Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Information Sciences 578 (2021) 559–573

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

RCTE: A reliable and consistent temporal-ensembling framework for semi-supervised segmentation of COVID-19 lesions

Weiping Ding [a,*], Mohamed Abdel-Basset [b], Hossam Hawash [b]

[a] School of Information Science and Technology, Nantong University, Nantong 226019, China
[b] Zagazig Univesitry, Shaibet an Nakareyah, Zagazig 2, 44519 Ash Sharqia Governorate, Egypt

ARTICLE INFO

Article history:
Received 16 March 2021
Received in revised form 17 June 2021
Accepted 17 July 2021
Available online 21 July 2021

Keywords:
COVID-19
Deep learning
Semi-supervised learning
CT scans
Temporal-ensembling

ABSTRACT

The segmentation of COVID-19 lesions from computed tomography (CT) scans is crucial to develop an efficient automated diagnosis system. Deep learning (DL) has shown success in different segmentation tasks. However, an efficient DL approach requires a large amount of accurately annotated data, which is difficult to aggregate owing to the urgent situation of COVID-19. Inaccurate annotation can easily occur without experts, and segmentation performance is substantially worsened by noisy annotations. Therefore, this study presents a reliable and consistent temporal-ensembling (RCTE) framework for semi-supervised lesion segmentation. A segmentation network is integrated into a teacher-student architecture to segment infection regions from a limited number of annotated CT scans and a large number of unannotated CT scans. The network generates reliable and unreliable targets, and to evenly handle these targets potentially degrades performance. To address this, a reliable teacher-student architecture is introduced, where a reliable teacher network is the exponential moving average (EMA) of a reliable student network that is reliably renovated by restraining the student involvement to EMA when its loss is larger. We also present a noise-aware loss based on improvements to generalized cross-entropy loss to lead the segmentation performance toward noisy annotations. Comprehensive analysis validates the robustness of RCTE over recent cutting-edge semi-supervised segmentation techniques, with a 65.87% Dice score.

1. Introduction

By the end of December 2019, the world was gripped by a new coronavirus epidemic, scientifically known as SARS-CoV-2, which causes acute viral pneumonia disease (COVID-19) [1]. The epidemic proliferated rapidly owing to human-to-human communication, with a reported 96,801,177 patients positively confirmed as COVID-19 and 2,069,763 mortalities as of January 01, 2021. The epidemic has become a worldwide threat to health and economic infrastructures. Hence the quick identification of COVID-19 threat considerations is essential for the remedy of infected patients and global infection containment.

In clinical practice, the reverse-transcription polymerase chain reaction (RT-PCR) is an important technique for screening COVID-19. However, to obtain test results takes up to two days due to insufficient resources and rigorous situational needs

* Corresponding author.
E-mail address: dwp9988@163.com (W. Ding).

https://doi.org/10.1016/j.ins.2021.07.059
0020-0255/

that limit the reliability and speed of screening patients. Moreover, the RT-PCR test exhibits a high ratio of false-negative samples. Clinicians and researchers have adopted the computed tomography (CT) scan as an efficient diagnostic tool, and demonstrated its efficiency at COVID-19 detection in terms of sensitivity and accuracy. Thus to leverage clinical findings and lung CT manifestations is the most appropriate way to realize a rapid and active diagnosis in terms of follow-up assessment and disease progression monitoring [2].

Advances in artificial intelligence (AI) technologies have facilitated disease forecasting [3–5] and AI-enabled computer-aided diagnosis (CAD) applications in the healthcare field [6]. Among these AI techniques, deep learning (DL) has proved efficient at different tasks in automated medical image analysis, especially for lung disease diagnosis [7,8]. However, current supervised DL approaches are known to be data-hungry, as they require a vast number of annotated scans to realize accurate performance. Trustworthy pixel-level annotation of chest CT scans is typically time-consuming, and it requires the laborious effort of an experienced radiological specialist. Owing to the rapid spread of COVID-19, annotating such enormous numbers of CT scans is impractical due to tight timelines and the intense workloads of the healthcare community [3].

To tackle these limitations, annotation-efficient DL for medical image segmentation has attracted growing research to relax the requirement of a large-scale, pixel-level annotated CT dataset for training, inspiring training techniques with partial or no supervision, such as unsupervised domain adaptation (UDA) [9], weakly supervised learning (WSL) [10], semi-supervised learning (SSL) [1,11,12], self-supervised learning (S-SL) [13], and active learning [14]. It is imperative that these techniques prevent overfitting of a network with scant annotated data. The present study investigates the semi-supervised segmentation (SS-seg) of COVID-19 lesions where a large amount of unannotated labeled CT data can be aggregated effortlessly.

SSL has been widely used to improve learning performance in scarce-annotation situations, which is a vital and challenging task with a marked effect on medical applications. Solutions of the SS-seg problem incrementally include segmentation maps from unannotated samples in the training data to improve segmentation performance. Other SSL approaches employ generative adversarial networks (GANs) [15], variational autoencoders (VAEs) [16], and ensemble learning for segmentation or classification purposes [17]. Recently, several SSL approaches have implemented a consistently imposing paradigm [18] that benefits from the unannotated data by regularizing model predictions, which will be consistent in case of applying various disturbances to the inputs and parameters of the model. Consistency-based approaches have emphasized the improvement of superiority of consistency targets. For instance, the method of temporal ensembling (TE) [19,20] computes consistency targets as exponential moving averages (EMAs) of estimates in several epochs. However, this requires a large prediction matrix throughout model training. The mean teacher (MT) [21] approach solves this issue through an ensembled teacher architecture to provide training consistency entities that have shown robust performance in a number of studies [10–14]. However, this approach ignores information about relations between samples, which the self-ensembling (SE) framework [22] addresses through a sample relation consistency matrix.

Despite recent improvements in automated DL-based segmentation from CT scans, most current studies employ standard convolutional networks (i.e., U-Nets) that are trained normally by ignoring noisy annotations. Several studies have investigated learning from noisy labels in medical image classification problems [23,24]. In view of this, to develop a noise-aware cost function has great potential in many medical applications and has motivated much research, as it lessens the need for image refinement techniques and complex models, and can be easily integrated with any learning strategy [25]. However, to apply these functions to segmentation may lead to poor efficiency because of the typical imbalance between the numbers of background and foreground pixels [18]. Therefore, this study investigates a noise-aware semi-supervised TE framework to efficiently reduce the effect of noisy annotations during COVID-19 lesion segmentation.

1.1. Challenges and motivation

While several AI models have been developed to facilitate the automation of COVID-19 diagnosis, there has been little study of COVID-19 lesion segmentation. To detect regions of interest (ROIs) from CT scans is an interesting and challenging task for several reasons.

- Large divergence in the characteristics of lesions in terms of scope, location, shape, and quality makes them difficult to classify (see Fig. 1). For instance, consolidations are small and precise, and often cause erroneous detection.
- Small inter-class divergence means that the margins of ground-glass opacity (GGO) predominantly exhibit clouded manifestation and low contrast, which complicates the detection process. This makes it impossible to aggregate a satisfactory amount of annotated data; hence domain experts must devote much time and effort to generate reliable annotations.
- Noisy annotation is inevitable for rare or new diseases (e.g., COVID-19), which decreases segmentation efficiency. However, the current DL literature focuses on noisy labels in the context of the classification problem.
- SS-seg techniques often generate both reliable and unreliable pseudo-annotations and treat them equally when computing the unsupervised loss, which likely degrades performance due to unreliable targets.

1.2. Novelty and contributions

To address the above challenges, this study presents a DL approach to use both annotated and unannotated CT images to segment COVID-19 infection lesions.

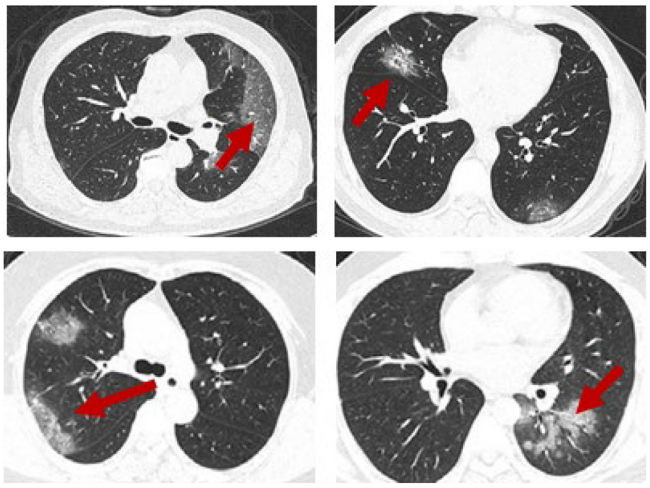

Fig. 1. Complicated manifestations of COVID-19 lesions from axial slice of CT scans of different patients.

The primary contributions of our study are summarized below.

