Active Deep Learning from a Noisy Teacher for Semi-supervised 3D Image Segmentation: Application to COVID-19 Pneumonia Infection in CT

Mohammad Arafat Hussain[a], Zahra Mirikharaji[a], Mohammad Momeny[b], Mahmoud Marhamati[c], Ali Asghar Neshat[c], Rafeef Garbi[d], Ghassan Hamarneh[a]

[a]*Medical Image Analysis Lab, Simon Fraser University, Burnaby, BC V5A 1S6, Canada*
[b]*Yazd University, Iran*
[c]*Esfarayen Faculty of Medical Science, Iran*
[d]*BiSICL, University of British Columbia, Vancouver, BC V6T 1Z4, Canada*

Abstract

Supervised deep learning has become a standard approach to solving medical image segmentation tasks. However, serious difficulties in attaining pixel-level annotations for sufficiently large volumetric datasets in real-life applications have highlighted the critical need for alternative approaches, such as semi-supervised learning, where model training can leverage small expert-annotated datasets to enable learning from much larger datasets without laborious annotation. Most of the semi-supervised approaches combine expert annotations and machine-generated annotations with equal weights within deep model training, despite the latter annotations being relatively unreliable and likely to affect model optimization negatively. To overcome this,

Email addresses: arafat@ece.ubc.ca (Mohammad Arafat Hussain), zmirikha@sfu.ca (Zahra Mirikharaji), mohamad.momeny@gmail.com (Mohammad Momeny), marhamatim@gmail.com (Mahmoud Marhamati), en.neshat@gmail.com (Ali Asghar Neshat), rafeef@ece.ubc.ca (Rafeef Garbi), hamarneh@sfu.ca (Ghassan Hamarneh)

we propose an active learning approach that uses an example re-weighting strategy, where machine-annotated samples are weighted (i) based on the similarity of their gradient directions of descent to those of expert-annotated data, and (ii) based on the gradient magnitude of the last layer of the deep model. Specifically, we present an active learning strategy with a query function that enables the selection of reliable and more informative samples from machine-annotated batch data generated by a noisy teacher. When validated on clinical COVID-19 CT benchmark data, our method improved the performance of pneumonia infection segmentation compared to the state of the art.

Keywords: Deep learning, semi-supervised learning, active learning, segmentation, noisy teacher, COVID-19, pneumonia

1. Introduction

Supervised learning using deep neural networks has been extensively used for volumetric medical image segmentation in recent years. However, adequate training of deep segmentation models for volumetric medical data, e.g., computed tomography (CT), requires prohibitive amounts of annotated data at the pixel/voxel level that are often very difficult to achieve (Fan et al., 2020) in practical clinical settings. For example, segmentation of pneumonia infection regions in CT can be beneficial as a first step within detection and analysis methods based on convolutional neural networks (CNN) for coronavirus disease 2019 (COVID-19) (Zhou et al., 2020; Gao et al., 2020a; Amyar et al., 2020; Harmon et al., 2020; Zhang et al., 2020; Paluru et al., 2021; Tilborghs et al., 2020; Ghomi et al., 2020; Voulodimos et al., 2021a;

Hasan et al., 2021; Voulodimos et al., 2021b; Ranjbarzadeh et al., 2021; Elharrouss et al., 2020; Singh et al., 2021; Yan et al., 2021). Health condition of many hospitalized COVID-19 patients often deteriorates and requires mechanical ventilation or high-flow oxygenation (Lassau et al., 2021). Therefore, identifying a patient with COVID-19 at increased risk of developing severity can help healthcare professionals plan ahead and make a justified decision on the allocation of resources in the intensive care unit (ICU) (Phua et al., 2020). The severity of patients with COVID-19 is positively correlated with the size and spread of lung infections of various forms (e.g., ground glass opacity (GGO), consolidation, interstitial thickening, air bronchograms, pleural effusion, fibrous strips, etc.) (Xiong et al., 2020). The size and spread of lung infection can be estimated from chest CT scans by segmenting the infection (Wang et al., 2020). However, using supervised learning for COVID-19 infection segmentation is challenged by the scarcity of pixel-level expert-annotated data, hindering the development of models that can reliably perform as well in the wild (i.e., generalize to data from various other sources/scanners) as on the limited data used to train them (Ma et al., 2020a).

To address the complexity of attaining sufficient pixel-level annotated data for supervised learning, various types of weakly supervised learning approaches have been proposed that can be grouped based on the level and/or quality of supervision used (Zhou, 2018): (a) incomplete supervision (Zhang et al., 2019), where only a subset of training data comes paired with labels; (b) inexact supervision (Xu and Lee, 2020), where the training data includes only coarse-grained labels; (c) inaccurate supervision (Zhang

et al., 2019), where the training data includes labels that are not necessarily correct (i.e., noisy). For example, to bypass the lack of sufficient expert-annotated COVID-19 CT data at pixel-level, Laradji et al. (2021) proposed an infection segmentation approach with 'inexact supervision,' where weak annotations, in the form of points, clicked within infection areas, were instead used. Nonetheless, even such weak supervision remains challenging and time-consuming, as annotators must go through all image slices of volumetric medical data for manual labeling. In addition, intra- and inter-annotator reproducibility are often poor, which causes variability in the manually annotated segmentation masks.

Alternatively, in semi-supervised learning, e.g., (Cheplygina et al., 2019; Abdel-Basset et al., 2020; Yang et al., 2021; Li et al., 2021), limited pixel-level expert-annotated data are combined with a large pool of machine-annotated data for training deep segmentation models. Furthermore, most semi-supervised learning approaches were designed for only 2D data, for example (Fan et al., 2020; Ma et al., 2020a; Laradji et al., 2021; Abdel-Basset et al., 2020; Yang et al., 2021; Taghanaki et al., 2019a; Souly et al., 2017; Lee et al., 2019; Hong et al., 2015; Wang et al., 2020), however, native 3D segmentation would be superior to the reconstruction of 3D segmentation from stacked slice-based 2D segmentation masks, as native 3D analysis captures the true 3D spatial context of the underlying structures within the imaged field of view (Çiçek et al., 2016). In addition, stacked slice-based 2D segmentation approaches assume that the slice thickness (i.e., the distance between two axial slices) is isotropic, but on many occasions, it may not be true.

Machine-generated annotations (i.e., pseudo labels) are generally less re-

liable and typically require further correction by experts (Marzahl et al., 2020). Thus, uncorrected (by experts) machine-generated annotations are likely to lead to incorrect predictions being reinforced during network optimization, which in turn leads to worse task performance at test time. To address this problem, a semi-supervised active learning (SSAL) strategy is sometimes used, which generally uses a pipeline of (i) query function for selecting "informative" samples from the annotation-free data pools, (ii) forwarding those to oracle annotators for generating ground truth annotation, and subsequently (iii) adding those new annotated data to the training data pool (Zhao et al., 2021; Gao et al., 2020b; Calma et al., 2018; Lv et al., 2022; Bull et al., 2018). However, such oracle annotation systems share limitations similar to those of expert supervision in medical imaging applications, namely, the time and labor requirements placed upon expert radiologists who are rarely available or interested in such manual dense annotation tasks, as well as the poor intra- and inter-annotator reproducibility.

To overcome the challenge of producing time and labor-intensive expert annotations for informative 3D volumetric medical images from the annotation-free data pools in the active learning framework, we propose a sample re-weighting-based (Xu et al., 2021) (i.e., a way to emphasize and pick informative data only during training) semi-supervised learning approach, namely, 'SSAL from a noisy teacher,' and showcase its utility for pneumonia infection segmentation in clinical CT scans of COVID-19 patients. The proposed method uses alternatives to the conventional human oracle-based active learning steps discussed in the previous paragraph. The proposed method consists of several steps. First, it generates voxel-level annotations

(pseudo-annotation) using supervised deep learning. Second, since machine-generated annotations are less reliable, we generate gradient-based relative sample weights that reflect the "trustworthiness" of the samples during training. These sample weights are estimated from the similarity of the gradient directions between the annotation-free batch data and the expert-annotated validation data. Third, because the weighting of samples based on the gradient similarity may lead to an underestimation of a more diversely informative data sample, we adopt a gradient magnitude-based strategy (Ash et al., 2019) to generate another set of sample weights that reflect the "informativeness" of the samples during training. Fourth, we generate an overall sample weight by combining the sample weights of "trustworthiness" and "informativeness." Finally, we use a query mechanism to choose more informative and trustworthy samples in a batch of annotation-free data by rectification of the combined weight per sample, and subsequently use these combined sample weights during the model optimization.

Our sample re-weighting based adaptive data sampling strategy can be viewed as a pool-based SSAL strategy, in which a few annotation-free training samples are adaptively chosen in each training cycle based on some preset criteria and presented for annotation to an oracle annotator, i.e., an expert teacher, albeit a noisy teacher in our case. A concise summary of the contributions of this paper is:

1. We propose active learning from the noisy teacher approach that uses an example re-weighting strategy in using expert-annotation-less data in deep model training.
2. Our re-weighting strategy uses 'gradient similarity' and 'gradient mag-

nitude' in determining the sample weights to reflect the 'trustworthiness' and the 'informativeness,' respectively, of machine-annotated data.

