b>	<b>61.42</b>	<b>61.08</b>
	significance	n.a.	n.a.	n.a.	No	No	No	Yes+	Yes+	Yes+	Yes+

256 × 256 images. The images are further divided into different centers according to their original image sizes. In such a way, images at different centers have different styles, which reflects data heterogeneity in a realistic manner. In this work, a binary classification of diabetic retinopathy is evaluated. Five centers are chosen for the experiments in this work and the details of the data at each center are displayed in Tab. 6. Each center has a data split of 64%, 16%, and 20% for training, validation, and test. Some exemplary images are displayed in Fig. 7 along with their original image size and diabetic retinopathy classification. Note that for images from the source of KDRDD, diabetic retinopathy negative images may contain other unclassified retinal diseases like Fig. 7(f). Therefore, the images from KDRDD are noisier than those from RFMiD. The symptoms of diabetic retinopathy include

small bulges in blood vessels, blood leakages, deposits of cholesterol, irregular beading, obliteration of blood vessels and growth of new blood vessels [57]. For mild cases, the symptoms are difficult to observe for human being (e.g., Figs. 7(d) and (e)) and hence the automatic classification of diabetic retinopathy has important clinical value.

## 2) NETWORK

The same small VGG-9 network for the Tiny ImageNet data is used for diabetic retinopathy classification. For SWT, all the models are trained for 100 epochs at each center. For monitoring overfitting, the learning rate is decayed into half after every  $E_{val} = 5$  epochs and early stopping is executed after  $E_{stop} = 20$  epochs. For CWT, a weight transfer



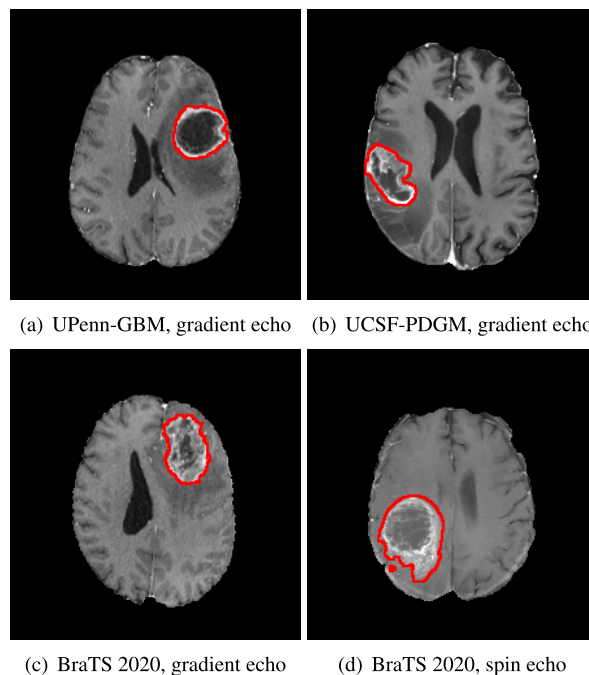
**FIGURE 7.** Exemplar retinal fundus images from different cameras (same size from the same camera). The original image sizes before preprocessing are displayed in the subcaptions, followed by the corresponding diabetic retinopathy classification (0: negative; 1: positive). In (f), other unclassified retinal disease exists, although it is diabetic retinopathy negative.

frequency of  $E_{transfer} = 15$  is investigated with a total number of 5 iterations. The batch size is 10. The Adam optimizer is used for training with an initial learning rate of 0.001. Other training parameters are the same as those for the Tiny ImageNet data. It is worth noting that the weighted random sampler from Pytorch is applied to get a balanced class distribution for each training and validation batch, otherwise it is difficult for the network to learn due to class imbalance.

### B. RESULTS

The SWT results with different algorithms are displayed in Tab. 7. The joint model with mixed training data achieves an average accuracy of 79.69% among the 30 repeats. IT achieves an average accuracy of 64.86% (the model in each center is trained and tested with its own data), which is relatively high. This is because each local model is optimized for the specific data from the same center. But such a model has very poor generalizability to test data in other centers, e.g., the local model from Center 5 achieves an average accuracy of 44.69% for test datasets from all centers. With all random initializations, FT achieves an average accuracy of 57.21%. LWF, EBLL, IMM-mean and IMM-mode have achieved higher average accuracies than FT, but without statistical significance. Picking a fixed high-accuracy (65.29%) model for the first center and continuing stochastic training subsequently for 30 repeats, LWF, EBLL, IMM-mean and IMM-mode have significantly better accuracies than FT, which is consistent with the observation from the Tiny ImageNet experiments.

The CWT results are displayed in Tab. 8. With CWT, FT achieves a better accuracy of 72.45%, which is 15.24% higher than that with SWT. The importance weight constrained methods (EWC, SI, and MAS) again have comparable accuracies to FT. Consistently, LWF, EBLL, IMM-mean and IMM-mode have higher accuracies than



**FIGURE 8.** Exemplar T1 contrast enhanced MRI images from different centers in our experiments. Center 1 to 3 ((a) to (c)) use gradient echo MRI sequences, while Center 4 uses a spin echo MRI sequence. The UPenn-GBM (a) and BraTS 2020 (c) images are noisier than the UCSF-PDGM (a) images. The red curves contour the reference tumor cores.

FT, among which LWF, IMM-mean and IMM-mode have statistical significance ( $p < 0.05$ ).

## VI. EXPERIMENTS ON GLIOMA SEGMENTATION

To demonstrate the efficacy of continual learning methods on real medical image segmentation tasks (in addition to the image classification tasks), the methods are also investigated on the glioma segmentation using multicenter data.