1) We introduce a reliable, consistent temporal-ensembling (RCTE) framework that leverages unannotated CT data for efficient semi-supervised segmentation of COVID-19 lesions from limited numbers of annotated CT scans. RCTE can be generalized for SS-seg applications from 2D medical images.
2) A noise-aware loss function is introduced to mitigate the impact of noisy annotations on segmentation performance. The new loss is an improvement of GCE loss [26], which is vigorous for noisy annotations and less responsive for the imbalance between background and foreground.
3) Comprehensive experiments on public datasets of COVID-19 CT images demonstrate the ability of RCTE to realize more efficient segmentation over recent cutting-edge SS-seg approaches while avoiding the effect of noisy annotations.

1.3. Study structure

The remainder of this paper is structured as follows. Section 2 discusses associated work involving semi-supervised DL approaches and noise-aware DL models. In Section 3, we present a detailed description of our methodology of COVID-19 lesion segmentation. Results, comparisons, analysis, and discussion are presented in Section 4. Section 5 summarizes our conclusions and identifies future research directions.

2. Related Studies

2.1. COVID-19 lesion segmentation

Despite the criticality of lesion segmentation for numerical evaluations in the task of COVID-19 diagnosis, few studies have investigated efficient and automated segmentation of COVID-19 lesions from CT scans. Fan et al. [11] developed a segmentation network called Inf-Net for automated segmentation of infected areas in lung CT slices. It learns a high-level representation using a concurrent partial decoder, and employs reverse and edge attention to enhance boundary detection capability. Wang et al. [1] introduced a DL approach called COPLE-Net to segment infection areas of different sizes and manifestations. Zhou et al. [8] reduced the complexity of segmentation by decomposing the 3D segmentation task into three 2D segmentation tasks, whose results are averaged. Adel et al. [2] presented an end-to-end segmentation method based on CT scan enhancement. Gao et al. [27] introduced a dual-stream approach with an in-between lesion attention module for the classification and segmentation of COVID-19 lesions from input CT slices. Mahmud et al. [28] segmented COVID-19 lesions using an attention-enabled segmentation model (TA-SegNet), repetitively employing a tri-level attention module at different network positions to aggregate pixel-, channel-, and spatial-wise representations during training. All these supervised approaches were validated on private or small-scale data, making them unreliable for the real world. Their performance must be investigated on sufficient amounts of labeled data. Despite the ease of obtaining a large amount of unlabeled CT data, to aggregate a large set of pixel-level annotated CT scans was impossible at the time of the outbreak, which motivates us to tackle the limited annotation problem through this study.

2.2. Semi-supervised self-ensembling

Owing to difficulties in aggregating large-scale annotated radiological data, SSL techniques have experienced wide adoption, as they enable the improvement of the performance of DL models even with a limited number of annotated images and a large number of unannotated images [17,29,30]. Semi-supervised SE approaches have been commonly employed to train DL networks such as to reduce the supervision and regularization costs on annotated and unannotated images, respectively [1,20,31]. Other studies have proposed regularization improvement through examining learned expertise, including TE reliance and pseudo-annotations [32]. However, these approaches neglect the relations among images, which provide valuable semantic information from unannotated data. Liu et al. [22] addressed this limitation and introduced an SSL approach for medical image classification that exploits an SE model to generate a target of superior consistency for unannotated images using relation information fused by a sample relation consistency mechanism. Similarly, Shi et al. [20] introduced an SE framework to leverage the semantic information of annotated and unannotated images to classify histopathology images. It maps the ensemble targets of each class to the same cluster for further enhancement. The above studies concentrate on the classification task, and to model the relationships between samples requires a large relation matrix or clustering map, which can retard the training process.

Fu et al. [31] developed an SS-Seg technique for medical images using a teacher-student model that seeks to minimize the weighted mixture of supervised loss using annotated inputs and an unsupervised loss that uses both unannotated and annotated images. Their model encourages consistent predictions for the same input under diverse perturbations (e.g., rotation, scale, dropout, and Gaussian noise) using a transformation-consistent paradigm to improve the regularization outcome for pixel-level inferences. Similarly, Wang et al. [1] designed a segmentation architecture, COPLE-NET, and integrated it into the SE framework wherever the input was simply disrupted with Gaussian noise. A major shortcoming of the above techniques is to disregard the dependability of pseudo-annotations. The generated pseudo-targets of the unannotated samples might be noisy and erroneous, which can have a negative effect on training the segmentation. These methods apply simple random transformations to ensure consistency without considering the choice of the best transformations for a specific dataset.

2.3. Noise-aware segmentation

Most DL studies address the problem of noisy labels in the context of medical image classification [18]. Shi et al. [20] adopted the graph TE-based classification approach to validate empirical superiority for a small ratio of classification with noisy labeling. Karim et al. [18] investigated techniques to solve noisy-label problems and showed that the mean absolute error (MAE) and generalized cross-entropy (GCE) were effective for noisy annotation. They also investigated data reweighting methods to eliminate erroneously labeled samples. Other studies have employed iterative training, ensemble techniques, and consistency regularization to learn from noisy labels. Only some of these techniques have been investigated for segmentation.

For medical image segmentation, Tajbakhsh et al. [25] reviewed and categorized the main techniques to solve the noisy-annotation problem using noise-resilient loss functions. Wang et al. [33] introduced a label denoising technique to iteratively train a DL network for male pelvic organ segmentation. Wang et al. [1] introduced a noise-robust Dice loss function that can improve lesion segmentation from weakly annotated CT scans. These techniques must set aside part of the training data for clean annotations, and the experiments employ fake noisy annotation, making real-world noisy annotations a critical challenge in medical image segmentation. Lui et al. [34] addressed the noisy-annotation problem with a three-stage image quality assessment technique that employs a hierarchical residual model to provide slice-, volume-, and subject-level assessments for diffusion magnetic resonance images. Lin et al. [35] introduced a synergistic grouping loss to increase the tolerance of a DL model to noisy annotations, including fuzzy or analogous lesions.

3. Proposed framework

We discuss the proposed RCTE for COVID-19 infection lesion segmentation. We arbitrarily select data x_i from the training data of labeled and unlabeled CT slices, and task-driven transformations are applied to these slices. Our model uses a student architecture that learns to minimize a certain loss value, and a teacher architecture is one of successive student architectures. The training of the student architecture uses the augmented images as an input. Its output is computed with the *softmax* function, compared with the ground truth (GT) mask using noise-aware loss, as shown in Fig. 2, and also compared with the output of the teacher utilizing the consistency regularization loss. Once the RCTE starts training, the parameters of the teacher architecture are upgraded by utilizing the exponential moving average (EMA) after the gradient descent updates the parameters of the student architecture. Thus the GT characteristics are broadcast to the unannotated CT slices by ensuring the consistency of model outcomes with the unannotated CT slices.

3.1. Problem Formulation

To facilitate the model description, we formulate our task of COVID-19 lesion segmentation as an SS-Seg problem. In this context, the model training set contains a number N of CT slices, which comprise M annotated CT images and $N - M$ unan-

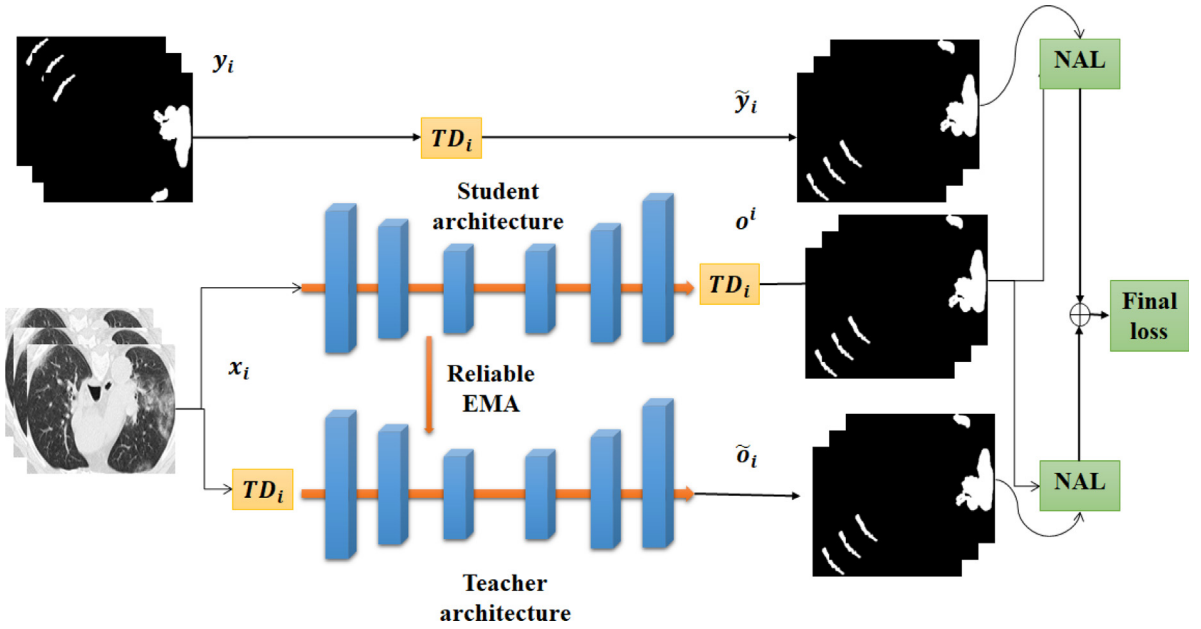

Fig. 2. RCTE architecture for SS-Seg of COVID-19 infection from 2D CT images. The teacher and student architectures have identical backbones, and the parameters of the teacher architecture are the EMA of the student architecture. The student optimizes the final loss that combines the DIL loss, CEL, and MSE loss (using labeled and unlabeled images). TD_I represents the task-driven transformation applied to the images.

notated CT images. The annotated set is $A = \{(x_i, y_i)\}_{i=1}^{M}$, and the unannotated set is $U = \{x_i\}_{i=M+1}^{N}$, where the input CT slices $x_i \in \mathbb{R}^{W\times H\times 3}$ and $y_i \in \{0,1\}\mathbb{R}^{W\times H}$ are a GT mask. Hence the overall SS-Seg tasks can be trained to optimize the parameter θ by minimizing the loss,

$$min(\theta) = \sum_{i=1}^{M} L(f(x_i,\theta), y_i) + \lambda R(\theta, A, U) \tag{1}$$

where L is the function that calculates the supervised loss, R is the function that computes the unsupervised loss, $f(\cdot)$ is the segmentation network with weight θ, and λ is the weighting agent that determines the regularization strength.