3. Our active learning strategy uses a query function that enables the selection of reliable and more informative samples from machine-annotated batch data.
4. We show the effectiveness of our proposed approach on clinical COVID-19 CT benchmark CT data.

2. Methodology

2.1. Method Overview

We design the working pipeline of the proposed method using the following steps:

1. Initially, we generate voxel-level annotations (pseudo annotation) using supervised deep models while considering the fact that these machine-generated annotations are less reliable (noisy teacher) than human expert annotations.
2. Then we generate a relative weight based on gradients per sample based on its "trustworthiness" during training. A sample weight is estimated from the similarity of the gradient directions between the annotation-free sample data and the expert-annotated validation data. Gradient similarity-based sample re-weighting approaches have previously been explored for deep learning from inaccurate labels on 2D RGB images (Ren et al., 2018; Mirikharaji et al., 2019), however, they have not been explored for volumetric radiographic images.

3. As the primary aim of an active learning algorithm is to identify and label only maximally-informative samples, gradient similarity-based sample weighting may lead to underestimation of a more diversely informative data sample. Ash et al. (2019) showed the efficacy of using the gradient magnitude, with respect to parameters in the final CNN layer, as a measure of the model's uncertainty. The higher magnitude of the gradient of the last layer, resulting from a higher loss of training, implies that the interrogated training sample contains newer information (Ash et al., 2019) that the model has not yet seen. In our proposed approach, we adopt this gradient magnitude-based strategy and generate another set of sample weights based on their "informativeness" during training.
4. Afterwards, we generate an overall sample weight by combining the "trustworthiness" and "informativeness" sample weights.
5. Finally, we use a query mechanism to choose more informative and trustworthy samples in a batch of annotation-free data by rectification (i.e., choosing more useful data in a batch) of the combined sample weight, and subsequently use these combined sample weights in the batch during the model optimization.

2.2. Data Partition

Our method uses a small set of expert-annotated volumetric imaging data $(X_e, Y_e) : \{(x^i, y^i); 1 \leq i \leq N_e\}$ to produce pseudo-annotations for a much larger set of data $X_p : \{x_p^m; 1 \leq m \leq N_p\}$ lacking annotations Y_p. We start by dividing the small cohort of expert-annotated data into training set $(X_{tr}, Y_{tr}) : \{(x_{tr}^j, y_{tr}^j); 1 \leq j \leq N_{tr}\}$, validation set $(X_v, Y_v) : \{(x_v^k, y_v^k); 1 \leq k \leq N_v\}$, and hold out test set $(X_t, Y_t) : \{(x_t^l, y_t^l); 1 \leq l \leq N_t\}$, where

$N_{tr} + N_v + N_t = N_e$ and $N_e \ll N_p$. We then train a deep 3D segmentation model on this small training set and use the resulting model to generate pseudo-annotations Y_p on the large pool of annotation-free scans. We then combine these generated pseudo-annotated data with the expert-annotated validation data to train an example re-weighted active learning model, in which we assign varying weights to each pseudo-annotated training example based on its gradient direction (see Fig. 1).

2.3. Supervised Learning for Pseudo-Label Generation

Using the training set (X_{tr}, Y_{tr}), our objective is to train a 3D CNN $\Phi_{sl}(X_{tr}, \theta) : x_{tr} \rightarrow y_{tr}$ in a fully supervised fashion (Step 1 in Fig. 1) which is then used to generate pseudo-labels Y_p for X_p, where θ is the set of learnable parameters. For model optimization, we adopt the combo loss (Taghanaki et al., 2019b) after we extend it with a sample-specific weighting:

$$\begin{aligned} \mathcal{L} &= \sum_{c=1}^{C} w_c \mathcal{L}_c \\ &= \sum_{c=1}^{C} w_c \left(1 - \frac{2 \sum_{b=1}^{B} \sum_{d=1}^{D} p_b^d y_{tr,b}^d}{\sum_{b=1}^{B} \sum_{d=1}^{D} (p_b^d + y_{tr,b}^d)} - \frac{1}{D} \sum_{d=1}^{D} \sum_{b=1}^{B} y_{tr,b}^d \log(p_b^d) \right), \end{aligned} \tag{1}$$

where C is the number of volumes in the training mini-batches, w_c is the scalar weight associated with each training example, B is the total number of class labels, D is the total number of voxels in each volume, and $p_b = Pr(y_{tr,b}^d | x_{tr,b}^d; \theta)$ are the probability vectors of the output label corresponding to the vector encoded ground truth label vectors $y_{tr,b}$. For $\Phi_{sl}(X_{tr}, \theta)$, we use $w_c = 1$. After training Φ_{sl}, we generate pseudo-labels for x_p as $y_p = \Phi_{sl}(x_p, \theta)$ (Step 2 in Fig. 1).

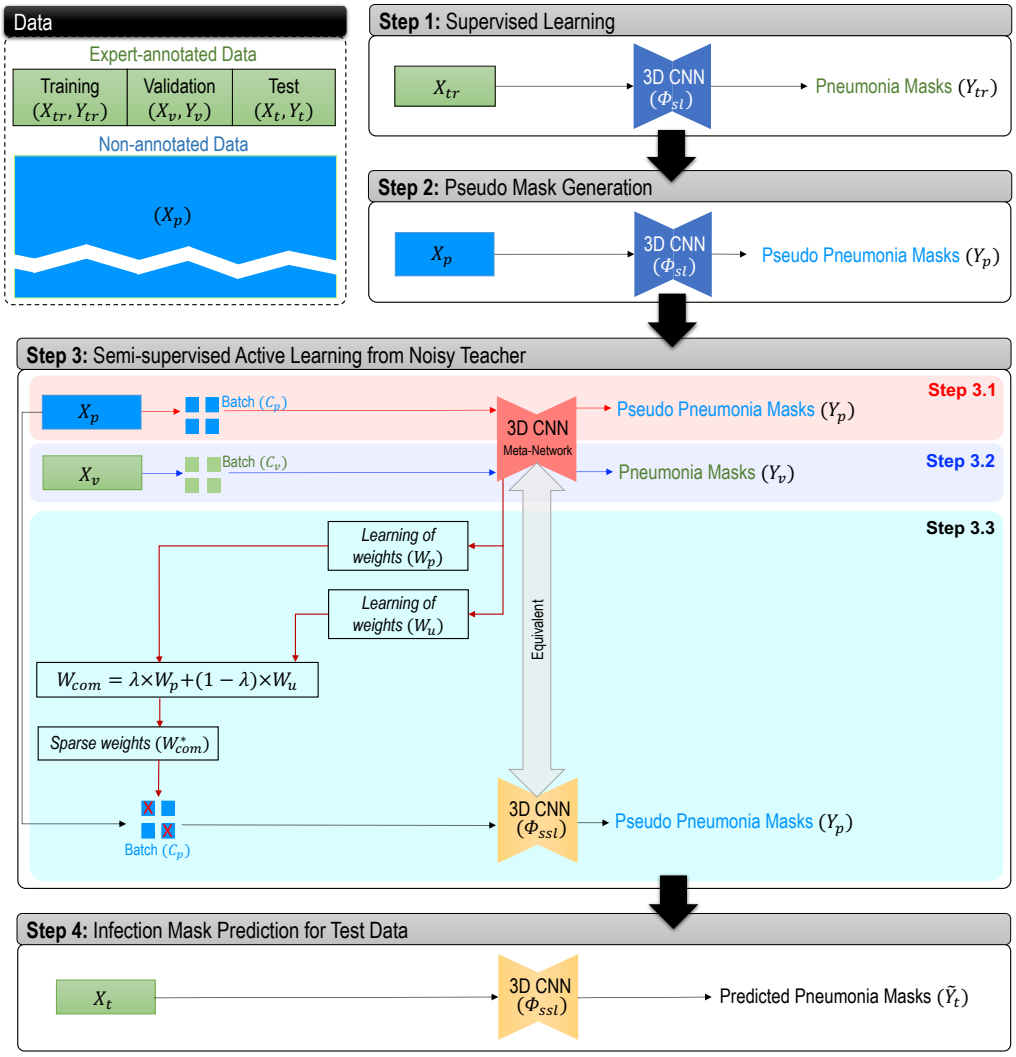

Figure 1: Schematic diagram of the proposed example re-weighted active learning from noisy teacher approach for 3D image segmentation.