### A. EXPERIMENTAL SETUP

#### 1) DATASET

The glioma datasets are from three sources: BraTS 2020 data [58], the UCSF-PDGM dataset [59], and the UPenn-GBM dataset [60]. The BraTS 2020 dataset contains multicentric data, but such center information for each patient's data is not available to users. Therefore, in our study, the BraTS 2020 dataset is simply split into two virtual centers based on the two basic T1 contrast enhanced MRI sequences gradient echo (e.g., MPRAGE) and spin echo (e.g., SPACE). In spin-echo images, the contrast between gray/white matters and the contrast for blood vessels are less than those in gradient echo images, as displayed in Fig. 8(c) and Fig. 8(d), respectively. The images from UCSF-PDGM and UPenn-GBM both use the gradient echo T1 contrast sequences. However, the UCSF-PDGM dataset is less noisy than the UPenn-GBM dataset and the BraTS 2020 dataset, as displayed in Fig. 8. All the volumes have  $240 \times 240 \times 155$  voxels with an isotropic spacing of 1 mm. The BraTS glioma segmentation challenge aims to segment

normal background tissues, necrotic tumor core, contrast-enhancing tumor, and peritumoral edema. The tumor core consists of necrotic tumor core and contrast-enhancing tumor. As tumor core frequently represents the only relevant tumor compartment for clinical treatment planning [61], [62], [63], in this work we focus on tumor core segmentation on contrast-enhanced T1-weighted images. To save computation, the volumes are cropped to  $224 \times 224$  2D images with an isotropic spacing of 1 mm and one slice among 10 neighboring slices is selected to avoid similarity. The four centers have total numbers of 374, 884, 486, and 402 images. Each center has a data split of 40%, 15%, and 45% approximately for training, validation, and test (patients exclusively for training, validation, and test).

## 2) NETWORK

The U-Net from [64] is used for glioma segmentation. It consists of 5 levels. The Dice coefficient loss function is used for training and validation. For SWT, all the models are trained for 100 epochs at each center. For monitoring overfitting, the learning rate is decayed into half after every  $E_{val} = 5$  epochs and early stopping is executed after  $E_{stop} = 20$  epochs. For CWT, a weight transfer frequency of  $E_{transfer} = 15$  is investigated with a total number of 5 iterations. The batch size is 16. The Adam optimizer is used for training with an initial learning rate of 0.001. A constant  $\lambda$  value of 0.1 instead of 1 for LWF, EWC, SI and MAS is used. Due to the similar performance of EBLL to LWF but with much more computation for training each encoder at each iteration, for the glioma segmentation EBLL is omitted. Other training parameters are the same as those for the Tiny ImageNet data.

## B. RESULTS

The SWT results with different algorithms are displayed in Tab. 9. The joint model with mixed training data achieves an average Dice coefficient of 0.7381 among the 30 repeats. Among all the methods, LWF, IMM-mean and IMM-mode have higher average Dice coefficients than FT, which is consistent with our observation from the Tiny ImageNet data and the retinal data. Among them, LWF is significantly better than FT, while IMM-mean and IMM-mode do not have significance due to their large variations. The exemplary plots of different SWT methods are displayed in Fig. 9. For the FT curve in Center 1, its Dice coefficient increases after subsequent training in Center 2 and Center 3, but it has a drastic decrease after training in Center 4. In contrast, LWF, IMM-mean and IMM-mode do not have such drastic drops after training in Center 4 (because data in Center 4 uses the T1 spin echo instead of T1 gradient echo sequence). Note that Center 3 (BraTS 2020, gradient echo) has relatively low Dice coefficients for all models because the BraTS 2020 gradient echo dataset contains gliomas of different stages and meanwhile the data is noisy. In contrast, the UPenn-GBM dataset contains glioblastomas which typically

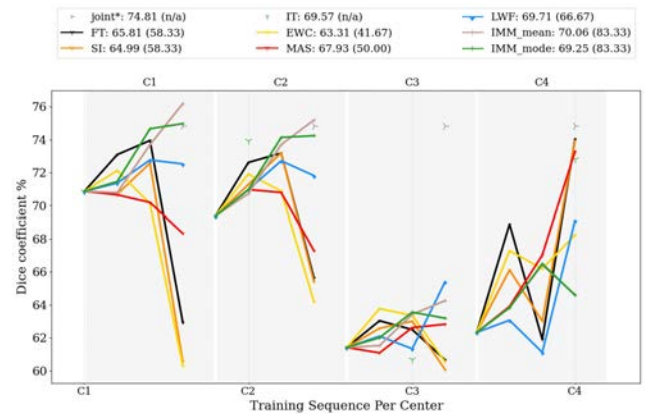


FIGURE 9. The exemplary SWT results of continual learning regularization methods on glioma data.

have larger sizes than those in the BraTS dataset, and the UCSF-PDGM dataset is less noisy.

The CWT results are displayed in Tab. 9 as well. LWF has a significantly better average Dice coefficient 0.7345 in comparison to that of FT 0.6708. Although IMM-mean and IMM-mode have better average Dice coefficients, their values are very close to that of FT. The importance weight constrained methods (EWC, SI, and MAS) again have comparable Dice coefficients to FT. The exemplary plots of different CWT methods are displayed in Fig. 10. EWC has reduced oscillations compared with FT. However, its final Dice coefficient is not better than FT. LWF has reduced oscillations with stably improved Dice coefficients. It is interesting to observe that IMM-mean and IMM-mode have very large performance drops at the early transfers, while it recovers to a relatively stable state in the late transfers. For segmentation tasks, the model needs a pixel-level classification instead of an image-level classification. Therefore, the final output of a segmentation model is much more sensitive to the model weights than that of a classification model. For IMM-mean and IMM-mode with CWT, these methods merge model weights from previous centers and hence there is a risk that the merged models have a low performance. Since each model reaches a relatively better local minimum in SWT than that in CWT in the early transfers, such performance drop in SWT is less frequently observed than that in CWT in our experiments for IMM.