In this SS-Seg, the softness hypothesis means that neighboring data points in the CT image are expected to be adjacent in the corresponding GT mask [32]. These techniques employ SE and leverage various perturbations, aiming to refine the quality of the target through input augmentation (i.e., dropout and noise). Thus the unsupervised loss improves the prediction consistency and the model's predictive performance. The unsupervised loss R is often formulated as mean squared loss, i.e.,

$$R(\theta, L, U) = \sum_{i=1}^{N} \mathbb{E}_{\pi,\varphi} \| f(x_i;\theta,\pi) - f(x_i;\theta,\varphi) \|^2 \tag{2}$$

where π and φ represent transformations applied to the input CT image. This study adopts an equivalent concept by applying a variety of augmentations to the input CT images. That is, consistency loss based on the regularization term is employed to improve segmentation outputs when applying a variety of transformations (e.g., geometric and intensity transformation, dropout layers) to the same data.

Many experiments with the classification/segmentation network two times aim to obtain two projections under distinct transformations. In this way, the segmentation network has roles of both teacher and student. In the student role, it learns normally, as previously mentioned; as a teacher, it creates the targets to be utilized by itself in student learning. The fact that the network creates the targets by itself can be problematic and erroneous, particularly when extreme weights are assigned to newly created targets. To address this issue and to enable the creation of more reliable targets, we employ the architecture design of the mean teacher (MT) [21], in which the teacher architecture $f_{\theta'}$ uses the EMA parameters of the student architecture f_{θ}, as

$$\theta'_t = \alpha.\theta'_{t-1} + (1-\alpha).\theta_t \tag{3}$$

where θ_t *and* θ'_t, respectively, represent the parameters of the student and teacher architecture, and α is a smoothing coefficient that determines to what extent the teacher architecture depends on the parameter of the existing student architecture. The larger the value of α the more reliance there is on the preceding teacher architecture. This depends on experimental evidence [21] indicating that the best model performance was achieved when $\alpha = 0.999$. Hence we set $\alpha = 0.999$ in our experiments.

3.2. Noise-aware Loss (NAL)

Early-phase COVID-19 lesions frequently occupy a small CT scan area, which might bias the network prediction to the background, especially if the training utilizes traditional classification models and typical objective functions. The Dice loss,

$$DIL = 1 - \frac{2 \times \sum_{x_i} o_i . y_i}{\sum_{x_i} o_i + \sum_{x_i} y_i} \tag{4}$$

where o_i and y_i represent the model prediction and GT, respectively, for input x_i, has been shown to overcome this limitation by indirectly establishing a balance between background and foreground regions. DIL can be considered a different form of MSE, whose numerator allocates larger scores to pixels with more estimation errors. MSE has been demonstrated to be ineffective for noisy annotations, and mean absolute error (MAE) can have higher effectiveness than CEL and MSE under assumptions that treat all data points more equally [18]. However, MAE leads to poor performance of deep CNNs because of the equality assumption [21]. It also deals inefficiently with the foreground-background disparity in segmentation [36]. Additionally, the stochastic characteristic of DL optimizers causes the MAE to down-weight hard samples with accurate annotations, which significantly increases training times and decreases test accuracy [36]. Charoenphakdee et al. [37] tried to address this with symmetric learning loss (SLL) that integrates CEL and reverse CEL (RCEL),

$$CEL = -\sum_{k=1}^{K} y_i log o_i \tag{5}$$

$$RCEL = -\sum_{k=1}^{K} o_i log y_i \tag{6}$$

$$SLL = CEL + RCEL \tag{7}$$

where y_i and o_i are the GT and model predictions, respectively. Motivated by this, we introduce noise-aware loss that takes advantage of DIL and generalized SLL to form a composite loss function,

$$NAL = SLL + DIL \tag{8}$$

NAL is employed to implement the supervised loss L and unsupervised loss R.

3.3. Reliable and consistent temporal-ensembling framework

Most SE techniques apply different input transformations for semi-supervised medical classification. However, applying this to get reliable lesion segmentation is a challenge, especially for COVID-19 diagnosis, because the transformation is invariant in the case of classification and equivariant in the case of segmentation [39]. In other words, in classification, CNN only distinguishes the existence or non-existence of an entity; hence the classification decision is unchanged regardless of the image transformations applied. In contrast, in the segmentation task, transformations applied to the input image must be applied to the corresponding GT mask. However, the convolutions typically are not transformation-equivariant, which means that applying these transformations to the input of CNN does not necessitate the transformation of convolutional maps in the same way [41]. Thus the CNNs are not equivariant. Formally, each transformation $\pi \in \prod$ of input x links with the output transformation $\varphi \epsilon \Psi$, where $\varphi[f(x)] = f[\pi(x)]$ and $\pi \neq \varphi$. This restricts the unsupervised impact of input augmentation on segmentation performance [21]. Several studies have tried to improve the regularization for efficient utilization of unannotated images using consistency-based SE models. However, they simply employ standard random augmentation techniques (scaling, rotation, and Gaussian noise), which does not guarantee the best segmentation performance at all times [31].

To tackle this problem, we propose to augment the input data by automated task-driven augmentation (*TD*) [38], which enables the selection of the best transformations according to the data characteristics [39,40]. In particular, two conditional generator networks are designed and trained to generate intensity and *deformation* (i.e., non-affine) transformations for the input image. The task-driven deformation generatorTD^D receives CT images as an input, plus an arbitrarily sampled vector from a uniform distribution to generate a condensed pixel-wise deformation map V. Then the input slice and its GT mask are warped by applying bilinear interpolation centered on the map V to produce the augmented slice-mask pair. In the same way, the task-driven intensity generator $TD^I$is trained to perform additive intensity augmentation. It takes a CT image as an input combined with an arbitrarily sampled vector from a uniform distribution to yield a preservative intensity mask that is subsequently employed on the input image to generate the transformed image.

Under the RCTE framework, every CT image x_i is passed into the model for two-fold assessment to attain dualistic outcomes o_i and o_i. The RCTE framework comprises three TD_i operations, as shown in Fig. 2. In the earliest assessment, the function TD_i is performed on the input, and applied on the output at the second assessment. Through the two evaluations, arbitrary perturbations are applied in the model. The model is trained to be transformation-consistent by diminishing the gap between $z_i$and z_i (by unsupervised loss) to regularize the model to be more consistent, and thus to improve the gener-

alization performance. Of note, the regularization loss is estimated on both annotated and unannotated CT images. To use the annotated sample $x_i \in A$, the same augmentation TD_i is applied to the mask y_i and trained with the NAL. Finally, the RCTE framework is trained to the final consistency loss,

$$Final\ \ Loss = L(TD_i(f_\theta(x_i)), y_i) + \lambda(E)R\big(TD_i(f_\theta(x_i)), f_{\theta'}(TD_i(x_i))\big) \tag{9}$$

where L and R represent the supervised and unsupervised loss segmentation terms, respectively, which are implemented using the NAL computed with equation (8). $\lambda(E)$ represents a weighting factor for both losses, considered a Gaussian ramp-up curve and calculated as

$$\lambda(E) = k * e^{\left(-5(1-E)^2\right)} \tag{10}$$

where E is the training epoch, and k scales the supreme score of the weighting function. We adopt $k = 1.0$ [31]. When the model starts training, the value of $\lambda(E)$ is small, and the supervised loss L takes control of the training using annotated images. The model effectively learns to fuse important and precise information from annotated images. The model's reliability progressively increases during training, and it becomes capable of generating output for the unannotated CT images.