2.4. Semi-supervised Active Learning from Noisy Teacher

Our active learning strategy aims to assign weights to pseudo-annotated data samples and subsequently select more reliable examples for training mini-batch. We train a 3D CNN, $\Phi_{ssl}(X, \theta) : X \in X_v \cup X_p$ in a semi-

supervised fashion (Step 3 in Fig. 1). In a standard training loss function $\mathcal{L}$ (in our case, the combo loss in Eq. 1), all input data are equally weighted, i.e., $w_c = 1$ in Eq. 1. Given that our expanded training dataset now consists of both an expert-annotated set (X_v, Y_v) and a pseudo-annotated set (X_p, Y_p), where Y_p is generally less reliable than Y_v, our method instead learns a data re-weighting strategy, where we minimize the weighted combo loss $(\mathcal{L}^p)$ for a mini-batch (C_p) of set (X_p, Y_p) as follows:

$$\theta^*(w) = \underset{\theta}{\operatorname{argmin}} \sum_{c=1}^{C_p} w_c \mathcal{L}_c^p(x_p, y_p; \theta). \tag{2}$$

In deep learning, the model parameters θ are updated using gradient descent:

$$\hat{\theta}_{t+1}(w) = \theta_t(w) - \alpha \nabla \left(\sum_{c=1}^{C_p} w_c \mathcal{L}_c^p(x_p, y_p; \theta)|_{\theta=\theta_t} \right), \tag{3}$$

where α is the step size and t denotes the training step. Given the reduced trustworthiness of the generated pseudo labels, our method inspects the gradient descent direction of each training minibatch of the pseudo-annotated dataset in each iteration, on the training loss surface. Instead of equal weighting, we reweight $(W_p = \{w_1, w_2, ..., w_{C_p}\})$ according to their similarity to the descent direction of the validation combo loss $(\mathcal{L}^v)$ surface as:

$$W_{p,t}^* = \underset{W_p, W_p \geq 0}{\operatorname{argmin}} \frac{1}{N_v} \sum_{n=1}^{N_v} \mathcal{L}_n^v(x_v, y_v; \theta_{t+1}(w)). \tag{4}$$

However, solving Eq. 4 to find optimal W_p for each update step of the network parameters θ is computationally intensive as it requires two nested loops of optimization. Therefore, to approximate W_p for every gradient descent step, a meta-learning procedure is used. At step t, we perform a single gradient

descent step on a mini-batch C_v of validation samples (shown as Step 3.2 in Fig. 1) with respect to $W_{p,t}$ (obtained in Step 3.1 in Fig. 1), followed by rectifying the output to generate a non-negative weight (shown as Step 3.3 in Fig. 1):

$$\tilde{W}_{p,t} = g(\max\left(0, -\eta \frac{\partial}{\partial W_{p,t}} \sum_{n=1}^{C_v} \mathcal{L}_n^v(x_v, y_v; \theta_{t+1}(w))|_{W_{p,t}=0}\right)), \tag{5}$$

where η is the descent step size on $W_{p,t}$, and g is the normalization function to ensure $\sum_{c=1}^{C_p} w_c = 1$.

We also estimate the embedding of the gradient $g_{em,t}$ in step t for pseudo-annotated data samples from the magnitude of the gradient with respect to parameters θ_{out} in the final layer of the meta-network as (Ash et al., 2019):

$$g_{em,t} = \frac{\partial}{\partial \theta_{out}} \left(\sum_{c=1}^{C_p} \mathcal{L}_c^p(x_p, y_p; \theta)|_{\theta=\theta_t} \right). \tag{6}$$

We then generate the gradient magnitude-based sample weights from the gradient embedding $g_{em,t}$ (Eq. 6) as:

$$\tilde{W}_{u,t} = g(\max(0, g_{em,t})). \tag{7}$$

To ensure a balance between the "trustworthiness" and "informativeness" of a particular pseudo-annotated data sample, we combine $\tilde{W}_{p,t}$ and $\tilde{W}_{u,t}$ using a relative weight λ as:

$$\tilde{W}_{com,t} = \lambda \tilde{W}_{p,t} + (1 - \lambda)\tilde{W}_{u,t}. \tag{8}$$

Although the optimum set of weights $\tilde{W}_{com,t}$ for a mini-batch is expected to have a positive value, which allows all the image volumes in the mini-batch to contribute to the optimization of θ of Φ_{ssl}, the constituent weight

per sample in $\tilde{W}_{com,t}$ is different based on the sample's descent direction similarity to those of the validation data, as well as its gradient embedding $g_{em,t}$. Therefore, we choose more reliable and informative pseudo-annotated samples from the mini-batch whose descent direction is most similar to those of the validation data and which introduces newer information. This approach makes the optimization more robust on the noisy pseudo-annotated samples and also mimics an active learning strategy. To achieve this, we further rectify $\tilde{W}_{com,t}$ so that samples with weights greater than the uniform value $(1/C_p)$ are selected as:

$$\tilde{W}^*_{com,t} = \begin{cases} \tilde{W}_{com,t} & \text{if } \tilde{W}_{com,t} \geq 1/C_p \\ 0 & \text{otherwise} \end{cases}$$

After learning the adaptive weights ($\tilde{W}^*_{com} = \{w^*_1, w^*_2, ..., w^*_{C_p}\}$), we perform a final backward pass to estimate the gradient and update the network parameters as:

$$\theta_{t+1}(w) = \theta_t(w) - \alpha \nabla \left(\sum_{c=1}^{C_p} w^*_c \mathcal{L}^p_c(x_p, y_p; \theta)|_{\theta=\theta_t} \right). \tag{9}$$

We also present a description of our method in Algorithm 1. This algorithm describes the technical steps in each training iteration. The iteration starts by loading the pseudo- and expert-annotated batch data (lines 1 and 2, respectively). Then the 3D meta-network loads current parameters from the main 3D CNN model of identical architecture (line 4). Afterward, steps to learn the weights per pseudo-annotated sample based on its gradient similarity and gradient magnitude are shown in lines 5-16. Finally, the sample re-weighted loss calculation and updating of parameters of the main 3D CNN

model are shown in lines 17-20.

3. Data

We used three clinical COVID-19 CT datasets to evaluate our proposed method, two of which are publicly available and include expert annotations at the voxel-level of infections. The third database is private and is accessed with proper institutional ethics approval (IR.ESFARAYENUMS.REC.1398.019, Esfarayen Faculty of Medical Sciences, 2020-03-18; 2020s0128, Simon Fraser University).

3.1. COVID-19 CT Benchmark Dataset

The first public database is the COVID-19 CT Benchmark dataset of (Ma et al., 2020b) (henceforth, "Benchmark" refers to these data), which contains 20 CT volumes from 20 patients with expert annotations at the voxel-level of COVID-19 infection in the lungs. The proportion of COVID-19 pneumonia infection in the lungs ranges from 0.01% to 59%. The left and right lungs, and pneumonia infection in these data were annotated in three steps: first, junior annotators with 1-5 years of experience annotated the data slice-by-slice using ITKSnap in the axial direction, which was refined by two radiologists with 5-10 years experience. Finally, a senior radiologist with more than 10 years of experience verified and refined the annotations.

3.2. COVID-19 Lung CT Lesion Segmentation Challenge 2020 Dataset

The second public database is the COVID-19 Lung CT Lesion Segmentation Challenge 2020 database of (An et al., 2020) (henceforth, "Challenge" refers to these data), which contains 199 chest CT scans from 199 patients

Algorithm 1 Learning to Re-weight Pseudo-annotated Samples

Require: θ_0, (X_v, Y_v), (X_p, Y_p), C_v, C_p, λ

Ensure: θ_T

for $t = 0...T-1$ **do**

$\{x_p, y_p\} \leftarrow \text{MiniBatch}((X_p, Y_p), C_p)$ ▷ Pseudo-annotated data

$\{x_v, y_v\} \leftarrow \text{MiniBatch}((X_v, Y_v), C_v)$ ▷ Expert-annotated data

$\theta_t^m \leftarrow \theta_t$ ▷ Assigning model parameters at t to meta-net

$\hat{y}_p \leftarrow \text{Forward}(x_p, y_p, \theta_t^m)$ ▷ Forward pass of pseudo-annotated data

$w \leftarrow 0; l_p \leftarrow \sum_{i=1}^{C_p} w_i \mathcal{L}_n^p(y_{p,i}, \hat{y}_{p,i})$ ▷ Loss calculation; refers to Eq. 2

$\nabla\theta_t^m \leftarrow \text{BackwardAD}(l_p, \theta_t^m)$ ▷ Backward automatic differentiation

$\hat{\theta}_t^m \leftarrow \theta_t^m - \alpha\nabla\theta_t^m$ ▷ Updating the meta-net parameters

$\hat{y}_v \leftarrow \text{Forward}(x_v, y_v, \hat{\theta}_t^m)$ ▷ Forward pass of expert-annotated data

$l_v \leftarrow \frac{1}{C_v} \sum_{i=1}^{C_v} \mathcal{L}_n^v(y_{v,i}, \hat{y}_{v,i})$ ▷ Loss calculation; refers to Eq. 4

$\nabla w \leftarrow \text{BackwardAD}(l_v, w)$ ▷ Backward automatic differentiation

$\tilde{w}_p \leftarrow \max(-\nabla w, 0); \tilde{W}_p \leftarrow \frac{\tilde{w}_p}{\sum \tilde{w}_p + \delta(\sum_j \tilde{w}_{p,j})}$ ▷ Refers to Eq. 5