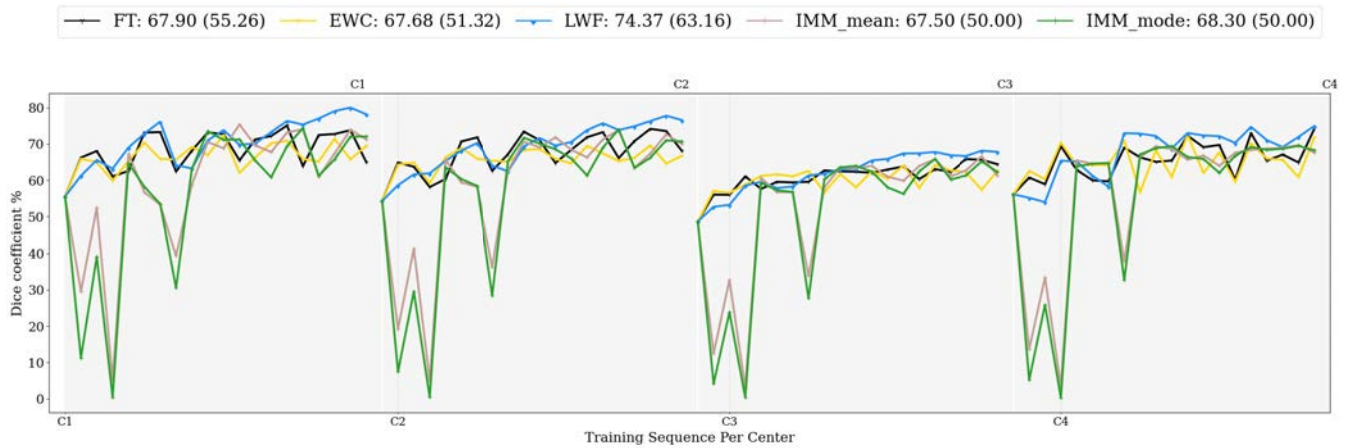
## VII. DISCUSSION

### A. DATA HETEROGENEITY AND CENTER ORDER

Our experiments demonstrate that with our proposed ITL framework, a certain level of forgetting occurs for FT in the IID-data case, but this forgetting is not catastrophic as that in task incremental learning. However, with non-IID data, catastrophic forgetting occurs when the model is trained in the centers with heterogeneous data. In practice, whether the datasets from multiple centers are heterogeneous or not can be determined by prior information before training (e.g., the imaging device and preprocessing pipeline

**TABLE 9.** The average performances of different methods with SWT and CWT for glioma data (%) (30 repeats).

Weight transfer	Metric	joint	IT	FT	EWC	SI	MAS	LWF	IMM-mean	IMM-mode
SWT	Dice coefficient	73.81	59.03 (68.31)	65.56	64.23	65.02	64.44	<b>66.44</b>	<b>67.07</b>	<b>66.38</b>
	standard deviation	1.33	1.26 (1.01)	1.53	1.50	1.85	2.71	1.36	3.66	4.87
	monotonicity	n.a.	n.a.	53.12	53.0	53.67	53.17	55.24	<b>69.27</b>	<b>70.83</b>
	significance	n.a.	n.a.	n.a.	No	Yes-	Yes-	<b>Yes+</b>	No	No
CWT	Dice coefficient	n.a.	n.a.	67.08	66.56	67.06	66.76	<b>73.45</b>	68.06	67.85
	standard deviation	n.a.	n.a.	2.04	2.46	2.72	3.42	1.19	4.52	4.36
	monotonicity	n.a.	n.a.	54.47	55.06	55.89	52.38	<b>64.96</b>	52.15	53.50
	significance	n.a.	n.a.	n.a.	No	No	No	<b>Yes+</b>	No	No



**FIGURE 10.** The exemplary CWT results of continual learning regularization methods on glioma data. LWF demonstrates a lower degree of oscillations than FT. IMM-mean and IMM-mode have severe oscillations for in the early transfers. The curves of SI and MAS are similar to that of EWC and hence are omitted for better visualization.

**TABLE 10.** The average performances of different methods with CWT for IID Tiny ImageNet data (%) (30 repeats).

Initialization	Metric	joint	IT	FT	EWC	SI	MAS	LWF	EBLL	IMM-mean	IMM-mode
Low (35.0)	accuracy	n.a.	n.a.	53.10	52.71	52.37	47.75	53.65	50.57	<b>54.90</b>	<b>54.52</b>
	standard deviation	n.a.	n.a.	1.71	1.72	1.52	1.85	1.40	2.82	1.77	1.59
	monotonicity	n.a.	n.a.	60.92	60.75	60.64	64.88	66.33	77.47	68.94	67.36
	significance	n.a.	n.a.	n.a.	No	No	Yes-	No	Yes-	<b>Yes+</b>	<b>Yes+</b>
High (45.2)	accuracy	n.a.	n.a.	54.77	54.17	53.11	50.25	54.57	51.47	<b>56.02</b>	<b>55.20</b>
	standard deviation	n.a.	n.a.	1.45	1.38	1.53	0.81	1.16	2.16	1.51	1.34
	monotonicity	n.a.	n.a.	57.64	57.83	57.27	65.50	65.23	75.60	65.30	63.77
	significance	n.a.	n.a.	n.a.	No	Yes-	Yes-	No	Yes-	<b>Yes+</b>	No

**TABLE 11.** The average performances of different methods with CWT for non-IID Tiny ImageNet data (noisy data in Center 5) (%) (30 repeats).