In earlier SE approaches [20] for SSL, the values of α and λ were manually assigned a static number or were altered progressively based on the training iteration t, regardless of factors such as the performance of the f_θ and $f_{\theta'}$ networks, diversity in data distributions, and presence of noisy annotations. During training, the network could exhibit degraded performance at an iteration, owing to noisy annotations, and upgrading the $f_{\theta'}$ network with a previously determined α might cause it to be distorted through the noisy annotations. Moreover, there is no assurance that the f_θ network permanently achieves superior results than the other network. Additionally, to apply L when the $f_{\theta'}$ network achieves inferior results to the f_θ network likely reduces the latter's efficiency. A reliable TE architecture is presented to address these issues, in which the f_θ and $f_{\theta'}$ networks are continuously upgraded depending on their performance. Specifically, we introduce a reliable $f_{\theta'}$ network by subduing the impacts of the f_θ network to it (i.e., EMA) when the S network exhibits bad performance (i.e., large training loss), potentially caused by noisy annotations [5]. This is realized by estimating the value of α according to the training loss of the $f_\theta$network,

$$\alpha = \begin{cases} \alpha', & L(TD_i(f_\theta(x_i)), y_i) < \gamma \\ 1.0, & otherwise \end{cases} \tag{11}$$

where α' is the distinctive EMA parameter when the f_θ network achieves moderately reliable performance (i.e., small training loss), and γ is an active threshold value corresponding to the S network loss, which is assigned the percentile value (p^{th}) of the loss through the latter k iterations of training. Once the noisy annotations result in an f_θ loss that surpasses γ at a particular iteration, the $f_{\theta'}$ network is not upgraded by the f_θ network. Thus the impact of noisy annotations on the $f_{\theta'}$ network is repressed. As big loss values can be caused by noisy annotations or accurate annotations of complex instances, to make the f_θ / $f_{\theta'}$ network neglect samples with high loss makes them likely to discard complex images. This may cause the model to learn just from simple samples, which would be undesirable. To solve this issue, a conventional strategy is adopted to enable the consideration of complex samples during the training of the f_θ network [5].

Therefore, we develop a reliable f_θ network by subduing the deep supervision of the $f_{\theta'}$ network on the f_θ network if the $f_{\theta'}$ network exhibits poorer performance than the f_θ network. This is realized by estimating the value of λ according to the f_θ / $f_{\theta'}$ network performance,

$$\lambda = \begin{cases} \lambda', & if\ \ L > R \\ \lambda'', & otherwise \end{cases} \tag{12}$$

where λ' is set to 0.01 [31] when the $f_{\theta'}$ network outperforms the f_θ network.λ'' = $0.1\lambda'$ is a minimum value that defeats the parameters of the regularization term R when the $f_{\theta'}$ network does not perform better than the f_θ network.

Conventional SE techniques generate pseudo-annotations through experimentation on the training set without upgrading the model parameter θ [18,22]. Solo-model experimentation might be unreliable or noisy. The TE-based classification approach mitigates this issue by accumulating the forecasts of several former model experiments into the outcomes of the ensemble. Thus the reliability of the generated labels is not influenced by a specific prediction [22]. We improve the design of RCTE to consider generated ensemble targets from the earlier training history based on a momentum factor.

Algorithm 1 shows the steps of the proposed RCTE approach. N^T, W^T, H^T represent the count of training instances, input width, and input height, respectively. Throughout the training process, CT slices of every mini-batch are initially transformed by $TD(x)$ augmentation. The augmented inputs are passed to the segmentation model, as previously described. The objective/loss function is computed, and the model parameters are upgraded using the Adam optimizer [41]. Following every training phase, the generated ensemble targets o^i are aggregated into ensemble targets O^i for upgrading as

$$O^i = \beta O_i + (1-\beta)o_i \tag{13}$$

where β is a momentum factor that regulates the amount of history to be considered by the ensemble during training. O_i is a matrix of dimension $N^T \times W^T \times H^T$ that is comprised of the generated ensemble targets of all images of the training data. o_i is the mini-batch in O_i with dimensions $B^T \times W^T \times H^T$, where B^T is the count of CT slices per mini-batch. Based on equation (13), O is upgraded from the initial value of zero, and progressively includes a weighted mean of all preceding generated targets throughout the training, which has minor weights in early epochs and a higher weight during later epochs. That O is upgraded from the initial value can be problematic because of the bias toward the start-up. A bias adjustment scheme [34] is adopted to rectify the bias that produces the pseudo-targets o,

$$o \leftarrow \frac{O}{1-\beta^k} \tag{14}$$

where k is the training epoch. Each mini-batch of targets o^i can be marked from o to compute the unsupervised loss.

Algorithm 1: Pseudocode of RCTE framework

Input: $x_i \in (A+U), y_i \in A$

1: $f_\theta(\cdot) \leftarrow$Student segmentation network
2: $f_{\theta'}(\cdot) \leftarrow$Teacher segmentation network
3: $TD(\cdot) \leftarrow$Task-driven augmentation generator
4: $\alpha \leftarrow$Smoothing factor.
5: $\lambda(E) \leftarrow$ Regularization weight ramp-up function
6: $\beta \leftarrow$Momentum factor
7: **for** $E = 1$ to the number of epochs E **do**:
8: **for** every training batch B **do**:
9: effectuate a suitable update $TD(x)$
10: $o_{i\in B} \leftarrow TD(f_\theta\ (x^{i\in B}))$
11: $o_{i\in B} \leftarrow f_{\theta'}(TD(x^{i\in B}))$
12: $loss = NAL + \lambda(T)NAL$
13: $\alpha = \begin{cases} \alpha', & L(TD_i(f_\theta(x_{i\in B})), y_{i\in B}) < \gamma \\ 1.0, & otherwise \end{cases}$
14: Upgrade θ *via Adam* optimizer
15: Upgrade $\theta'_t = \alpha.\theta'_{t-1} + (1-\alpha).\theta_t$
16: **Terminate for**
17: $O = \beta O + (1-\beta)\, o$
18: $\tilde{o} \leftarrow O/\left(1-\beta^k\right)$
19: **Terminate for**
20: **Return**θ

4. Experiments and analysis

4.1. Dataset description

MosMedData [42], a public dataset of 1100 COVID-19 chest CT scans with seriousness labels and COVID-19-associated outcomes from Moscow, Russia, was used in experiments to evaluate the segmentation performance of RCTE. Every CT volume was obtained from distinct subjects, with 30–46 slices per volume. Among them, 50 CT volumes were manually annotated by an experienced radiologist for COVID-19 infection lesions. Samples of the training set were annotated according to the human-in-the-loop scheme, where a preliminary architecture was trained on the manually annotated subset, and later employed to generate elementary annotations for training samples that were subsequently purified via inexperienced researchers as the training ground truth [43]. Hence the inter- and intra-observer variations, vague lesion boundaries, and prospective tendency toward the early model are plentiful reasons to consider these as noisy annotations. The GT masks of the images from the validation and testing sets were constructed using the manually annotated 50 CT scans by an expert radiologist, where the validation set contained 20 CT scans and the test set contained 30 CT scans (cross-validated). To estimate the degree of noise in the training masks, an arbitrary sample of 100 images was manually annotated by professionals to obtain their actual GT masks; the similarity between the actual and noisy masks was measured using the Dice score and found to have an average value of 0.87 ± 0.15.

4.2. Implementation details

All implementations were performed using the Python PyTorch library, Windows 10, and an Nvidia Quadro GPU. The details of experimental implementations are as follows. We employed the Dense U-Net model [49] as the backbone of the teacher and student architecture models. The architecture of the adopted Dense U-Net is described in Table 1. In all experiments, the model was trained for 6000 steps with 0.0001 as the initial learning rate. We eliminated the transformation procedures and conducted one single test with the original images to obtain equivalent comparisons by the phase of model testing. Once the model's prediction map was received, a thresholding operation with a constant value 0.5 was applied to produce the binary segmentation outcome while applying the morphology process to acquire the final segmentation outcome. RCTE had a total training time of 23.7 h, and an average inference time of 16.24 ± 5.79 s per single scan. It worth mentioning that the full radiological CT diagnosis and RT-PCR test consumed around 21.5 min and 4 h, respectively, whether the underlying patient was infected or not.

4.3. Evaluation metrics

Given the true-positive (TP), false-negative (FN), false-positive (FP), and true-negative (TN) samples, the evaluation indicators employed in this paper can be defined as follows.

i. Sensitivity

$$Sensitivity(SE) = \frac{TP}{TP + FN} \tag{15}$$

ii. Specificity

$$Specificity(SP) = \frac{TN}{FP + TN} \tag{16}$$

iii. Dice similarity coefficient (DSC): To estimate the commonality between the segmentation results, denoted by the set S,and the ground truth signified using the set GT, the DSC was calculated as

$$DSC = \frac{2.TP}{2.TP + FN + FP} \tag{17}$$

iv. Region-based Jaccard Index

$$JI = \frac{TP}{TP + FP + TN} \tag{18}$$

Table 1
Architecture of Dense U-Net for COVID-19 lesion segmentation.