$g_{em} = \frac{\delta}{\delta\theta_{out}^m} l_p$ ▷ θ_{out}^m: parameters of the output layer; refers to Eq. 6

$\tilde{w}_u \leftarrow \max(g_{em}, 0); \tilde{W}_u \leftarrow \frac{\tilde{w}_u}{\sum \tilde{w}_u + \delta(\sum_j \tilde{w}_{u,j})}$ ▷ Refers to Eq. 7

$\tilde{W}_{com} \leftarrow \lambda\tilde{W}_p + (1-\lambda)\tilde{W}_u$ ▷ Refers to Eq. 8

$\tilde{W}_{com}^* \leftarrow \tilde{W}_{com}$ if $\tilde{W}_{com} \geq 1/C_p$; else 0 ▷ Rectification od weights

$\hat{y}_p \leftarrow \text{Forward}(x_p, y_p, \theta_t)$ ▷ Forward pass to actual model in training

$\tilde{l}_p \leftarrow \sum_{i=1}^{C_p} \tilde{w}_{com,i}^* \mathcal{L}_n^p(y_{p,i}, \hat{y}_{p,i})$ ▷ Loss calculation with re-weighting

$\nabla\theta_t \leftarrow \text{BackwardAD}(\tilde{l}_p, \theta_t)$ ▷ Backward automatic differentiation

$\theta_{t+1} \leftarrow \theta_t - \alpha\nabla\theta_t$ ▷ Updating the model parameters; refers to Eq. 9

end for

with ground truth pixel-level annotations of COVID-19 lesions in the lungs. These data were acquired without intravenous contrast enhancement from COVID-19 patients confirmed by reverse transcription polymerase chain reaction (RT-PCR) in China. COVID-19 infection in these CT volumes was initially segmented using a previously trained model to segment the COVID-19 lesion (Yang et al., 2021). Later, a group of experienced radiologists used the initial segmentation as a starting point for the subsequent ITKSnap-based adjudication and correction of infection masks.

3.3. COVID-19 CT Private Dataset

The third database is composed of 1,473 CT scans of 623 patients of Imam Khomeini Hospital, Esfarayen, Iran (henceforth, "hospital" refers to these data). We accessed these data with all required ethics approvals in place (IR.ESFARAYENUMS.REC.1398.019, Esfarayen Faculty of Medical Sciences, 2020-03-18; 2020s0128, Simon Fraser University). Of these scans, 567 were confirmed to be from COVID-19 patients and 906 are from non-COVID patients. None of the scans in this third database had pixel-level annotation of lung infections. These data were acquired using the Toshiba Alexion CT scanner (Toshiba, Minato City, Tokyo, Japan). The axial pixel dimension ranges between 0.571 and 0.763 mm. The thickness of the slice was set to 7 mm.

4. Implementation Details

To standardize the clinical datasets, we resampled all CT volumes to have a common voxel dimension of $1.6\times1.6\times3.2$ mm^3 by trilinear interpolation. We used a modified version of 3D UNet (Çiçek et al., 2016) as CNN (for

both Φ_{sl} and Φ_{ssl}), which has residual connections around the convolutional blocks as in (Kerfoot et al., 2018). We show our 3D UNet architecture as a tabular form in Table 1, which also mentions the number of trainable parameters in each layer of the network. We use Adam optimizer with an initial learning rate of 0.01 to train our networks. We carried out two experiments, one using Benchmark and Hospital data and the other using Challenge and Hospital data. We used Challenge and Hospital data for the ablation study (i.e., experiment 2) and the Benchmark and Hospital data for the comparison of our performance with respect to the state-of-the-art approaches (i.e., experiment 1). In both cases, the hospital scans constituted the data without voxel-level annotation (i.e., (X_p, Y_p)). Before running the experiments, we augmented the Benchmark and Challenge datasets by flipping left-right and up-down (i.e., the sample size was increased 4×). For experiment 1, we used 16 volumes from the Benchmark dataset as X_{tr} during supervised learning (Step 1 in Fig. 1). We used the remaining 64 Benchmark CT volumes in 4-fold cross-validation, where we used 48 volumes as X_v (i.e., used in gradient similarity estimation during training) and 16 volumes as X_t (i.e., held out for testing) in each fold. For experiment 2, we used the pseudo annotations from experiment 1, thus used 636 (i.e., 4×159; used in gradient similarity estimation during training) volumes as X_v and 160 (i.e., 4×40; held out for testing) volumes as X_t in 5-fold cross-validation. In both experiments, the entire third dataset was used as annotation-free data X_p. We ensured that the augmented data of a patient was never split between the training, validation, and test sets. Training Φ_{sl} and Φ_{ssl} are scheduled to run for 500 epochs each in both experiments, which is supposed to take about 1 day

and 8 days, respectively. However, we often stopped training early if the training error was found to be saturated. We also chose $C_v = C_p = 4$ (i.e., the size of the training batch) and $\lambda = 0.5$. We implemented our method in PyTorch version 1.6.0 and Python version 3.7.4. The training was performed on a workstation with an Intel E5-2650 v4 Broadwell 2.2 GHz processor, an Nvidia P100 Pascal GPU with 16 GB of VRAM, and 64 GB of RAM.

5. Results and Discussion

In this section, we first present our ablation study on the Challenge dataset and then compare the pneumonia infection segmentation performance of our proposed method with that of the state-of-the-art methods on the Benchmark dataset.

5.1. Ablation Study on the Challenge Data

Here, we present the results of our ablation study in Table 2 to demonstrate the incremental contributions by different modules of our proposed method. In this table, we present 5-fold cross-validated Dice scores and Hausdorff distances by different approaches such as our fully supervised method, a semi-supervised approach adopting the training strategy from (Fan et al., 2020), the proposed gradient similarity-based sample re-weighting method (RGS; re-weighting with $\lambda = 1$ but without rectification of $\tilde{W}_{com,t}$), the proposed last layer gradient magnitude-based sample re-weighting method (RGM; re-weighting with $\lambda = 0$ but without rectification of $\tilde{W}_{com,t}$), proposed gradient similarity and last layer gradient magnitude-based sample re-weighting (RGS&M; re-weighting with $\lambda = 0.5$ but without rectification of $\tilde{W}_{com,t}$), and the proposed gradient similarity- and last layer gradient

Table 1 Our 3D UNet architecture. Any number with * represents the skip connection between the encoder and decoder sides of the network. Acronyms-BN: batch normalization, Conv3D: 3D convolution, Conv3D-Res: Conv3D is used for residual connection, PReLU: parametric rectified linear unit, and I: identity connection.

Block type	Conv3D Kernel	Stride	Activation function	BN	Repeat	Input size	Output size	# trainable parameters
Conv3D	3^3	2	PReLU	Yes	2	$96^3 \times 1$	$48^3 \times 16$	897
Conv3D	3^3	1	PReLU	Yes	1	$48^3 \times 16$	$48^3 \times 16$	6,929
Conv3D-Res	3^3	2	-	-	1	$48^3 \times 16$	$48^3 \times 16$	0
Conv3D	3^3	2	PReLU	Yes	2	$48^3 \times 16$	$24^3 \times 32$	27,713
Conv3D	3^3	1	PReLU	Yes	1	$24^3 \times 32$	$24^3 \times 32$	27,681
Conv3D-Res	3^3	2	-	-	1	$24^3 \times 32$	$24^3 \times 32$	0
Conv3D	3^3	2	PReLU	Yes	2	$24^3 \times 32$	$12^3 \times 64$	110,720
Conv3D	3^3	1	PReLU	Yes	1	$12^3 \times 64$	$12^3 \times 64$	110,657
Conv3D-Res	3^3	2	-	-	1	$12^3 \times 64$	$12^3 \times 64$	0
Conv3D	3^3	2	PReLU	Yes	2	$12^3 \times 64$	$6^3 \times 128$	442,625
Conv3D	3^3	1	PReLU	Yes	1	$6^3 \times 128$	$6^3 \times 128$	442,497
Conv3D-Res	3^3	2	-	-	1	$6^3 \times 128$	$6^3 \times 128$	0
Conv3D	3^3	1	PReLU	Yes	2	$6^3 \times 128$	$6^3 \times 256$	918,017
Conv3D	3^3	1	PReLU	Yes	1	$6^3 \times 256$	$6^3 \times 256$	1,769,729
Conv3D-Res	1^3	1	-	-	1	$6^3 \times 256$	$6^3 \times 256$	0
Conv3D	3^3	2	PReLU	Yes	1	$6^3 \times (256+128^*)$	$12^3 \times 64$	663,617
Conv3D	3^3	1	PReLU	Yes	1	$12^3 \times 64$	$12^3 \times 64$	110,657
Conv3D-Res	I	-	-	-	1	$12^3 \times 64$	$12^3 \times 64$	0
Conv3D	3^3	2	PReLU	Yes	1	$12^3 \times (64+64^*)$	$24^3 \times 32$	110,621
Conv3D	3^3	1	PReLU	Yes	1	$24^3 \times 32$	$24^3 \times 32$	27,681
Conv3D-Res	I	-	-	-	1	$24^3 \times 32$	$24^3 \times 32$	0
Conv3D	3^3	2	PReLU	Yes	1	$24^3 \times (32+32^*)$	$48^3 \times 16$	27,665
Conv3D	3^3	1	PReLU	Yes	1	$48^3 \times 16$	$48^3 \times 16$	6,929
Conv3D-Res	I	-	-	-	1	$48^3 \times 16$	$48^3 \times 16$	0
Conv3D	3^3	2	PReLU	Yes	1	$48^3 \times (16+16^*)$	$96^3 \times 2$	1,731
Conv3D	3^3	1	-	-	1	$96^3 \times 2$	$96^3 \times 2$	110
Conv3D-Res	I	-	-	-	1	$96^3 \times 2$	$96^3 \times 2$	0
							Total	**4,806,481**

magnitude-based sample re-weighting with AL method (RGS&M+AL; re-weighting with $\tilde{W}^*_{com,t}$ and $\lambda = 0.5$). Note that for the semi-supervised approach, we only adopted the semi-supervision strategy from (Fan et al., 2020) but not their deep model, as it was designed for 2D data. This semi-