Initialization	Metric	joint	IT	FT	EWC	SI	MAS	LWF	EBLL	IMM-mean	IMM-mode
Low (34.4)	accuracy	n.a.	n.a.	52.11	52.24	51.81	44.65	52.49	<b>53.45</b>	<b>53.63</b>	<b>53.36</b>
	standard deviation	n.a.	n.a.	1.65	0.99	1.52	1.92	1.74	1.68	1.23	1.25
	monotonicity	n.a.	n.a.	59.08	60.19	60.67	58.69	<b>69.47</b>	<b>69.39</b>	<b>70.92</b>	<b>68.83</b>
	significance	n.a.	n.a.	n.a.	No	No	Yes-	No	<b>Yes+</b>	<b>Yes+</b>	<b>Yes+</b>
High (43.8)	accuracy	n.a.	n.a.	50.75	50.95	50.01	48.83	<b>52.50</b>	<b>52.3</b>	<b>52.99</b>	<b>52.67</b>
	standard deviation	n.a.	n.a.	1.23	0.90	1.19	0.56	0.95	1.00	1.03	0.87
	monotonicity	n.a.	n.a.	55.97	55.89	56.08	60.03	<b>64.78</b>	<b>65.14</b>	<b>65.50</b>	<b>65.67</b>
	significance	n.a.	n.a.	n.a.	No	No	Yes-	<b>Yes+</b>	<b>Yes+</b>	<b>Yes+</b>	<b>Yes+</b>

information) or be computed in feature space (e.g. using t-distributed stochastic neighbor embedding (t-SNE) [65] to visualize feature distributions of datasets from different centers, or training a small encoder like EBLL does). In addition to this, some SWT or CWT trials using FT can be performed to check the data heterogeneity. Based

on the FT results (e.g., Fig. 5, Fig. 6, Fig. 9 and Fig. 10), if the model performance drops frequently at certain centers, the datasets from these centers are likely to have different features or styles from the majority of centers. Therefore, they should be placed at the end of the center sequence for training.

### B. ROBUSTNESS TO DATA PRIVACY ATTACKS

Although beyond the main focus of this work, it is worth mentioning the robustness of ITL to data privacy attacks. In federated learning, only model parameters or training gradients are shared, but there is still risk of data leakage. Recent attack techniques, e.g. model inversion attacks, have the capability of reverse engineering private data at federated training phases [66] and federated inference phases [67], [68]. However, such attack methods typically work on C2PFL with split network architectures, where intermediate features of input data are shared among different participants. The layer depth of shared features is one key indicator on the vulnerability: the shallower the easier [66]. In addition, such attacks typically are effective for low-resolution images, like the MNIST and CIFAR-10 datasets. Restoration of high-resolution images from deep networks is much more challenging and requires accurate training statistics, e.g., the batch normalization statistics. In this work, because of the peer-to-peer model sharing strategy, no intermediate features of input (test nor training) images are shared. Therefore, restoring private data information is extremely challenging for existing attack algorithms.

### VIII. CONCLUSION

The following conclusions can be drawn from this work:

- With the multi-head setting, naive weight transfer has the risk of catastrophic forgetting for both IID data and non-IID data, and all the continual learning regularization methods are effective to avoid catastrophic forgetting.
- For naive weight transfer, the catastrophic forgetting problem in the multi-head setting can be reduced to a large degree by using the single-head setting.
- For IID data, because of the reduced forgetting problem by the single-head setting, the continual learning regularization methods cannot obtain significantly better performance than FT (both SWT and CWT).
- For non-IID data, apparent forgetting pattern is observed after training on centers with heterogeneous data. LWF, EBLL, IMM-mean and IMM-mode are effective to reduce performance oscillations and achieve better overall performance than FT for both SWT and CWT.
- For SWT, a high-accuracy initial model trained from the first center is preferred for the guidance of the subsequent training. On the contrary, for CWT a non-convergent (low-accuracy) initial model is preferred.
- For non-IID data, the order of centers matters. In particular, the centers with heterogeneous data are preferred to be placed in the last positions for training.
- For both SWT and CWT with non-IID data, the heterogeneous data is still beneficial to improve the overall performance of the shared model.
- CWT is better than SWT in the overall accuracy, at the cost of more frequent weight transfers (communications among centers).

- Despite of performance improvement compared with naive transfer learning, the performance of ITL is still lower than that of joint training with pooled data.

### APPENDIX

#### A. ADDITIONAL CWT RESULTS ON TINY IMAGENET DATA

The CWT results on the Tiny Image data with converged models as the initialization are displayed in Tab. 10 and Tab. 11 for IID data and non-IID data, respectively.

#### B. SI WITH THE ADAM OPTIMIZER

Taking Adam as an example, Adam keeps two bias-corrected moments  $\mathbf{m}(t)$  and  $\mathbf{v}(t)$  to compute adaptive learning rates for different parameters,

$$\begin{aligned} \mathbf{m}(t) &= (\beta_1 \mathbf{m}(t-1) + (1 - \beta_1) \mathbf{g}(t)) / (1 - \beta_1^t), \\ \mathbf{v}(t) &= (\beta_2 \mathbf{v}(t-1) + (1 - \beta_2) \mathbf{g}^2(t)) / (1 - \beta_2^t), \end{aligned} \quad (11)$$

where  $\mathbf{g}(t) = \partial L_{\text{task}} / \partial \boldsymbol{\theta}(t)$  is the gradient of  $L_{\text{task}}$  with respect to the real-time parameter  $\boldsymbol{\theta}(t)$ , and  $\beta_1$  and  $\beta_2$  are two parameters for  $\mathbf{m}(t)$  and  $\mathbf{v}(t)$ , respectively. The  $k$ -th parameter  $\theta_k(t)$  is updated as the following,

$$\theta_k(t+1) = \theta_k(t) - \frac{r}{\sqrt{v_k(t)} + \epsilon} m_k(t), \quad (12)$$

where  $r$  is the fixed learning rate and  $\epsilon$  is a small value to avoid division by zero, while  $\frac{r}{\sqrt{v_k(t)} + \epsilon}$  is an adaptive learning rate automatically adjusted for each parameter.