Layers	Output size	Kernel size
Input Layer	128 × 512	–
Convolution	64 × 256	conv(3 × 3), stride 2
Pooling	32 × 128	3 × 3max pool, stride 2
1st Dense Block 1	32 × 128	$\begin{bmatrix} \text{conv}(1\times1) \\ \text{conv}(3\times3) \end{bmatrix} \times 6$
1st BottleneckLayer 1	32 × 128	conv(1 × 1)
	16 × 64	average pooling (2 × 2, stride = 2)
2nd Dense Block 2	16 × 64	$\begin{bmatrix} \text{conv}(1\times1) \\ \text{conv}(3\times3 \end{bmatrix} \times 12$
2nd Bottleneck Layer 2	16 × 64	conv(1 × 1)
	8 × 32	average pooling (2 × 2, stride = 2)
3rd Dense Block	8 × 32	$\begin{bmatrix} \text{conv}(1\times1) \\ \text{conv}(3\times3) \end{bmatrix} \times 24$
3rd Bottleneck Layer	8 × 32	conv(1 × 1)
	4 × 16	2 × 2average pool, stride 2
4th Dense Block	4 × 16	$\begin{bmatrix} \text{conv}(1\times1) \\ \text{conv}(3\times3 \end{bmatrix} \times 16$
1st Upsampling Layer	8 × 32	Upsampling (2 × 2), $[conv(3\times3)\times512]$
2nd Upsampling Layer	16 × 64	Upsampling (2 × 2), $[conv(3\times3)\times256]$
3rd Upsampling Layer	32 × 128	Upsampling (2 × 2), $[conv(3\times3)\times96]$
4th Upsampling Layer	64 × 256	Upsampling (2 × 2), $[conv(3\times3)\times96]$
5th Upsampling Layer	128 × 512	Upsampling (2 × 2), $[conv(3\times3)\times64]$
Output Layer	128 × 512	$[conv(1\times1)\times2]$

4.4. Results

We discuss the performance from experiments with RCTE when trained using 20% of the training set as an annotated subset and 80% as an unannotated set. Table 2 lists the evaluation results under some experimental settings, i.e., supervised setting, augmented supervised setting, and proposed SSL setting. The same backbone architecture was employed in all settings to guarantee fair comparisons. The Dice loss was employed to optimize the model in the first experiment using 2D slices of the annotated part of the dataset, whereas the final consistency loss was employed in the other three experiments. The traditional supervised setting had the worst segmentation performance across all measures (JI: 53.83%; DSC: 55.84%; AC: 88.91%; SE: 82.34%; SP: 95.31%), owing to the small size of the training set and the negative effect of the noisy labels. Applying random augmentation to the input (i.e., Rand + Supervised) realized improvements (JI: 1.79%; DSC: 4.59%; AC: 4.06%; SE: 4.69%; SP: 2.13%) over the supervised counterpart. This explains the effectiveness of input augmentation to improve the consistency of segmentation. It can be observed that task-driven augmentation resulted in more improvements (JI: 3%–4%; DSC: 4%–6%; AC: 3%–4%; SE: 4%–7%; SP: 1%–3%%) over the supervised segmentation performance, which explains the effectiveness of data-driven augmentation compared to its random counterpart. Finally, it is notable that RCTE realized many performance improvements (JI: 4%; DSC: 4.2%; AC: 1.05%; SE: 3.46%; SP: 0.72%) compared to other settings. This demonstrates the efficiency of RCTE in leveraging labeled and unlabeled CT scans for COVID-19 lesion segmentation.

4.5. Comparative analysis

We compare the performance of RCTE to recent cutting-edge SS-Seg approaches. Table 3 presents the segmentation performance of the competing approaches across different measures by implementing them under the same experimental settings. MT showed the lowest performance across different measures. UA-MT [36] realized desirable performance improvements over the MT model, which explains the effectiveness of measuring the uncertainty of output to improve segmentation performance. The Inf-Net architecture showed good segmentation performance. However, the multistage training of Semi-Inf-Net [11] limited the realization of the optimal performance. Compared with MT [21], SE-COPLE-Net showed large performance improvements owing to the adaptive knowledge transfer from the teacher network to the student network. In contrast with MT [21], TCSM_v2 [31] achieved much better performance due to the imposed transformation to improve the regularization effect, and hence the performance. More importantly, RCTE attained robust segmentation performance with performance improvements over competing approaches (JI: 2.95%; DSC: 2.66%; AC: 1.35%; SE: 1.26%; SP: 1.05%). Fig. 3 shows a graphical comparison of the segmentation results of different real-world COVID-19 axial slices.

To assess the performance of the new loss function, noisy annotated training images were obtained to delineate the infection lesion with non-specialist researchers [1]. Hence these annotations became unavoidably noisy because of the vague lesion borders, inter- and intra-observer inconsistencies, and the possible tendency of the preliminary pattern. Additional comparative analysis was performed to contrast the introduced NAL with three well-known noise-aware loss functions, i.e., noise-robust Dice (NR-Dice), GCE loss, and MAE loss [44]. NAL was also compared with the standard Dice and CE loss functions. The quantitative findings of these comparative experiments are presented in Table 4. CE and Dice loss had the lowest segmentation performance. The noise-robust losses (i.e., MAE, GCE, NR-Dice) seemed to be more efficient in comparison with the standard losses. Contrasted with the noise-robust losses, the performance attained by NAL significantly surpassed that of the other losses by large margins (JI: 1.1%; DSC: 1.3%), thus validating the effectiveness of NAL at reducing the effect of noisy labeled areas.

4.6. Statistical analysis

A paired sample t-test analysis was conducted to investigate the statistical significance of the results of RCTE against the competing SS-Seg approaches. Two paired-sample t-test experiments were performed on the test set using the measures of JI, DSC, accuracy, sensitivity, and specificity. These experiments were implemented using the SciPy scientific computing library [46]. The significance threshold was set to $5e^{-2}$, where a $p-value$ less than 0.05 indicates statistical significance of the results. Table 5 presents the computed $P-values$ for the test set. Most were less than 5.00E-03, implying that the results from RCTE were distinct from those of the competing SS-Seg approaches. This validates the effectiveness of RCTE.

Table 2
Comparison of supervised and semi-supervised learning (20% annotated/80% unannotated) on test set of MosMedData.

Methods	JI (%)	DSC (%)	AC (%)	SE (%)	SP (%)
Supervised	53.83 ± 11.46	55.84 ± 9.66	88.91 ± 10.35	82.34 ± 14.29	95.31 ± 9.18
Rand + Supervised	55.62 ± 29.03	58.08 ± 24.09	90.93 ± 21.55	85.57 ± 19.24	96.23 ± 25.11
TD^I + Supervised	57.87 ± 9.15	60.43 ± 16.12	92.97 ± 18.34	87.03 ± 15.11	97.44 ± 8.13
TD^D + Supervised	57.05 ± 10.14	61.67 ± 12.34	92.54 ± 12.66	90.11 ± 14.37	96.98 ± 16.21
Proposed RCTE	**61.87 ± 10.66**	**65.87 ± 9.11**	**94.02 ± 7.91**	**93.57 ± 5.46**	**98.16 ± 9.23**

Table 3
Segmentation results of different SSL models on the test set of MosMedData.

Method	JI (%)	DSC (%)	AC (%)	SE (%)	SP (%)
MT [21]	55.67 ± 19.16	59.24 ± 15.42	89.77 ± 8.17	90.01 ± 8.68	96.13 ± 7.24
UA-MT [36]	56.13 ± 21.44	60.73 ± 17.63	91.03 ± 12.49	91.16 ± 10.34	96.48 ± 8.57
Semi-Inf-Net [11]	57.73 ± 13.62	61.56 ± 9.74	90.48 ± 13.72	89.86 ± 7.02	96.31 ± 9.34
SE-COPLE-Net [1]	58.92 ± 17.23	62.73 ± 14.09	91.51 ± 10.13	92.31 ± 9.11	97.11 ± 8.67
TCSM_v2 [31]	58.11 ± 12.37	63.21 ± 12.28	92.67 ± 12.35	92.09 ± 10.06	96.89 ± 10.07
Proposed RCTE	**61.87 ± 10.66**	**65.87 ± 9.11**	**94.02 ± 7.91**	**93.57 ± 5.46**	**98.16 ± 9.23**