Table 2 5-fold cross-validation performance in terms of Dice scores and Hausdorff distances using our methods implemented for segmenting COVID-19 pneumonia infection in **Challenge data**. The upward arrow (↑) indicates that 'higher is better, and the downward arrow (↓) indicates that 'lower is better.' Values indicated by the colors blue and red indicate the best performance in terms of the Dice score and the Hausdorff distance, respectively. (Acronyms: RGS: sample re-weighting based on gradient similarity only, RGS+AL: sample re-weighting based on gradient similarity followed by active learning, RGS&M+AL: sample re-weighting based on both gradient similarity and last layer gradient magnitude followed by active learning, Met: metrics, DS: Dice score, HD: Hausdorff distance.)

Method	Met.	Fold-1	Fold-2	Fold-3	Fold-4	Fold-5	Average
Fully Supervised	DS↑	0.573±0.240	0.565±0.236	0.589±0.225	0.642±0.214	0.573±0.229	0.588±0.228
	HD↓	35.68±10.27	37.50±09.54	34.71±10.50	32.80±13.28	34.41±11.23	35.02±10.96
Semi-supervised	DS↑	0.574±0.228	0.577±0.256	0.588±0.223	0.637±0.211	0.576±0.220	0.590±0.227
	HD↓	35.50±09.13	35.05±10.45	38.59±11.66	33.21±11.13	33.58±11.20	35.18±10.71
Proposed RGS	DS↑	0.588±0.215	0.583±0.231	0.592±0.217	0.644±0.220	0.583±0.214	0.598±0.219
	HD↓	37.52±09.51	36.06±08.61	35.13±11.99	33.58±10.64	34.26±11.35	35.31±10.42
Proposed RGM	DS↑	0.592±0.228	0.583±0.243	0.626±0.208	0.657±0.198	0.577±0.226	0.607±0.221
	HD↓	33.59±10.33	34.85±11.71	36.85±11.72	33.86±10.73	37.22±09.66	35.27±10.83
Proposed RGS&M	DS↑	0.593±0.224	0.604±0.232	0.622±0.203	0.660±0.203	0.599±0.217	0.616±0.216
	HD↓	34.06±09.06	36.06±09.18	35.54±12.71	32.37±12.56	36.19±10.63	34.84±10.82
Proposed RGS&M+AL	DS↑	0.600±0.224	0.613±0.222	0.631±0.197	0.661±0.200	0.600±0.210	0.621±0.211
	HD↓	32.33±11.55	33.28±10.23	34.37±12.39	34.91±10.34	37.36±11.12	34.45±11.12

supervision strategy used machine-generated annotation (i.e., noisy annotation) for the annotation-free data pool during training. As a result, we see in Table 2 that the performance of this approach in terms of the mean Dice and mean Hausdorff distance is worse than that of our fully supervised approach for folds 3 and 4. The overall performance of the semi-supervised method is slightly better than that of the fully supervised approach in terms of average

Dice; however, the opposite is seen in terms of the mean Hausdorff distance. Next, we see in Table 2 that the sample re-weighting strategy using the gradient similarity and the last-layer gradient magnitude (used in the RGS and RGM methods, respectively) leads to better segmentation performance in terms of Dice score than fully supervised and semi-supervised approaches because of incorporating the "trustworthiness" (in terms of $\tilde{W}_{p,t}$) and "informativeness" (in terms of $\tilde{W}_{u,t}$) of pseudo-annotated samples into the model loss calculation. However, the Hausdorff distance by the RGS method is seen to be worse than that by the semi-supervised approach for folds 1, 2, 4, and 5. A similar but marginally worse performance trend is seen in the case of the RGM approach for folds 4 and 5. We further see in Table 2 that the RGS&M method, where we combined the similarity of the gradients and the magnitude of the gradient of the last layer with $\lambda = 0.5$, improves the segmentation performance in terms of the Dice score more than the RGS or RGM method alone. However, mixed performance is seen in terms of the Hausdorff distance for folds 1 to 5, compared to semi-supervised, RGS, and RGM approaches. Nonetheless, the average Hausdorff distance performance by the RGS&M method is better than that by the fully supervised, semi-supervised, RGS, and RGM approaches. Finally, we see in Table 3 that our active learning strategy using weight rectification (i.e., $\tilde{W}^{*}_{com,t}$ with $\lambda = 0.5$), which completely removes the contribution of less trustworthy and informative samples in batch-wise loss calculation, leads to the best segmentation performance in terms of the Dice score. Furthermore, the average Hausdorff distance by the proposed RGS&M+AL method outperforms all other methods. We also performed the two-sample t-test for the 5-fold mean Dice

scores between our proposed RGS&M+AL and other methods mentioned in Table 2. The estimated p-values are 0.0027, 0.0048, 0.0330, 0.1963, and 0.6404 for the fully supervised, semi-supervised, RGS, RGM, and RGS&M methods, respectively. Since, RGS, RGM, and RGS&M methods are intrinsic parts of our proposed RGS&M+AL method, differences in segmentation performance by these approaches may be non-significant. However, as expected, the proposed RGS&M+AL showed statistically significant improvements ($p < 0.01$) in terms of Dice score compared to fully supervised and semi-supervised approaches.

In Fig. 2, we demonstrate the qualitative performance comparison of fully supervised, semi-supervised, RGS, RGM, RGS&M, and RGS&M+AL methods. Here, we show the axial CT slices and corresponding expert-annotated pneumonia infection masks for seven COVID-19-positive patients. We see in this figure that all methods performed reasonably well in identifying pneumonia infections in the lung. However, for more irregular and complex infection patterns (e.g., patients I, IV, V, and VII), the masks produced by the proposed RGS&M+AL method, shown in the last row, are the best match to the expert-annotated infection masks, shown in the second row. We also see in the last three columns that there are considerable false positives (indicated with blue arrows) and false negatives (indicated with yellow arrows) in infection masks produced by different methods except for the proposed RGS&M+AL approach. This evidence further supports the efficacy of the proposed RGS&M+AL method.

Figure 2: Qualitative performance comparison by our implemented methods in pneumonia infection segmentation on the **Challenge data**. The first row shows the axial CT slices of seven COVID-19-infected patients. The second row shows the expert-generated infection mask overlaid on the corresponding CT slices. The third to eighth rows show infection segmentation masks generated by different approaches. Blue arrows indicate false positives and yellow arrows indicate false negatives.

5.2. Performance Comparison on the Benchmark Data

In Table 3, we show the Dice scores and Hausdorff distances achieved in 4-fold cross-validation by our fully supervised method, proposed gradient

Table 3 4-fold cross-validation performance in terms of Dice scores and Hausdorff distances by our implemented methods for segmenting pneumonia infection in the **Benchmark data**. The upward arrow (↑) indicates that higher is better, and the downward arrow (↓) indicates that lower is better. The values indicated by the colors blue and red indicate the best performance in terms of dice score and Hausdorff distance, respectively. (Acronyms: RGS: sample re-weighting based on gradient similarity only, RGS+AL: sample re-weighting based on gradient similarity followed by active learning, RGS&M+AL: sample re-weighting based on both gradient similarity and last layer gradient magnitude followed by active learning, Met: metrics, DS: Dice score, HD: Hausdorff distance.)