When continual learning regularization is included, the gradient  $\mathbf{g}(t)$  in Eqn. (11) need to be replaced by the new gradient with regularization  $\mathbf{g}'_k(t)$ . Taking SI as an example,  $\mathbf{g}'_k(t)$  is modified from  $\mathbf{g}(t)$  as follows according to Eqn. (3),

$$\mathbf{g}'_k(t) = \mathbf{g}_k(t) + 2\lambda \boldsymbol{\Omega}_k^{(\mu-1)} (\theta_k^{(\mu-1)} - \theta_k). \quad (13)$$

In addition, the contribution of each network parameter to the change of the task loss  $L_{\text{task}}$  need to be tracked in the variable  $\mathbf{w}_k^{(\mu)}$  in the Adam optimizer,

$$\mathbf{w}_k^{(\mu)} = \int_{t^{\mu-1}}^{t^\mu} \mathbf{g}_k(\boldsymbol{\theta}(t)) \cdot \boldsymbol{\theta}'_k(t) dt, \quad (14)$$

where  $\boldsymbol{\theta}'(t) = \partial \boldsymbol{\theta} / \partial t$  is the network parameter change over time. After training in the current center  $\mu$ , the parameter importance factor  $\boldsymbol{\Omega}^{(\mu)}$  is updated as the following for SI [32],

$$\boldsymbol{\Omega}_k^{(\mu)} = \sum_{\nu < \mu} \frac{\mathbf{w}_k^\nu}{(\theta_k^\nu - \theta_k^{\nu-1})^2 + \epsilon}, \quad (15)$$

where  $\nu$  is the index of a previous center and  $\theta^\nu$  is the parameter set of the final trained model at center  $\nu$ . Hence, the knowledge learned from previous centers are accumulated in this importance factor.

