



Semi-Supervised Segmentation of COVID-19 Infection on CT-Scan Lung Using Dynamic Mutual Training

Ersa Resita¹ Tita Karlita^{1*} Riyanto Sigit¹ Eko Mulyanto Yuniarno²
 I Ketut Eddy Purnama² Mauridhi Hery Purnomo^{2*}

¹ *Department of Informatics and Computer Engineering,
 Electronics Engineering Polytechnic Institute of Surabaya, Indonesia*

² *Department of Computer Engineering, Faculty of Intelligent Electrical and Informatics Technology,
 Sepuluh Nopember Institute of Technology, Indonesia*

* Corresponding author's Email: tita@pens.ac.id, hery@ee.its.ac.id

Abstract: Corona virus disease 2019 (COVID-19) 's global pandemic has caused the world to face a health crisis. Automated detection of COVID-19 infection from computed tomography (CT-scan) images has improved healthcare for treating COVID-19. However, segmentation of infected areas on CT-scan images of the lungs faces several challenges: detailed infection characteristics and low contrast differences between CT scans of infected lungs. It has a low data scale with a doctor's statement because it is still a new case, with a lot of data with pseudo labels, while pseudo labels have a low confidence level and a high error rate. Therefore, using the data of 1600 pseudo label images and 50 doctor label images, we apply pseudo supervision as the core idea, mutual training between two different models with a dynamic loss function called dynamic mutual training (DMT). DMT will do multi-training on pseudo labels with doctor's labels to be trusted in area segmentation. The results obtained are the most superior value of 91.32% with a loss value of 0.19 dice score 0.23, IOU 0.781, precision 0.843, sensitivity 0.753, and specificity 0.845. We also compare our method with other segmentation methods such as UNET, which is highly preferred in terms of medical images, and mask RCNN, which shows the best method in terms of segmentation. This comparison indicates that DMT provides the best experimental incentive with a dice score value of 2-30%, superior to cases segmentation areas affected by COVID-19 on CT scans of the lungs.

Keywords: COVID-19, CT-scan, Infection segmentation, Semi-supervised.

1. Introduction

Image segmentation is an essential process in computer vision. It involves separating visual input into segments to simplify image analysis. A segment represents an object or part of an object and comprises a collection of pixels, or "superpixels." Image segmentation sorts pixels into more significant components, eliminating the need to consider individual pixels as observation units. It is the process of grouping something larger into smaller parts based on specific characteristics. The image segmentation process separates the object (foreground) from the background. In general, the output of image segmentation is a binary image where the desired object (foreground) is white (1), while the

background to be removed is black (0). In the case of COVID-19 segmentation in recent studies [1, 2], typical signs of infection can be observed from CT slices, for example, ground-glass opacity (GGO) in the early stages and pulmonary consolidation in the late stages. Qualitative evaluation of infection and longitudinal changes in CT slices can provide essential benefits and information in the fight against COVID-19. However, manually depicting lung infections is tedious and time-consuming [3]. In addition, the explanation of infection by the radiologist is a highly subjective task, often influenced by individual bias and clinical experience [4].

From the above problems, we propose a segmentation of the COVID-19 system based on CT-

scan lung images using the CNN method, called DMT. This system analyzes the location and the region of the COVID-19 virus spread in the lungs using CT-scan images by segmentation. We are automatically using machine learning (CNN). In this experiment, we used the DMT method as a basic model for segmenting the lungs in areas affected by COVID-19. The main contributions of our work are:

- We carried out various data augmentation methods to improve the consistency of the results, such as random scale, random horizontal flip, and random cropping.
- In the segmentation process, we use the CNN architecture widely used in the medical image segmentation process. This architecture is similar to the U-net architecture.
- The training was carried out using the semi-supervised DMT method, which managed to outperform the segmentation results because it took into account small labeled data.
- For the evaluation matrix using IOU, Dice, Precision, sensitivity, and precision which are proven to be able to display good evaluations in segmentation cases

As we know that a COVID-19 pandemic is an event of the spread of the coronavirus disease which has spread very quickly in almost all countries in the world. This disease is caused by a severe acute respiratory syndrome that can be transmitted, namely coronavirus 2 (SARS-CoV2). As of December 10, 2021, more than 268 million cases have been reported from 219 countries worldwide, resulting in more than 5.28 million deaths and over 231 million recoveries [5]. With this large amount of data, the world health organization (WHO) urges all countries and communities to immediately carry out more effective prevention and treatment by conducting more tests, especially on suspected patients [6].