Method	Met.	Fold-1	Fold-2	Fold-3	Fold-4	Average
Fully Supervised	DS↑	0.751±0.075	0.696±0.113	0.705±0.038	0.768±0.038	0.730±0.066
	HD↓	39.78±10.71	49.97±05.53	45.51±06.20	41.10±04.75	44.08±06.79
Proposed RGS	DS↑	0.755±0.126	0.708±0.010	0.741±0.084	0.814±0.040	0.754±0.065
	HD↓	39.31±11.64	50.67±06.23	46.88±09.26	41.19±05.71	44.51±08.21
Proposed RGS+AL	DS↑	0.758±0.119	0.711±0.099	0.741±0.082	0.814±0.038	0.756±0.084
	HD↓	39.45±11.40	50.41±06.63	45.85±08.99	40.68±04.89	44.09±07.97
Proposed RGS&M+AL	DS↑	0.767±0.065	0.727±0.089	0.744±0.079	0.815±0.042	0.763±0.068
	HD↓	39.26±11.24	50.07±06.60	44.95±09.26	42.25±04.85	43.91±07.98

similarity-based sample re-weighting method (RGS; re-weighting but no rectification of $\tilde{W}_{com,t}$), proposed gradient similarity-based sample re-weighting with AL (RGS+AL; re-weighting with $\tilde{W}^{*}_{com,t}$ and $\lambda = 1$), and proposed gradient similarity- and last layer gradient magnitude-based sample re-weighting with AL method (RGS&M+AL; re-weighting with $\tilde{W}^{*}_{com,t}$ and $\lambda = 0.5$). Since we also compared our performance to those of the state-of-the-art methods on the same benchmark data set, and we had to use part of the data as

X_{tr} to generate pseudo annotation, we had to use 4-fold cross-validation so that we have the same number of data samples in the test cohort as in the state-of-the-art. Table 3 shows that the proposed RGS method, where we used CT volumes with noisy annotation during training, performs better in terms of the Dice score than the fully supervised approach. This performance confirms that the gradient similarity between the data with expert annotation and noisy annotation helps to automatically emphasize more trustworthy samples in a training batch in the loss calculation and thus update the model parameters during the back-propagation. We further see in Table 3 that the RGS+AL method performs better in terms of Dice score than the RGS method alone in all folds except Fold 4, which demonstrates that the complete removal of the contribution from less trustworthy samples (i.e., AL in terms of rectification of $\tilde{W}_{com,t}$ with $\lambda = 1$) in batch-wise loss calculation improves the model's segmentation performance. We further see in Table 3 that the RGS&M+AL method performs the best in terms of the Dice score compared to all other approaches in all folds. It proves that the re-weighting of a machine-annotated sample, based on its gradient similarity and last layer gradient magnitude reflecting its "trustworthiness" and "informativeness," respectively, improves the accuracy of the deep segmentation model. Here also, after incorporating both the data informativeness and label trustworthiness into the sample re-weighting, followed by the complete removal of the contribution of less trustworthy and informative samples (i.e., AL in terms of rectification of $\tilde{W}_{com,t}$ with $\lambda = 0.5$) in batch-wise loss calculation, we obtain a better-performing model. We also observe in Table 3 that the average Hausdorff distance by the proposed RGS&M+AL method is the

best among all other techniques. We further performed the two-sample t-test for the 5-fold mean Dice scores between our proposed RGS&M+AL and other methods mentioned in Table 3. The estimated p-values are 0.0021, 0.3399, and 0.5180 for the fully supervised, RGS, and RGS+AL methods, respectively. Since RGS and RGS+AL methods are intrinsic parts of our proposed RGS&M+AL method, differences in segmentation performance by these approaches may not be significant. However, as expected, the proposed RGS&M+AL showed statistically significant improvements ($p < 0.01$) in terms of Dice score compared to the fully supervised approach.

In Fig. 3, we present the qualitative performance comparison of the fully supervised, RGS, RGS+AL, and RGS&M+AL methods. Here, we show the axial CT slices and corresponding expert-annotated pneumonia infection masks for five COVID-19-positive patients. Similar to the segmentation performance on the challenge data, here we see in this figure that all the methods performed reasonably well in identifying pneumonia infections in the lung. However, for more irregular and complex infection patterns (i.e., patients I, III, IV, and V), the masks produced by the proposed RGS&M+AL method (shown in the last row) match the best with the expert-annotated infection masks (shown in the second row). Therefore, this qualitative performance further supports the efficacy of the proposed RGS&M+AL approach.

Next, in Table 4, we show the Dice scores achieved by different methods to segment pneumonia infection in the benchmark dataset. Here, we show results for two fully supervised learning approaches (Isensee et al. (2019) and our 3D UNet implementation), five semi-supervised learning approaches (Chen et al. (2020), Yu et al. (2019), Ma et al. (2020a), Fan et al. (2020), and

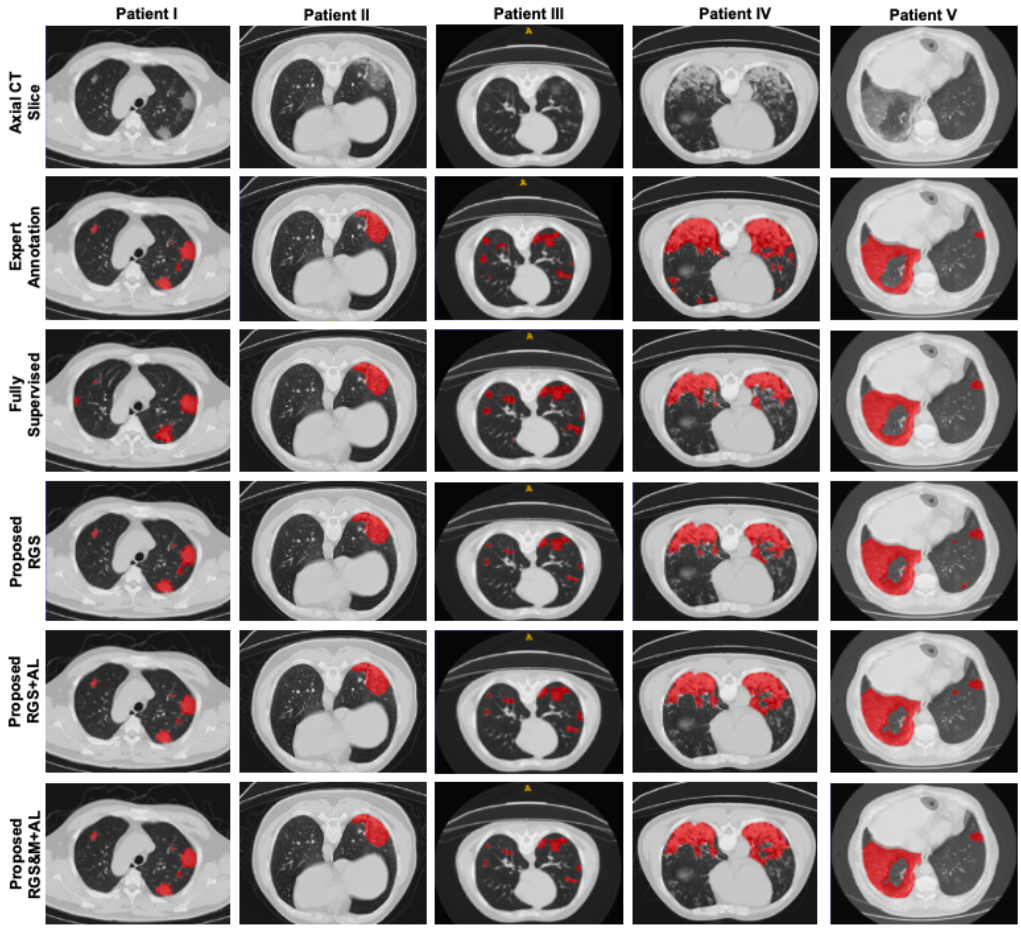

Figure 3: Qualitative performance comparison by our implemented methods in pneumonia infection segmentation on the **Benchmark data**. The first row shows the axial CT slices of five COVID-19-infected patients. The second row shows the expert-generated infection mask overlaid on the corresponding CT slices. The third to sixth rows show infection segmentation masks generated by different approaches.

proposed RGS method), and two active learning by noisy teacher approaches, namely, 'proposed RGS+AL' and 'proposed RGS&M+AL.' Note that Ma et al. (2020a) are the curator and publishers of this Benchmark dataset (Ma

Table 4 Dice scores achieved by contrasting methods in the segmentation of pneumonia infections in **Benchmark data**. (Acronyms: RGS: sample re-weighting based on gradient similarity only, RGS+AL: sample re-weighting based on gradient similarity followed by active learning, RGS&M+AL: sample re-weighting based on both gradient similarity and last layer gradient magnitude followed by active learning.)