## REFERENCES

- [1] Y. Fu, Y. Lei, T. Wang, W. J. Curran, T. Liu, and X. Yang, "Deep learning in medical image registration: A review," *Phys. Med. Biol.*, vol. 65, no. 20, 2020, Art. no. 20TR01.
- [2] Y. Huang, A. Preuhs, M. Manhart, G. Lauritsch, and A. Maier, "Data extrapolation from learned prior images for truncation correction in computed tomography," *IEEE Trans. Med. Imag.*, vol. 40, no. 11, pp. 3042–3053, Nov. 2021.
- [3] T. Weissmann, Y. Huang, S. Fischer, J. Roesch, S. Mansoorian, H. A. Gaona, A.-O. Gostian, M. Hecht, S. Lettmaier, L. Deloch, B. Frey, U. S. Gaipl, L. V. Distel, A. Maier, H. Iro, S. Semrau, C. Bert, R. Fietkau, and F. Putz, "Deep learning for automatic head and neck lymph node level delineation provides expert-level accuracy," *Frontiers Oncol.*, vol. 13, Feb. 2023, Art. no. 1115258.
- [4] Y. Huang, C. Bert, P. Sommer, B. Frey, U. Gaipl, L. V. Distel, T. Weissmann, M. Uder, M. A. Schmidt, A. Dörfler, A. Maier, R. Fietkau, and F. Putz, "Deep learning for brain metastasis detection and segmentation in longitudinal MRI data," *Med. Phys.*, vol. 49, no. 9, pp. 5773–5786, Sep. 2022.
- [5] M. Avanzo, J. Stancanello, G. Pirrone, and G. Sartor, "Radiomics and deep learning in lung cancer," *Strahlentherapie und Onkologie*, vol. 196, no. 10, pp. 879–887, Oct. 2020.
- [6] R. Beckers, Z. Kwade, and F. Zanca, "The EU medical device regulation: Implications for artificial intelligence-based medical device software in medical physics," *Phys. Medica*, vol. 83, pp. 1–8, Mar. 2021.
- [7] N. Rieke, "The future of digital health with federated learning," *NPJ Digit. Med.*, vol. 3, no. 1, pp. 1–7, Sep. 2020.
- [8] I. Dayan, "Federated learning for predicting clinical outcomes in patients with COVID-19," *Nat. Med.*, vol. 27, no. 10, pp. 1735–1743, Oct. 2021.
- [9] J. Xu, B. S. Glicksberg, C. Su, P. Walker, J. Bian, and F. Wang, "Federated learning for healthcare informatics," *J. Healthcare Informat. Res.*, vol. 5, no. 1, pp. 1–19, Mar. 2021.
- [10] T. Li, A. K. Sahu, A. Talwalkar, and V. Smith, "Federated learning: Challenges, methods, and future directions," *IEEE Signal Process. Mag.*, vol. 37, no. 3, pp. 50–60, May 2020.
- [11] T. Wink and Z. Nocht, "An approach for peer-to-peer federated learning," in *Proc. 51st Annu. IEEE/IFIP Int. Conf. Dependable Syst. Netw. Workshops (DSN-W)*, Jun. 2021, pp. 150–157.
- [12] K. Chang, N. Balachandar, C. Lam, D. Yi, J. Brown, A. Beers, B. Rosen, D. L. Rubin, and J. Kalpathy-Cramer, "Distributed deep learning networks among institutions for medical imaging," *J. Amer. Med. Inform. Assoc.*, vol. 25, no. 8, pp. 945–954, Aug. 2018.
- [13] M. J. Sheller, G. A. Reina, B. Edwards, J. Martin, and S. Bakas, "Multi-institutional deep learning modeling without sharing patient data: A feasibility study on brain tumor segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput. Assist. Intervent.*, 2018, pp. 92–104.
- [14] M. J. Sheller, B. Edwards, G. A. Reina, J. Martin, S. Pati, A. Kotrotsou, M. Milchenko, W. Xu, D. Marcus, R. R. Colen, and S. Bakas, "Federated learning in medicine: Facilitating multi-institutional collaborations without sharing patient data," *Sci. Rep.*, vol. 10, no. 1, pp. 1–12, Jul. 2020.
- [15] M. McCloskey and N. J. Cohen, "Catastrophic interference in connectionist networks: The sequential learning problem," *Psychol. Learn. Motiv.*, vol. 24, pp. 109–165, Jan. 1989.
- [16] T. Lesort, V. Lomonaco, A. Stoian, D. Maltoni, D. Filliat, and N. Díaz-Rodríguez, "Continual learning for robotics: Definition, framework, learning strategies, opportunities and challenges," *Inf. Fusion*, vol. 58, pp. 52–68, Jun. 2020.
- [17] M. De Lange, R. Aljundi, M. Masana, S. Parisot, X. Jia, A. Leonardis, G. Slabaugh, and T. Tuytelaars, "A continual learning survey: Defying forgetting in classification tasks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 7, pp. 3366–3385, Jul. 2022.
- [18] G. M. van de Ven, T. Tuytelaars, and A. S. Tolias, "Three types of incremental learning," *Nature Mach. Intell.*, vol. 4, no. 12, pp. 1185–1197, Dec. 2022.
- [19] M. Masana, X. Liu, B. Twardowski, M. Menta, A. D. Bagdanov, and J. van de Weijer, "Class-incremental learning: Survey and performance evaluation on image classification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 5, pp. 5513–5533, May 2023.
- [20] K. Li, L. Yu, and P.-A. Heng, "Domain-incremental cardiac image segmentation with style-oriented replay and domain-sensitive feature whitening," *IEEE Trans. Med. Imag.*, vol. 42, no. 3, pp. 570–581, Mar. 2023.
- [21] D. Isele and A. Cosgun, "Selective experience replay for lifelong learning," in *Proc. AAAI*, vol. 32, no. 1, 2018, pp. 3302–3309.
- [22] D. Rolnick, A. Ahuja, J. Schwarz, T. Lillicrap, and G. Wayne, "Experience replay for continual learning," in *Proc. Adv. Neural. Inf. Process. Syst.*, vol. 32, 2019, pp. 1–11.
- [23] A. Chaudhry, M. Rohrbach, M. Elhoseiny, T. Ajanthan, P. K. Dokania, P. H. Torr, and M. Ranzato, "Continual learning with tiny episodic memories," in *Proc. ICML*, 2019, pp. 1–13.
- [24] A. A. Rusu, N. C. Rabinowitz, G. Desjardins, H. Soyer, J. Kirkpatrick, K. Kavukcuoglu, R. Pascanu, and R. Hadsell, "Progressive neural networks," 2016, *arXiv:1606.04671*.
- [25] A. Mallya and S. Lazebnik, "PackNet: Adding multiple tasks to a single network by iterative pruning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7765–7773.
- [26] Y. Wu, Y. Chen, L. Wang, Y. Ye, Z. Liu, Y. Guo, and Y. Fu, "Large scale incremental learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 374–382.
- [27] C. Fernando, D. Banarse, C. Blundell, Y. Zwols, D. Ha, A. A. Rusu, A. Pritzel, and D. Wierstra, "PathNet: Evolution channels gradient descent in super neural networks," 2017, *arXiv:1701.08734*.
- [28] R. Aljundi, P. Chakravarty, and T. Tuytelaars, "Expert gate: Lifelong learning with a network of experts," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 7120–7129.
- [29] J. Bang, H. Kim, Y. Yoo, J.-W. Ha, and J. Choi, "Rainbow memory: Continual learning with a memory of diverse samples," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 8214–8223.
- [30] Z. Li and D. Hoiem, "Learning without forgetting," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 12, pp. 2935–2947, Dec. 2018.
- [31] K. James, "Overcoming catastrophic forgetting in neural networks," *Proc. Nat. Acad. Sci. USA*, vol. 114, no. 13, pp. 3521–3526, Mar. 2017.
- [32] F. Zenke, B. Poole, and S. Ganguli, "Continual learning through synaptic intelligence," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 3987–3995.
- [33] H. Shin, J. K. Lee, J. Kim, and J. Kim, "Continual learning with deep generative replay," *Adv. Neural. Inf. Process. Syst.*, vol. 30, 2017.
- [34] S. Niu, Y. Liu, J. Wang, and H. Song, "A decade survey of transfer learning (2010–2020)," *IEEE Trans. Artif. Intell.*, vol. 1, no. 2, pp. 151–166, Oct. 2020.
- [35] F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, and Q. He, "A comprehensive survey on transfer learning," *Proc. IEEE*, vol. 109, no. 1, pp. 43–76, Jan. 2021.
- [36] C. Zhang, Y. Xie, H. Bai, B. Yu, W. Li, and Y. Gao, "A survey on federated learning," *Knowl. Based Syst.*, vol. 216, Mar. 2021, Art. no. 106775.
- [37] Q. Li, Z. Wen, Z. Wu, S. Hu, N. Wang, Y. Li, X. Liu, and B. He, "A survey on federated learning systems: Vision, hype and reality for data privacy and protection," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 4, pp. 3347–3366, Apr. 2023.
- [38] Z. Mai, R. Li, J. Jeong, D. Quispe, H. Kim, and S. Sanner, "Online continual learning in image classification: An empirical survey," *Neuro-computing*, vol. 469, pp. 28–51, Jan. 2022.
- [39] X. Yao and L. Sun, "Continual local training for better initialization of federated models," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2020, pp. 1736–1740.
- [40] J. Yoon, W. Jeong, G. Lee, E. Yang, and S. J. Hwang, "Federated continual learning with weighted inter-client transfer," in *Proc. 38th Int. Conf. Mach. Learn.*, vol. 139, Jul. 2021, pp. 12073–12086.
- [41] A. Usmanova, F. Portet, P. Lalanda, and G. Vega, "Federated continual learning through distillation in pervasive computing," in *Proc. IEEE Int. Conf. Smart Comput. (SMARTCOMP)*, Jun. 2022, pp. 86–91.
- [42] F. E. Casado, D. Lema, M. F. Criado, R. Iglesias, C. V. Regueiro, and S. Barro, "Concept drift detection and adaptation for federated and continual learning," *Multimedia Tools Appl.*, vol. 81, no. 3, pp. 3397–3419, Jan. 2022.
- [43] H. Guan and M. Liu, "Domain adaptation for medical image analysis: A survey," *IEEE Trans. Biomed. Eng.*, vol. 69, no. 3, pp. 1173–1185, Mar. 2022.
- [44] C. Wachinger, "Importance driven continual learning for segmentation across domains," in *Proc. Int. Workshop Mach. Learn. Med. Imag.*, vol. 12436. Cham, Switzerland: Springer, 2020, p. 423.
- [45] C. You, J. Xiang, K. Su, X. Zhang, S. Dong, J. Onofrey, L. Staib, and J. S. Duncan, "Incremental learning meets transfer learning: Application to multi-site prostate mri segmentation," in *Proc. Int. Workshop Distrib. Collaborative, Federated Learn.* Cham, Switzerland: Springer, 2022, pp. 3–16.