All efforts and resources have been made to fight this pandemic, one of which is developing a diagnosis. RT-PCR has been developed for COVID-19 screening as the standard method for testing obtained specimens. RT-PCR has proven to be superior and widely used for COVID-19 sampling than other methods such as nasopharyngeal swabs, oropharyngeal swabs, bronchoalveolar lavage, or tracheal aspiration. RT-PCR can detect viral RNA and some new findings. RT-PCR can detect viral RNA and some new findings. However, RT-PCR testing has a low sensitivity, almost 71 percent, thus requiring repeated testing for a proper diagnosis [7]. Computed tomography (CT) imaging is essential in detecting lung infections associated with COVID-19. It is practical and has been widely used to assess and evaluate disease evolution, patchy ground-glass

opacity (GGO) with consolidation frequently found on CT images. As a sign of lung infection. Thus, quantitative assessment of these lung lesions can aid in the diagnosis. [8].

Artificial intelligence (AI) methods have significantly progressed in the last decade. The abundance of data has achieved high accuracy in many tasks, including machine vision and medical diagnosis. In the case of COVID-19, convolutional deep neural networks (CNN), a type of AI method developed to solve vision problems, may play an essential role in disease diagnosis using CT scans. By localizing and categorizing infections, misdiagnosis can be reduced and can assist clinicians in quantifying lesions and disease stages [9].

2. Related method

Recently, a deep learning system has been proposed to detect COVID-19 infected patients using radiological imaging [3, 4]. The popular semi-supervised learning method is described and practiced by many researchers. For example, Deng-Ping Fan et al. [10] used a method to reduce labeled data, namely a semi-supervised framework using the INF-Net architecture. This study obtained a dice value of 0.682, a sensitivity of 0.692, a specificity of 0.943, and an MAE of 0.082. These results are quite effective and can be developed further.

Due to the lack of datasets, Issam Laradji et al. [11] used three different data sets to categorize the monitored consistency by calculating the loss consistency. This study used an ImageNet architecture normalized in the preprocessing section with an IOU evaluation matrix, dice coefficient (F1 score), positive predictive value (PPV), sensitivity, and specificity yielded values of 0.58, 0.41, 0.46, 0.80, 0.82, respectively. The research has shown promising results, but there is no evaluation from doctors or experts. Therefore, the accuracy of the segmentation results cannot be trusted.

Tanvir Mahmud et al. [12], used three different data sets in the segmentation, namely two COVID-19 data and one city view drone data, using a joint optimization scheme in the preprocessing process, and the CovSegNet architecture with multi-encoder resulting in an IOU value of 64.6, dice 59.5, sensitivity 76.4, and specificity 87.7. The work in this research is still too complicated because 3D images are converted into 2D images. Still, overall it has produced good results, but it may be developed in a simplified way.

Naveen Paluru et al. [13] used three different data sets to amplify the segmentation annotation results.

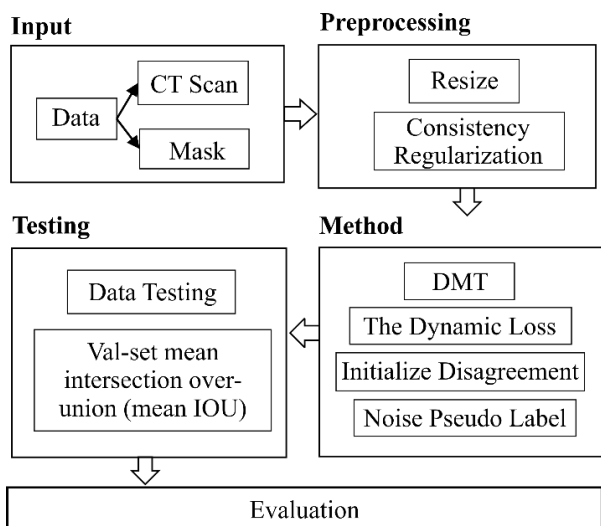


Figure. 1 System diagram

Table 1. Size of the dataset

Dataset				
Type	Total	Size		
CT-Scan	1600	336 x 493	332 x 486	231 x 384
	1600	336 x 492	333 x 488	231 x 383
	1600	334 x 491	235 x 320
Mask	1600	235 x 320	281 x 395	267 x 385
	1600	237 x 319	275 x 383	235 x 320
	1600	194 x 279	248 x 360
Edge	1600	235 x 320	281 x 395	252 x 374
	1600	237 x 319	275 x 383	235 x 320
	1600	194 x 279	248 x 360

They used the normalization of lung extraction segmentation abnormalities in the preprocessing section and introduced a new CNN architecture called Anam-Net (light-based anamorphic depth embedding). The result had an accuracy score of 0.990, a dice score of 0.972, a sensitivity of 0.959, and a sensitivity of 0.997.

Yifan Jiang et al. [14] proposed cGAN-based COVID-19 CT as an image synthesis method that can produce realistic CT images covering the two main types of infection; opacity ground-glass and consolidation. The proposed method takes the semantics segmentation map of the appropriate pulmonary CT images, and The cGAN structure studies the characteristics and information of the CT image. This study used 829 lung CT slices in the preprocessing process using various augmentation methods: random resizing and cropping, random rotation, Gaussian noise, and elastic.