Method Type	**Methods**	**Mean Dice**
Fully supervised	Isensee et al. (2019)	0.6728 ± 0.2220
	Our 3D UNet	0.7307 ± 0.0660
Semi-supervised	Chen et al. (2020)	0.6759 ± 0.2230
	Yu et al. (2019)	0.6962 ± 0.2030
	Ma et al. (2020a)	0.7225 ± 0.1989
	Fan et al. (2020)	0.5970
	Proposed RGS	0.7548 ± 0.06517
Active Learning from Noisy Teacher	Proposed RGS+AL	0.7562 ± 0.0848
	Proposed RGS&M+AL	$\mathbf{0.7635 \pm 0.0687}$

et al., 2020b) that we use in this paper for validation. Furthermore, the methods of Isensee et al. (2019), Chen et al. (2020), and Yu et al. (2019) were implemented and tested on the same Benchmark dataset of Ma et al. (2020a), where they adhered to the exact computation and pre-processing steps discussed in the respective articles. Therefore, in Table 4, we report the mean Dices of methods by Isensee et al. (2019), Chen et al. (2020), Yu et al. (2019), and Ma et al. (2020a), as reported by Ma et al. (2020a). In this way, we also avoided any deteriorated performance that could have resulted from our own implementation of these methods. We also plot the mean Dice by Fan et al. (2020) in Table 4, which was reported on the same Benchmark dataset although they did not report the standard deviation. Comparing

Dice scores by different methods in Table 4, we see that our base 3D UNet outperformed all other methods in infection segmentation. Further improvement in the mean Dice score is achieved by our proposed semi-supervised RGS method. Although trained with the machine-annotated data, the proposed RGS method outperformed our base 3D UNet model because of using the gradient similarity-based sample re-weighting. Finally, we see in Table 4 that the proposed RGS+AL method further improves the Dice score than the RGS method, and the proposed RGS&M+AL method performs the best among all methods. We also performed the two-sample t-test between our proposed RGS&M+AL and other state-of-the-art methods mentioned in Table 4, where the estimated p-values are 0.0002, 0.0001, 0.0003, 0.0023, 0.0567, and 3.86e-36 for the methods by Isensee et al. (2019), our 3D UNet, Chen et al. (2020), Yu et al. (2019), Ma et al. (2020a), and Fan et al. (2020), respectively. Except for the method by Ma et al. (2020a), the performance improvement by the proposed RGS&M+AL is statistically significant ($p < 0.01$) compared to other methods. This result again reinforces our claim that the re-weighting of pseudo-annotated samples, based on their gradient similarity and last layer gradient magnitude followed by the complete removal of the contribution from lesser trustworthy and informative samples (i.e., AL in terms of rectification of $\tilde{W}_{com,t}$) in batch-wise loss calculation, leads to better model optimization and segmentation performance.

Although the proposed RGS&M+AL approach showed better segmentation performance compared to other state of the art, this approach is slightly computationally expensive compared to training a 3D UNet under full supervision. The proposed approach needs an additional forward pass and gradient calculation for the meta-network in each iteration, although the parameters of the meta-network are not updated. The meta-network is an identical 3D UNet that loads the current state of parameters from the actual 3D UNet in training in each iteration. Despite the use of an additional meta-network, the total number of trainable parameters of the proposed method remains the same as shown in Table 1. Also because of an additional forward pass and gradient calculation for the meta-network in each iteration, the total training time is slightly higher for our proposed model training than for training a 3D UNet under full supervision. Despite a slightly longer training time and more computational complexity while incorporating larger training data without expert annotation, our approach showed better segmentation performance compared to other state-of-the-art approaches.

6. Conclusions

We proposed a new semi-supervised segmentation method that deploys a noisy teacher-based active deep learning strategy. We use an example reweighting scheme that adaptively weights pseudo-annotated training samples based on the similarity of their gradient directions to those of the expert-annotated validation data and the gradient magnitude of the last layer of the deep model. We incorporated the trustworthiness and informativeness of pseudo-annotated data samples within an active learning strategy by in-

corporating a query function in the re-weighting process that favors more trustworthy and more informative samples from batch data. We validated our approach using 3 different clinical CT databases of COVID-19 and non-COVID pneumonia lung images and demonstrated that our method outperformed state of the art in COVID-19 pneumonia infection segmentation. Our proposed method achieved the highest Dice score using a smaller number of expert-annotated data in the semi-supervised model training. The conventional deep learning framework often faces various challenges in maintaining a standard accurate predictability when training and testing data come from different sources, which is referred to as the 'domain shift.' Since our proposed approach utilized the gradient similarity between the training and validation data (from two different sources), it can be more robust in the domain shift problem. Additionally, our proposed approach showed efficacy in producing accurate pneumonia infection masks, although we had extremely limited expert-annotated data. Our method can significantly contribute to image-based diagnosis procedures in the clinical environment via leveraging the commonly available large pool of annotation-free data in the hospital records, as attaining expert annotation by radiologists is a common bottleneck in the volumetric medical image-based supervised learning framework. While demonstrating the best COVID-19 pneumonia infection segmentation performance compared to the state of the arts, our method has a few limitations that require further improvement in the future. For example, we empirically choose the value of λ. Our future plan includes developing an automatic data-driven process to set the value of λ on the fly during model training. We also plan to validate our method on a larger expert-annotated

data pool once available.

References

Abdel-Basset, M., Chang, V., Hawash, H., Chakrabortty, R.K., Ryan, M., 2020. FSS-2019-nCov: A deep learning architecture for semi-supervised few-shot segmentation of COVID-19 infection. Knowledge-Based Systems 212, 106647.

Amyar, A., Modzelewski, R., Li, H., Ruan, S., 2020. Multi-task deep learning based CT imaging analysis for COVID-19 pneumonia: Classification and segmentation. Computers in Biology and Medicine 126, 104037.

An, P., Xu, S., Harmon, S., Turkbey, E., Sanford, T., Amalou, A., Kassin, M., Varble, N., Blain, M., Anderson, V., Patella, F., Carrafiello, G., Turkbey, B., Wood, B., 2020. CT images in COVID-19. https://doi.org/10.7937/tcia.2020.gqry-nc81 .

Ash, J.T., Zhang, C., Krishnamurthy, A., Langford, J., Agarwal, A., 2019. Deep batch active learning by diverse, uncertain gradient lower bounds, in: International Conference on Learning Representations.

Bull, L., Worden, K., Manson, G., Dervilis, N., 2018. Active learning for semi-supervised structural health monitoring. Journal of Sound and Vibration 437, 373–388.

Calma, A., Reitmaier, T., Sick, B., 2018. Semi-supervised active learning for support vector machines: A novel approach that exploits structure information in data. Information Sciences 456, 13–33.

Chen, Z., Zhang, R., Zhang, G., Ma, Z., Lei, T., 2020. Digging into pseudo label: a low-budget approach for semi-supervised semantic segmentation. IEEE Access 8, 41830–41837.

Cheplygina, V., de Bruijne, M., Pluim, J.P., 2019. Not-so-supervised: a survey of semi-supervised, multi-instance, and transfer learning in medical image analysis. Medical image analysis 54, 280–296.

Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O., 2016. 3D U-Net: learning dense volumetric segmentation from sparse annotation, in: International conference on medical image computing and computer-assisted intervention, Springer. pp. 424–432.

Elharrouss, O., Subramanian, N., Al-Maadeed, S., 2020. An encoder-decoder-based method for COVID-19 lung infection segmentation. arXiv preprint arXiv:2007.00861 .

Fan, D.P., Zhou, T., Ji, G.P., Zhou, Y., Chen, G., Fu, H., Shen, J., Shao, L., 2020. Inf-Net: Automatic COVID-19 lung infection segmentation from CT images. IEEE Transactions on Medical Imaging 39, 2626–2637.

Gao, K., Su, J., Jiang, Z., Zeng, L.L., Feng, Z., Shen, H., Rong, P., Xu, X., Qin, J., Yang, Y., et al., 2020a. Dual-branch combination network (DCN): Towards accurate diagnosis and lesion segmentation of COVID-19 using CT images. Medical image analysis 67, 101836.

Gao, M., Zhang, Z., Yu, G., Arık, S.Ö., Davis, L.S., Pfister, T., 2020b. Consistency-based semi-supervised active learning: Towards minimizing

labeling cost, in: European Conference on Computer Vision, Springer. pp. 510–526.

Ghomi, Z., Mirshahi, R., Bagheri, A.K., Fattahpour, A., Mohammadiun, S., Gharahbagh, A.A., Djavadifar, A., Arabalibeik, H., Sadiq, R., Hewage, K., 2020. Segmentation of COVID-19 pneumonia lesions: A deep learning approach. Medical Journal of the Islamic Republic of Iran 34, 174.

Harmon, S.A., Sanford, T.H., Xu, S., Turkbey, E.B., Roth, H., Xu, Z., Yang, D., Myronenko, A., Anderson, V., Amalou, A., et al., 2020. Artificial intelligence for the detection of COVID-19 pneumonia on chest CT using multinational datasets. Nature communications 11, 1–7.

Hasan, M., Jawad, M., Hasan, K.N.I., Partha, S.B., Masba, M., Al, M., Saha, S., et al., 2021. COVID-19 identification from volumetric chest CT scans using a progressively resized 3D-CNN incorporating segmentation, augmentation, and class-rebalancing. arXiv preprint arXiv:2102.06169 .

Hong, S., Noh, H., Han, B., 2015. Decoupled deep neural network for semi-supervised semantic segmentation. Advances in neural information processing systems 28.