- [46] J. Xie, S. Yan, and X. He, "General incremental learning with domain-aware categorical representations," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 14331–14340.
- [47] Y. Huang, C. Bert, S. Fischer, M. Schmidt, A. Dörfler, A. Maier, R. Fietkau, and F. Putz, "Continual learning for peer-to-peer federated learning: A study on automated brain metastasis identification," 2022, *arXiv:2204.13591*.
- [48] S. Srivastava, M. Yaqub, K. Nandakumar, Z. Ge, and D. Mahapatra, "Continual domain incremental learning for chest X-ray classification in low-resource clinical settings," in *Proc. MICCAI Workshop Domain Adaptation Represent. Transf.* Cham, Switzerland: Springer, 2021, pp. 226–238.
- [49] A. Thakur, J. Armstrong, A. Youssef, D. Eyre, and D. A. Clifton, "Self-aware SGD: Reliable incremental adaptation framework for clinical AI models," *IEEE J. Biomed. Health Informat.*, vol. 27, no. 3, pp. 1624–1634, Mar. 2023.
- [50] R. Aljundi, F. Babiloni, M. Elhoseiny, M. Rohrbach, and T. Tuytelaars, "Memory aware synapses: Learning what (not) to forget," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 139–154.
- [51] A. Rannen, R. Aljundi, M. B. Blaschko, and T. Tuytelaars, "Encoder based lifelong learning," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1329–1337.
- [52] S.-W. Lee, J.-H. Kim, J. Jun, J.-W. Ha, and B.-T. Zhang, "Overcoming catastrophic forgetting by incremental moment matching," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–11.
- [53] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," in *Proc. NIPS*, 2015, pp. 1–9.
- [54] Y. Le and X. Yang, "Tiny imagenet visual recognition challenge," *CS 231N*, vol. 7, no. 7, p. 3, 2015.
- [55] S. T. Grossberg, *Studies of Mind and Brain: Neural Principles of Learning, Perception, Development, Cognition, and Motor Control*, vol. 70. Cham, Switzerland: Springer, 2012.
- [56] S. Pachade, P. Porwal, D. Thulkar, M. Kokare, G. Deshmukh, V. Sahasrabudhe, L. Giancardo, G. Quellec, and F. Mériaudeau, "Retinal fundus multi-disease image dataset (RFMiD): A dataset for multi-disease detection research," *Data*, vol. 6, no. 2, p. 14, Feb. 2021.
- [57] A.-O. Asia, C.-Z. Zhu, S. A. Althubiti, D. Al-Alimi, Y.-L. Xiao, P.-B. Ouyang, and M. A. A. Al-Qaness, "Detection of diabetic retinopathy in retinal fundus images using CNN classification models," *Electronics*, vol. 11, no. 17, p. 2740, Aug. 2022.
- [58] B. H. Menze, "The multimodal brain tumor image segmentation benchmark (BRATS)," *IEEE Trans. Med. Imag.*, vol. 34, no. 10, pp. 1993–2024, Oct. 2015.
- [59] E. Calabrese, J. E. Villanueva-Meyer, J. D. Rudie, A. M. Rauschecker, U. Baid, S. Bakas, S. Cha, J. T. Mongan, and C. P. Hess, "The university of California san Francisco preoperative diffuse glioma MRI dataset," *Radiology: Artif. Intell.*, vol. 4, no. 6, Nov. 2022, Art. no. e220058.
- [60] S. Bakas, "The university of Pennsylvania glioblastoma (UPenn-GBM) cohort: Advanced MRI, clinical, genomics, & radiomics," *Sci. Data*, vol. 9, no. 1, p. 453, Jul. 2022.
- [61] F. Putz, J. Grigo, T. Weissmann, P. Schubert, D. Hoefler, A. Goma, H. Ben Tkhayat, A. Hagag, S. Lettmaier, B. Frey, U. S. Gaipl, L. V. Distel, S. Semrau, C. Bert, R. Fietkau, and Y. Huang, "The segment anything foundation model achieves favorable brain tumor autosegmentation accuracy on MRI to support radiotherapy treatment planning," 2023, *arXiv:2304.07875*.
- [62] D. Bernhardt, "DEGRO practical guideline for central nervous system radiation necrosis part I: Classification and a multistep approach for diagnosis," *Strahlentherapie und Onkologie*, vol. 198, no. 10, pp. 873–883, Oct. 2022.
- [63] O. Wittenstein, F. Krause, M. Fischer, J. Domschikowski, M. Nitsche, C. Henkenberens, D. Habermehl, and J. Dunst, "The tumor core boost study: A feasibility study of radical dose escalation to the central part of large tumors with an integrated boost in the palliative treatment setting," *Strahlentherapie und Onkologie*, vol. 199, no. 3, pp. 258–267, Mar. 2023.
- [64] M. Buda, A. Saha, and M. A. Mazurkowski, "Association of genomic subtypes of lower-grade gliomas with shape features automatically extracted by a deep learning algorithm," *Comput. Biol. Med.*, vol. 109, pp. 218–225, Jun. 2019.
- [65] G. E. Hinton and S. Roweis, "Stochastic neighbor embedding," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 15, 2002, pp. 1–11.
- [66] Y. Huang, S. Gupta, Z. Song, K. Li, and S. Arora, "Evaluating gradient inversion attacks and defenses in federated learning," in *Proc. Annu. Conf. Neural Inf. Process. Syst. (NeurIPS)*, 2021, pp. 7232–7241.
- [67] Z. He, T. Zhang, and R. B. Lee, "Model inversion attacks against collaborative inference," in *Proc. 35th Annu. Comput. Secur. Appl. Conf.*, Dec. 2019, pp. 148–162.
- [68] J. Jia, A. Salem, M. Backes, Y. Zhang, and N. Z. Gong, "MemGuard: Defending against black-box membership inference attacks via adversarial examples," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, Nov. 2019, pp. 259–274.