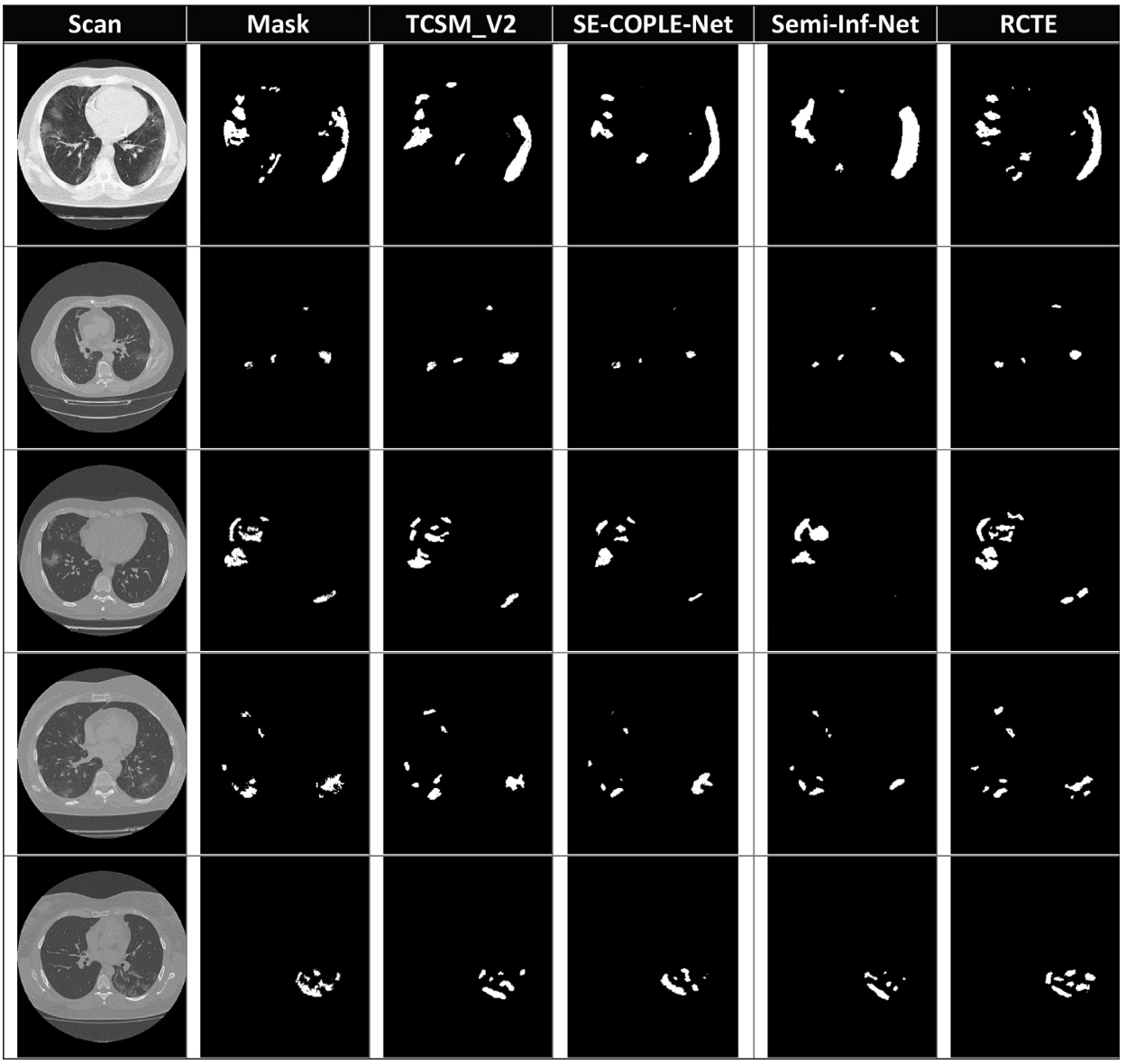

Fig. 3. Visual comparison of segmentation results produced by proposed RCTE against those generated by competing approaches.

4.7. Ablation experiments

For a deeper analysis of the performance of RCTE, ablation experiments were performed to enable understanding of the behavior of the model under different settings. In these experiments, the Dense U-Net optimized with Dice loss was selected as the baseline for RCTE.

Table 4
Comparison of segmentation performance using different loss functions on the test set of MosMedData.

Loss function	JI (%)	DSC (%)
CE	57.88 ± 14.21	61.88 ± 10.64
Dice [45]	59.61 ± 9.24	63.07 ± 13.02
GCE [19]	58.46 ± 13.42	64.46 ± 11.27
MAE [44]	57.97 ± 9.16	64.07 ± 10.47
NR-Dice [1]	60.77 ± 11.61	64.57 ± 9.78
NIL	**61.87** ± 10.66	**65.87** ± 9.11

Table 5
Results of paired t-tests for segmentation results of RCTE against competing approaches.

Experiment	JI (%)	DSC (%)	AC (%)	SE (%)	SP (%)
RCTE vs. MT	2.17E−03	5.06E−04	3.13E−02	2.24E−03	5.76E−03
RCTE vs. UA-MT	3.22E−03	2.66E−03	4.08E−03	3.27E−02	4.15E−02
RCTE vs. Semi Inf-Net	2.12E−02	8.79E−04	4.53E−02	5.17E−02	3.21E−02
RCTE vs. SE−COPLE−Net	3.37E−03	3.68E−03	1.56E−01	4.22E−02	5.64E−02
RCTE vs. TCSM_v2	4.02E−03	9.97E−03	3.47E−01	6.28E−02	4.27E−01

1) Impact of size of unannotated set

Table 6 lists the segmentation performance of RCTE when trained using different combinations of randomly selected annotated and unannotated CT scans. Semi-supervised training steadily achieved better performance than supervised training in all combinations of annotated/unannotated data settings, which validates that RCTE can effectively leverage unannotated scans to fine-tune segmentation performance. It is also observable that segmentation performance improves as a result of increasing the number of unannotated instances, which conforms to the authors' expectations. To realize such performance using small-scale annotated CT data is attributable to the employed consistency loss to essentially leverage extra knowledge from the unannotated data. Moreover, that splitting of data into 10% annotated and 90% unannotated will result in a similar performance as with 20:80 splitting scheme. Meanwhile the increasing the amount of noisy annotated samples to 30% is showing to gain slight performance improvements with 1% on JI and DSC measures.

2) Impact of mean-teacher design

The main purpose of this experiment was to validate the selection of the mean-teacher architecture for RCTE. According to the results presented in Table 7, the deployment of the baseline model in the MT architecture improved the segmentation performance by 14%, 1.53%, and 0.43% compared to JI, DSC and AC, respectively. This further justifies the significance of leveraging the annotated and unannotated data to maintain the consistency of segmentation.

3) Impact of reliable teacher

An experiment was performed to explore the influence of implementing the MT architecture with a reliable teacher network. The results are reported in the third row of Table 7. It can be seen that this resulted in significant efficiency improvements (JI: 0.59%; DSC: 2.46%; AC: 0.51%) over the baseline model. Compared with the standard MT design, it is apparent that the segmentation performance achieved good enhancements (JI: 0.45%; DSC: 0.93%; AC: 0.74%), which demonstrates the effectiveness of the reliable teacher in subduing the involvement of the student to the EMA as soon as the student exhibits higher training loss, thus preventing the possible negative impact of noisy annotations.

4) Impact of reliable student

Table 6
Evaluation results of RCTE on different combinations of annotated and unannotated data.

Annotated	Unannotated	JI (%)	DSC (%)	AC (%)	SE (%)	SP (%)
20% (212)	0% (0)	53.83 ± 11.46	55.84 ± 9.66	88.91 ± 10.35	82.34 ± 14.29	95.31 ± 9.18
20% (212)	40% (424)	56.64 ± 9.13	58.43 ± 10.87	90.09 ± 9.71	86.38 ± 10.61	96.82 ± 10.47
20% (212)	60% (636)	59.49 ± 11.29	61.87 ± 8.87	92.76 ± 9.89	90.02 ± 9.87	98.16 ± 9.23
20% (212)	80% (848)	**61.87 ± 10.66**	**65.87 ± 9.11**	**94.02 ± 7.91**	**93.57 ± 5.46**	**98.16 ± 9.23**
10% (106)	90% (954)	61.03 ± 8.79	64.97 ± 8.67	93.63 ± 10.21	93.01 ± 9.74	98.12 ± 8.97
30% (318)	70% (742)	62.47 ± 9.57	65.69 ± 10.74	94.13 ± 9.62	93.98 ± 9.35	98.79 ± 10.67

Table 7
Quantitative results of ablation experiments of RCTE for COVID-19 segmentation. Stated results indicate significant improvement from V2 ($p - value < 0.05$ according to paired t-test).

	Ablation Studies								Performance		
	MT	Reliable Teacher	Reliable Student	Random augm	TD^I	TD^D	$TD^I + TD^D$	NIL	JI (%)	DSC (%)	AC (%)
V1	✗	✗	✗	✗	✗	✗	✗	✗	53.83 ± 11.46	55.84 ± 9.66	88.91 ± 10.35
V2	✓	✗	✗	✗	✗	✗	✗	✗	53.97 ± 13.03	57.37 ± 11.42	88.68 ± 11.19
V3	✓	✓	✗	✗	✗	✗	✗	✗	*54.42 ± 12.43	58.3 ± 12.34	89.42 ± 7.02
V4	✓	✗	✓	✗	✗	✗	✗	✗	53.99 ± 9.87	*59.15 ± 16.42	89.86 ± 10.13
V5	✓	✓	✓	✗	✗	✗	✗	✗	*54.81 ± 10.78	*59.87 ± 10.97	90.18 ± 9.25
V6	✓	✓	✓	✓	✗	✗	✗	✗	*56.43 ± 8.97	*60.35 ± 15.31	91.08 ± 10.78
V7	✓	✓	✓	✗	✓	✗	✗	✗	*57.31 ± 13.88	*61.29 ± 12.11	*92.17 ± 8.62
V8	✓	✓	✓	✗	✗	✓	✗	✗	*57.91 ± 14.34	*62.84 ± 9.47	*92.81 ± 6.99
V9	✓	✓	✓	✗	✗	✗	✓	✗	*59.91 ± 11.23	*63.91 ± 10.38	*93.21 ± 9.78
V10	✓	✓	✓	✗	✗	✗	✓	✓	***61.87 ± 10.66**	***65.87 ± 9.11**	***94.02 ± 7.91**

An experiment was carried out to investigate the impact of redesigning the MT model using the reliable student architecture. The results are presented in the fourth row of Table 7. Contrasted with the baseline, this experiment resulted in significant efficiency improvements (JI: 0.16%; DSC: 3.13%; AC: 0.95%). Compared with the MT design, it is observable that the segmentation performance realized good improvements (JI: 0.02%; DSC: 1.78%; AC: 1.18%). These improvements can be justified by the capability of a reliable student to learn from the teacher when the loss of the teacher is better than the student's loss. This enables subduing of the impact of unreliable and noisy annotations. More importantly, it is notable that implementing the MT with a reliable teacher and reliable student network results in improvements of 0.84%, 2.5%, and 1.5% over JI, DSC, and AC, respectively.