Isensee, F., Jäger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H., 2019. Automated design of deep learning methods for biomedical image segmentation. arXiv preprint arXiv:1904.08128 .

Kerfoot, E., Clough, J., Oksuz, I., Lee, J., King, A.P., Schnabel, J.A., 2018. Left-ventricle quantification using residual U-Net, in: International

Workshop on Statistical Atlases and Computational Models of the Heart, Springer. pp. 371–380.

Laradji, I., Rodriguez, P., Manas, O., Lensink, K., Law, M., Kurzman, L., Parker, W., Vazquez, D., Nowrouzezahrai, D., 2021. A weakly supervised consistency-based learning method for COVID-19 segmentation in CT images, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 2453–2462.

Lassau, N., Ammari, S., Chouzenoux, E., Gortais, H., Herent, P., Devilder, M., Soliman, S., Meyrignac, O., Talabard, M.P., Lamarque, J.P., et al., 2021. Integrating deep learning CT-scan model, biological and clinical variables to predict severity of COVID-19 patients. Nature communications 12, 1–11.

Lee, J., Kim, E., Lee, S., Lee, J., Yoon, S., 2019. FickleNet: Weakly and semi-supervised semantic image segmentation using stochastic inference, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5267–5276.

Li, Y., Luo, L., Lin, H., Chen, H., Heng, P.A., 2021. Dual-consistency semi-supervised learning with uncertainty quantification for COVID-19 lesion segmentation from CT images. arXiv preprint arXiv:2104.03225 .

Lv, Y., Liu, B., Zhang, J., Dai, Y., Li, A., Zhang, T., 2022. Semi-supervised active salient object detection. Pattern Recognition 123, 108364.

Ma, J., Nie, Z., Wang, C., Dong, G., Zhu, Q., He, J., Gui, L., Yang, X., 2020a. Active contour regularized semi-supervised learning for COVID-19

CT infection segmentation with limited annotations. Physics in Medicine & Biology 65, 225034.

Ma, J., Wang, Y., An, X., Ge, C., Yu, Z., Chen, J., Zhu, Q., Dong, G., He, J., He, Z., et al., 2020b. Towards efficient COVID-19 CT annotation: A benchmark for lung and infection segmentation. arXiv preprint arXiv:2004.12537 .

Marzahl, C., Bertram, C.A., Aubreville, M., Petrick, A., Weiler, K., Gläsel, A.C., Fragoso, M., Merz, S., Bartenschlager, F., Hoppe, J., et al., 2020. Are fast labeling methods reliable? a case study of computer-aided expert annotations on microscopy slides, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 24–32.

Mirikharaji, Z., Yan, Y., Hamarneh, G., 2019. Learning to segment skin lesions from noisy annotations, in: Domain Adaptation and Representation Transfer and Medical Image Learning with Less Labels and Imperfect Data. Springer, pp. 207–215.

Paluru, N., Dayal, A., Jenssen, H.B., Sakinis, T., Cenkeramaddi, L.R., Prakash, J., Yalavarthy, P.K., 2021. Anam-Net: Anamorphic depth embedding-based lightweight CNN for segmentation of anomalies in COVID-19 chest CT images. IEEE Transactions on Neural Networks and Learning Systems 32, 932–946.

Phua, J., Weng, L., Ling, L., Egi, M., Lim, C.M., Divatia, J.V., Shrestha, B.R., Arabi, Y.M., Ng, J., Gomersall, C.D., et al., 2020. Intensive care

management of coronavirus disease 2019 (COVID-19): challenges and recommendations. The lancet respiratory medicine 8, 506–517.

Ranjbarzadeh, R., Jafarzadeh Ghoushchi, S., Bendechache, M., Amirabadi, A., Ab Rahman, M.N., Baseri Saadi, S., Aghamohammadi, A., Kooshki Forooshani, M., 2021. Lung infection segmentation for COVID-19 pneumonia based on a cascade convolutional network from CT images. BioMed Research International 2021.

Ren, M., Zeng, W., Yang, B., Urtasun, R., 2018. Learning to reweight examples for robust deep learning, in: International Conference on Machine Learning, PMLR. pp. 4334–4343.

Singh, V.K., Abdel-Nasser, M., Pandey, N., Puig, D., 2021. LungINFseg: Segmenting COVID-19 infected regions in lung CT images based on a receptive-field-aware deep learning framework. Diagnostics 11, 158.

Souly, N., Spampinato, C., Shah, M., 2017. Semi supervised semantic segmentation using generative adversarial network, in: Proceedings of the IEEE international conference on computer vision, pp. 5688–5696.

Taghanaki, S.A., Havaei, M., Berthier, T., Dutil, F., Di Jorio, L., Hamarneh, G., Bengio, Y., 2019a. InfoMask: Masked variational latent representation to localize chest disease, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 739–747.

Taghanaki, S.A., Zheng, Y., Zhou, S.K., Georgescu, B., Sharma, P., Xu, D., Comaniciu, D., Hamarneh, G., 2019b. Combo loss: Handling input and

output imbalance in multi-organ segmentation. Computerized Medical Imaging and Graphics 75, 24–33.

Tilborghs, S., Dirks, I., Fidon, L., Willems, S., Eelbode, T., Bertels, J., Ilsen, B., Brys, A., Dubbeldam, A., Buls, N., et al., 2020. Comparative study of deep learning methods for the automatic segmentation of lung, lesion and lesion type in CT scans of COVID-19 patients. arXiv preprint arXiv:2007.15546 .

Voulodimos, A., Protopapadakis, E., Katsamenis, I., Doulamis, A., Doulamis, N., 2021a. Deep learning models for COVID-19 infected area segmentation in CT images, in: The 14th PErvasive Technologies Related to Assistive Environments Conference, pp. 404–411.

Voulodimos, A., Protopapadakis, E., Katsamenis, I., Doulamis, A., Doulamis, N., 2021b. A few-shot U-Net deep learning model for COVID-19 infected area segmentation in CT images. Sensors 21, 2215.

Wang, G., Liu, X., Li, C., Xu, Z., Ruan, J., Zhu, H., Meng, T., Li, K., Huang, N., Zhang, S., 2020. A noise-robust framework for automatic segmentation of COVID-19 pneumonia lesions from CT images. IEEE Transactions on Medical Imaging 39, 2653–2663.

Xiong, Y., Sun, D., Liu, Y., Fan, Y., Zhao, L., Li, X., Zhu, W., 2020. Clinical and high-resolution CT features of the COVID-19 infection: comparison of the initial and follow-up changes. Investigative radiology .

Xu, X., Lee, G.H., 2020. Weakly supervised semantic point cloud segmen-

tation: Towards 10x fewer labels, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13706–13715.

Xu, Z., Lu, D., Wang, Y., Luo, J., Jayender, J., Ma, K., Zheng, Y., Li, X., 2021. Noisy labels are treasure: mean-teacher-assisted confident learning for hepatic vessel segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 3–13.

Yan, Q., Wang, B., Gong, D., Luo, C., Zhao, W., Shen, J., Ai, J., Shi, Q., Zhang, Y., Jin, S., et al., 2021. COVID-19 chest CT image segmentation network by multi-scale fusion and enhancement operations. IEEE Transactions on Big Data 7, 13–24.

Yang, D., Xu, Z., Li, W., Myronenko, A., Roth, H.R., Harmon, S., Xu, S., Turkbey, B., Turkbey, E., Wang, X., et al., 2021. Federated semi-supervised learning for COVID region segmentation in chest CT using multi-national data from China, Italy, Japan. Medical image analysis 70, 101992.

Yu, L., Wang, S., Li, X., Fu, C.W., Heng, P.A., 2019. Uncertainty-aware self-ensembling model for semi-supervised 3D left atrium segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 605–613.

Zhang, H.t., Zhang, J.s., Zhang, H.h., Nan, Y.d., Zhao, Y., Fu, E.q., Xie, Y.h., Liu, W., Li, W.p., Zhang, H.j., et al., 2020. Automated detection and quantification of COVID-19 pneumonia: CT imaging analysis by a

deep learning-based software. European journal of nuclear medicine and molecular imaging 47, 2525–2532.

Zhang, Z.Y., Zhao, P., Jiang, Y., Zhou, Z.H., 2019. Learning from incomplete and inaccurate supervision, in: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 1017–1025.

Zhao, Z., Zeng, Z., Xu, K., Chen, C., Guan, C., 2021. DSAL: Deeply supervised active learning from strong and weak labelers for biomedical image segmentation. IEEE Journal of Biomedical and Health Informatics .

Zhou, L., Li, Z., Zhou, J., Li, H., Chen, Y., Huang, Y., Xie, D., Zhao, L., Fan, M., Hashmi, S., et al., 2020. A rapid, accurate and machine-agnostic segmentation and quantification method for CT-based COVID-19 diagnosis. IEEE transactions on medical imaging 39, 2638–2652.

Zhou, Z.H., 2018. A brief introduction to weakly supervised learning. National science review 5, 44–53.