**YIXING HUANG** received the B.E. degree from the Department of Biomedical Engineering, Peking University, China, in 2013, the M.S. degree from Pattern Recognition Lab, Department of Computer Science, Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU), in 2016, and the Ph.D. degree from Pattern Recognition Lab in 2020, and did Postdoctoral Research with the Pattern Recognition Lab till April 2021. He is currently a Leading Computer Science Researcher with Department of Radiation Oncology, University Hospital Erlangen, Erlangen, Germany. His research interests include machine learning in medical imaging and radiation oncology.



**CHRISTOPH BERT** received the Ph.D. degree from Technische Universität (TU) Darmstadt, in 2006. Since 2012, he has been a Professor of medical radiation physics with FAU Erlangen-Nürnberg and the Head of medical physics of the Department of Radiation Oncology, University Hospital Erlangen. He studied physics with FAU and Imperial College London. His research focus aims at improving therapy techniques and their quality assurance. Current focus topics are the management of organ motion by tracking, surface-guided radiation therapy (SGRT), MRI in radiation oncology, and error detection in interstitial brachytherapy.



**AHMED GOMAA** received the B.Sc. degree from the Department of Electrical Engineering, King Fahd University of Petroleum and Minerals, Saudi Arabia, in 2017, and the M.Sc. degree in communication and multimedia engineering from FAU, Germany, in 2022, where he is currently pursuing the Ph.D. degree with the Department of Radiation Oncology, University Hospital Erlangen. His research focuses on the application of deep learning in brain medical images.



**RAINER FIETKAU** received the Ph.D. degree from the Institute of Medical Radiation Science, University of Würzburg, in 1985, and the Habilitation degree from the Department of Radiation Oncology, University Hospital Erlangen, FAU, Erlangen, Germany, in 1992. In 1989, he became a Senior Physician with the Department of Radiation Oncology, University Hospital Erlangen, FAU. In 1993, he was the Head of the Clinic and Polyclinic for Radiation Therapy,

University Göttingen. From 1996 to 1997, he was a Professor with the Department of Radiation Oncology, University Hospital Erlangen, FAU. From 1997 to 2008, he became a Professor with the Department of Radiation Oncology, Radiation Therapy and Radiation Biology, University Rostock. He has been a Professor and the Chair of the Department of Radiation Oncology, University Hospital Erlangen, FAU, since 2008. Since 2014, he has been a Speaker with the Research Group for Clinical Application of Hyperthermia in Oncology. Since 2019, he has become the President of German Society for Radiation Oncology (DEGRO). His major scientific research interests include chemoradiation therapy, the combination of radiotherapy with immunotherapy, and hyperthermia.



**ANDREAS MAIER** (Senior Member, IEEE) received the master's degree and the Ph.D. degree in computer science in 2005 and 2009, respectively. From 2005 to 2009, he was with the Pattern Recognition Lab, Computer Science Department, FAU. From 2009 to 2010, he started working on flat-panel C-arm CT as a Postdoctoral Fellow with the Radiological Sciences Laboratory, Department of Radiology, Stanford University. From 2011 to 2012, he joined Siemens Healthcare as the

Innovation Project Manager and was responsible for reconstruction topics in the Angiography and X-ray Business Unit. In 2012, he returned to FAU as

the Head of the Medical Reconstruction Group, Pattern Recognition Lab. In 2015, he became a Professor and the Head of the Pattern Recognition Lab. His major research subject was medical signal processing in speech data. His current research interests include medical imaging, image and audio processing, digital humanities, and interpretable machine learning and the use of known operators. Since 2016, he has been a member of the Steering Committee of European Time Machine Consortium. In 2018, he was awarded an ERC Synergy Grant "4D nanoscope".



**FLORIAN PUTZ** received the degree, in 2012, and the M.D. degree from FAU, in 2015. He is currently a Senior Physician and the Group Leader of the Department of Radiation Oncology, University Hospital Erlangen. He has initiated and is conducting multiple publicly and industrially-funded research projects, including clinical trials, aiming to develop and evaluate deep learning applications for clinical requirements. His current research interests include machine learning in medical

imaging, imaging in radiation oncology, and understanding the interplay between deep learning applications and human experts.

...