5) Impact of random augmentation

Motivated by the recent TSCM_V2 that randomly applied datasets for the input images, we investigate the impact of such transformations in the proposed RCTE. From Table 7, it can be noted that applying the random transformation improved the segmentation performance of MT by 2.46%, 2.98%, and 2.4% over JI, DSC, and AC, respectively. This observation indicates the effectiveness of augmenting the input image for improving the regularization power of pixel-level segmentation.

6) Impact of task-driven transformation

The proposed RCTE employs task-driven augmentations to automatically generate intensity or deformation transformations. We performed three experiments to evaluate the impact of each kind of transformation separately and together. In Table 7, it can be noted that intensity augmentation attained better performance than random augmentation, at 0.88%, 0.94%, and 1.09% over JI, DSC, and AC, respectively. Deformation augmentation led to more improvements (JI: 1.48%; DSC: 2.49%; AC: 1.73%) over the random augmentations. More importantly, applying both intensity and deformation together enabled better performance than applying them separately.

7) Impact of Noise-aware loss

This experiment analyzed the impact of NAL on segmentation performance, as shown in the last row of Table 7. It can be seen that training RCTE with NAL resulted in substantial performance improvements (JI: 2.0%; DSC: 1.07%; AC: 0.4%), which validates the effectiveness of NAL in dealing with noisy annotated COVID-19 images.

4.8. Merits and limitations

The advantages of RCTE can be summarized as follows: 1) it enables efficient segmentation of complex COVID-19 lesions from a limited amount of annotated data; 2) it enables the leveraging of unannotated training data to improve the segmentation performance of the model; 3) it prevents unreliable targets from negatively affecting training performance; and 4) it offers noise-aware loss that enables effective learning of lesion features from noisy annotated CT scans.

This study has some limitations. RCTE just considers two kinds of task-driven transformations to be applied to input images, which might lead to suboptimal transformations, and possibly to suboptimal performance. RCTE was not tested for multi-class segmentation owing to the availability of binary masks only. RCTE does not consider relationships between input instances that could help extract valuable semantic representations from unannotated images, as noted in recent studies [29]. RCTE is designed based on the hypothesis that the data come from a single domain with a shared data distribution. Therefore, to incorporate out-of-distribution data might negatively affect performance.

5. Conclusion and future directions

We introduced a reliable and consistent RCTE framework for efficient semi-supervised segmentation of COVID-19 lesions from 2D lung CT scans. A reliable teacher-student architecture was employed to improve the superiority and reliability of ensemble predictions and mitigate the effect of defective and unreliable pseudo-annotations on segmentation loss. Noise-aware loss (NAL) was introduced to deal with noisy annotated COVID-19 CT scans. RCTE was trained to minimize a subjective mixture of supervised and regularization loss, which was implemented using NAL. Empirical evaluations showed that NAL overcomes existing noise-aware loss functions, and RCTE realized superior performance over cutting-edge self-ensembling-based medical segmentation techniques.

Our future work will investigate the effectiveness of RCTE as a general segmentation approach for similar problems in the medical domain. Another future direction is to improve RCTE to address issues stemming from cross-modality data; domain adaption techniques can offer promising solutions to these issues [47]. Motivated by the success of RCTE, the development of semi-supervised multiple task/instance learning [29] is essential to provide the ultimate diagnosis framework for COVID-19 and similar pandemics. We also intend to investigate the diagnosis of COVID-19 from lung ultrasound frames/videos in our future work.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research was supported in part by the National Natural Science Foundation of China under 61300167 and 61976120, supported in part by the Natural Science Foundation of Jiangsu Province under Grant BK20191445, supported in part by the Six Talent Peaks Project of Jiangsu Province under Grant XYDXXJS-048, and sponsored by Qing Lan Project of Jiangsu Province.

References

[1] G. Wang, X. Liu, C. Li, Z. Xu, J. Ruan, H. Zhu, T. Meng, K. Li, N. Huang, S. Zhang, A noise-robust framework for automatic segmentation of COVID-19 Pneumonia Lesions from CT Images, IEEE Trans. Med. Imaging. (2020), https://doi.org/10.1109/TMI.2020.3000314.
[2] A. Oulefki, S. Agaian, T. Trongtirakul, A. Kassah Laouar, Automatic COVID-19 lung infected region segmentation and measurement using CT-scans images, Pattern Recognit. (2021), https://doi.org/10.1016/j.patcog.2020.107747.
[3] P. Melin, J.C. Monica, D. Sanchez, O. Castillo, Multiple ensemble neural network models with fuzzy response aggregation for predicting covid-19 time series: The case of mexico, Healthc. (2020), https://doi.org/10.3390/healthcare8020181.
[4] S. Boccaletti, W. Ditto, G. Mindlin, A. Atangana, Modeling and forecasting of epidemic spreading: The case of Covid-19 and beyond, Chaos, Solitons Fractals (2020), https://doi.org/10.1016/j.chaos.2020.109794.
[5] O. Castillo, P. Melin, Forecasting of COVID-19 time series for countries in the world based on a hybrid approach combining the fractal dimension and fuzzy logic, Chaos, Solitons Fractals (2020), https://doi.org/10.1016/j.chaos.2020.110242.
[6] T. Sun, Y. Wang, Modeling COVID-19 epidemic in Heilongjiang province, China, Chaos, Solitons Fractals. (2020), https://doi.org/10.1016/j.chaos.2020.109949.
[7] S. Varela-Santos, P. Melin, A new approach for classifying coronavirus COVID-19 based on its manifestation on chest X-rays using texture features and neural networks, Inf. Sci. (2021), https://doi.org/10.1016/j.ins.2020.09.041.
[8] L. Zhou, Z. Li, J. Zhou, H. Li, Y. Chen, Y. Huang, D. Xie, L. Zhao, M. Fan, S. Hashmi, F. Abdelkareem, R. Eiada, X. Xiao, L. Li, Z. Qiu, X. Gao, A. Rapid, Accurate and machine-agnostic segmentation and quantification method for CT-Based COVID-19 Diagnosis, IEEE Trans. Med. Imag. (2020), https://doi.org/10.1109/TMI.2020.3001810.
[9] Y. Zhang, Y. Wei, Q. Wu, P. Zhao, S. Niu, J. Huang, M. Tan, Collaborative unsupervised domain adaptation for medical image diagnosis, IEEE Trans. Image Process. (2020), https://doi.org/10.1109/TIP.2020.3006377.
[10] Y. Shen, N. Wu, J. Phang, J. Park, K. Liu, S. Tyagi, L. Heacock, S.G. Kim, L. Moy, K. Cho, K.J. Geras, An interpretable classifier for high-resolution breast cancer screening images utilizing weakly supervised localization, Med. Image Anal. (2021), https://doi.org/10.1016/j.media.2020.101908.
[11] D.P. Fan, T. Zhou, G.P. Ji, Y. Zhou, G. Chen, H. Fu, J. Shen, L. Shao, Inf-Net: Automatic COVID-19 Lung Infection Segmentation from CT Images, IEEE Trans. Med. Imag. (2020), https://doi.org/10.1109/TMI.2020.2996645.
[12] N. Kumar, P. Uppala, K. Duddu, H. Sreedhar, V. Varma, G. Guzman, M. Walsh, A. Sethi, Hyperspectral tissue image segmentation using semi-supervised NMF and hierarchical clustering, IEEE Trans. Med. Imag. (2019), https://doi.org/10.1109/TMI.2018.2883301.
[13] X. Li, M. Jia, M.T. Islam, L. Yu, L. Xing, Self-supervised feature learning via exploiting multi-modal data for retinal disease diagnosis, IEEE Trans. Med. Imag. (2020), https://doi.org/10.1109/TMI.2020.3008871.
[14] X. Wu, C. Chen, M. Zhong, J. Wang, J. Shi, COVID-AL: The diagnosis of COVID-19 with deep active learning, Med. Image Anal. (2021), https://doi.org/10.1016/j.media.2020.101913.
[15] B. Lei, Z. Xia, F. Jiang, X. Jiang, Z. Ge, Y. Xu, J. Qin, S. Chen, T. Wang, S. Wang, Skin lesion segmentation via generative adversarial networks with dual discriminators, Med. Image Anal. (2020), https://doi.org/10.1016/j.media.2020.101716.
[16] C. Baur, S. Denner, B. Wiestler, N. Navab, S. Albarqouni, Autoencoders for unsupervised anomaly segmentation in brain MR images: A comparative study, Med. Image Anal. (2021), https://doi.org/10.1016/j.media.2020.101952.
[17] Z. Yu, Y. Zhang, J. You, C.L.P. Chen, H.S. Wong, G. Han, J. Zhang, Adaptive semi-supervised classifier ensemble for high dimensional data classification, IEEE Trans. Cybern. (2019), https://doi.org/10.1109/TCYB.2017.2761908.
[18] D. Karimi, H. Dou, S.K. Warfield, A. Gholipour, Deep learning with noisy labels: Exploring techniques and remedies in medical image analysis, Med. Image Anal. (2020), https://doi.org/10.1016/j.media.2020.101759.
[19] S. Laine, T. Aila, Temporal ensembling for semi-supervised learning, in: 5th Int. Conf. Learn. Represent. ICLR 2017 - Conf. Track Proc., 2017.
[20] X. Shi, H. Su, F. Xing, Y. Liang, G. Qu, L. Yang, Graph temporal ensembling based semi-supervised convolutional neural network with noisy labels for histopathology image analysis, Med. Image Anal. (2020), https://doi.org/10.1016/j.media.2019.101624.

[21] A. Tarvainen, H. Valpola, Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results, Adv. Neural Inf. Process. Syst.(NIPS (2017).
[22] Q. Liu, L. Yu, L. Luo, Q. Dou, P.A. Heng, Semi-Supervised Medical Image Classification With Relation-Driven Self-Ensembling Model, IEEE Trans. Med. Imaging. (2020), https://doi.org/10.1109/TMI.2020.2995518.
[23] Z. Yu, Y. Zhang, C.L.P. Chen, J. You, H.S. Wong, D. Dai, S. Wu, J. Zhang, Multiobjective semisupervised classifier ensemble, IEEE Trans. Cybern. (2019), https://doi.org/10.1109/TCYB.2018.2824299.
[24] S. Wu, Q. Ji, S. Wang, H.S. Wong, Z. Yu, Y. Xu, Semi-supervised image classification with self-paced cross-task networks, IEEE Trans. Multimed. (2018), https://doi.org/10.1109/TMM.2017.2758522.
[25] N. Tajbakhsh, L. Jeyaseelan, Q. Li, J.N. Chiang, Z. Wu, X. Ding, Embracing imperfect datasets: A review of deep learning solutions for medical image segmentation, Med. Image Anal. (2020), https://doi.org/10.1016/j.media.2020.101693.
[26] Z. Zhang, M.R. Sabuncu, Generalized cross entropy loss for training deep neural networks with noisy labels, in, Adv. Neural Inf. Process. Syst. (NIPS 2018) (2018).
[27] K. Gao, J. Su, Z. Jiang, L.L. Zeng, Z. Feng, H. Shen, P. Rong, X. Xu, J. Qin, Y. Yang, W. Wang, D. Hu, Dual-branch combination network (DCN): Towards accurate diagnosis and lesion segmentation of COVID-19 using CT images, Med. Image Anal. (2021), https://doi.org/10.1016/j.media.2020.101836.
[28] T. Mahmud, M.J. Alam, S. Chowdhury, S.N. Ali, M.M. Rahman, S.A. Fattah, M. Saquib, CovTANet: A Hybrid Tri-level attention based network for lesion segmentation, diagnosis, and severity prediction of COVID-19 Chest CT Scans, IEEE Trans. Ind. Informatics. (2020), https://doi.org/10.1109/TII.2020.3048391.
[29] V. Cheplygina, M. de Bruijne, J.P.W. Pluim, Not-so-supervised: A survey of semi-supervised, multi-instance, and transfer learning in medical image analysis, Med. Image Anal. (2019), https://doi.org/10.1016/j.media.2019.03.009.
[30] Z. Yu, Y. Lu, J. Zhang, J. You, H.S. Wong, Y. Wang, G. Han, Progressive semisupervised learning of multiple classifiers, IEEE Trans. Cybern. (2018), https://doi.org/10.1109/TCYB.2017.2651114.
[31] X. Li, L. Yu, H. Chen, C.W. Fu, L. Xing, P.A. Heng, Transformation-consistent self-ensembling model for semisupervised medical image segmentation, IEEE Trans. Neural Networks Learn. Syst. (2021), https://doi.org/10.1109/TNNLS.2020.2995319.
[32] J. Liang, R. He, Z. Sun, T. Tan, Exploring uncertainty in pseudo-label guided unsupervised domain adaptation, Pattern Recognit. (2019), https://doi.org/10.1016/j.patcog.2019.106996.
[33] S. Wang, Q. Wang, Y. Shao, L. Qu, C. Lian, J. Lian, D. Shen, Iterative label denoising network: Segmenting male pelvic organs in CT from 3D Bounding Box Annotations, IEEE Trans. Biomed. Eng. (2020), https://doi.org/10.1109/TBME.2020.2969608.
[34] S. Liu, K.H. Thung, W. Lin, D. Shen, P.T. Yap, Hierarchical nonlocal residual networks for image quality assessment of pediatric diffusion MRI With Limited and Noisy Annotations, IEEE Trans. Med. Imaging. (2020), https://doi.org/10.1109/TMI.2020.3002708.
[35] H. Lin, H. Chen, X. Wang, Q. Wang, L. Wang, P.A. Heng, Dual-path network with synergistic grouping loss and evidence driven risk stratification for whole slide cervical image analysis, Med. Image Anal. (2021), https://doi.org/10.1016/j.media.2021.101955.
[36] L. Yu, S. Wang, X. Li, C.W. Fu, P.A. Heng, Uncertainty-Aware Self-ensembling Model for Semi-supervised 3D Left Atrium Segmentation, in: Lect. Notes Comput. Sci. (Including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), 2019. 10.1007/978-3-030-32245-8_67.
[37] N. Charoenphakdee, J. Lee, M. Sugiyama, On symmetric losses for learning from corrupted labels, ICML, 2019, p. 2019.
[38] K. Chaitanya, N. Karani, C.F. Baumgartner, E. Erdil, A. Becker, O. Donati, E. Konukoglu, Semi-supervised task-driven data augmentation for medical image segmentation, Med. Image Anal. (2021), https://doi.org/10.1016/j.media.2020.101934.
[39] E.D. Cubuk, B. Zoph, D. Mane, V. Vasudevan, Q. V. Le, Autoaugment: Learning augmentation strategies from data, in: Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., 2019. 10.1109/CVPR.2019.00020.
[40] S. Lim, I. Kim, T. Kim, C. Kim, S. Kim, Fast AutoAugment, Neural Inf. Process. Syst. (NeurIPS 2019) (2019).
[41] D.P. Kingma, J.L. Ba, Adam: A method for stochastic optimization, in: 3rd Int. Conf. Learn. Represent. ICLR 2015 - Conf. Track Proc., 2015.
[42] S.P. Morozov, A.E. Andreychenko, I.A. Blokhin, P.B. Gelezhe, A.P. Gonchar, A.E. Nikolaev, N.A. Pavlov, V.Y. Chernina, V.A. Gombolevskiy, MosMedData: data set of 1110 chest CT scans performed during the COVID-19 epidemic, Digit. Diagnost. (2020), https://doi.org/10.17816/dd46826.
[43] F. Shan, Y. Gao, J. Wang, W. Shi, N. Shi, M. Han, Z. Xue, D. Shen, Y. Shi, Abnormal lung quantification in chest CT images of COVID-19 patients with deep learning and its application to severity prediction, Med. Phys. (2021), https://doi.org/10.1002/mp.14609.
[44] A. Ghosh, H. Kumar, P.S. Sastry, Robust loss functions under label noise for deep neural networks, in: 31st AAAI Conf. Artif. Intell. AAAI 2017, 2017.
[45] F. Milletari, N. Navab, S.A. Ahmadi, V-Net: Fully convolutional neural networks for volumetric medical image segmentation, in: Proc. - 2016 4th Int. Conf. 3D Vision (3DV (2016), https://doi.org/10.1109/3DV.2016.79.
[46] P. Virtanen, R. Gommers, T.E. Oliphant, M. Haberland, T. Reddy, et al, Author Correction: SciPy 1.0: fundamental algorithms for scientific computing in Python (Nature Methods, (2020), 17, 3, (261-272), 10.1038/s41592-019-0686-2), Nat. Methods. (2020). 10.1038/s41592-020-0772-5.
[47] K. Li, S. Wang, L. Yu, P.-A. Heng, Dual-Teacher++: Exploiting Intra-domain and Inter-domain Knowledge with Reliable Transfer for Cardiac Segmentation, IEEE Trans. Med. Imag. (2020), https://doi.org/10.1109/tmi.2020.3038828.