Yu-Huan Wu et al. [15] built a large-scale COVID-19 classification and segmentation (COVID-CS) data set and also developed a joint classification and segmentation system (JCS) for COVID-19 diagnostics. In our system, the classification model

has identified whether the patient is a positive suspect for COVID-19 or not, accompanied by a convincing visual explanation. This study yielded a sensitivity of 95.0% and a specificity of 93.0% in our COVID-CS dataset classification assay.

3. Method

In this section, we discuss the methods used in the experiment. The segmentation process requires several stages image input, preprocessing, segmentation and evaluation. We present these stages in the form of a system diagram shown in Fig. 1.

Fig. 1 shows four main stages in the research, starting from data input, preprocessing, training, testing, and evaluation.

3.1 Input

The input data uses training data with two different data types: CT-Scan images of COVID-19 patients and images of COVID-19 segmentation results.

3.2 Pre-processing

The preprocessing stage has two stages, namely resizing and regulation consistency. Each preprocessing stage is described below:

Resize. The datasets collected previously have various sizes that are very diverse. For this reason, it is necessary to change the data size. One way to equalize the input image size is to resize it. Resize is done with a size of 235 x 320 pixels taken from the largest size in the dataset. Resizing is done to facilitate the next learning process. The size variations in our data are shown in Table 1.

Table 1 shows data size variations in training data types with pseudo labels with 1600 data on each CT-Scan, mask, and edge image. The data varies from the largest size of 336 x 486 pixels to the smallest of 231 x 383 pixels (with blue bold font). Based on these data, it was found that each type of image has the largest size, which is 235x320 pixels. On that basis, we resize to 235x320 pixels so that the data have the same size.

Consistency regulation. Common methods or techniques used for semi-supervised and self-supervised learning. Consistency regulation plays a similar role to data augmentation or data augmentation methods.

Like research done, Hendryckction et al. [16], recently conducted an experiment showing that

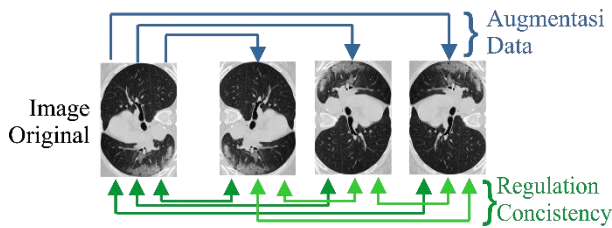


Figure. 2 Comparison of how data augmentation methods work in general and explicit consistency regulation methods

regularization naturally carries resistance at test time to corrupted data. This experiment studied the relevance of consistency regulation for training time resistance to noisy labels. Erick Engleson et al. [17] used two valuable observations on the consistency of networks trained with cross-standard entropy loss in noisy data sets. The networks trained on noisy data had lower consistency values, and the consistency was reduced more significantly around noise-labeled training data points than correctly labeled ones. This technique produces different levels and types of noise and achieves sophisticated results. Han Zhang [18] used the generative adversarial networks (GAN) method. GAN is a regularization technique to stabilize training. The experiment proposed a simple and effective training stabilizer based on the idea of consistency regularization—a popular technique in semi-supervised learning literature. Specifically, this study augmented the data that bypassed the GAN discriminator. Then, the data penalized the sensitivity of the discriminator to this augmentation. Then, a series of experiments demonstrated that consistency regularization works effectively with spectral normalization and various GAN architectures, loss functions, and optimizer settings. Our method achieves the best FID score for unconditional image creation compared to other regularization methods. Consistency regularization is done to get consistent prediction results when the data is disturbed by data augmentation. In this experiment, we employ a regular explicit consistency regulation method on data augmentation using random augmentation where random scaling, random truncation, and random reversal are shown in Fig. 2.

It can be observed that both methods have the same principle and similar results. In general, data augmentation methods are more straightforward than consistency regulation.

3.3 Training

At the training stage, we used the DMT deep learning method. In this DMT method, there is dynamic loss to calculate the loss value, Initialization

of disagreement to determine the value of disagreement between models, and pseudo label noise to filter data with excess noise. The following is an explanation of each method used:

DMT. This architecture uses a new perspective, reciprocal training between two models with a dynamically re-weighted loss function. By measuring the disagreement between models determined by comparing the predictions of two different models (model A and model B) to derive the value of losses in training dynamically. A greater disagreement indicates error probability and corresponds to a lower loss value.

We apply the architecture created by Zhengyang Feng et al. [19], which proved to be able to counteract pseudo surveillance noise with a re-weighted loss function based on the model disagreement and produce an IOU value of 7.85 ± 0.29 . Applying the DMT method to the Covid-19 CT-scan dataset and modifying it is expected to provide better values.

In this architecture, the focus is on semi-supervised learning to reduce labeled data. Semi-supervised learning only labels a small part of the data set and exploits the remaining part as unlabeled (pseudo) data. To learn pseudo labels, apply "bootstrapping" (pull yourself by bootstrapping yourself), i.e., using self-monitoring (pseudo). The disadvantages of pseudo labels are as follows:

- First, true pseudo labels with low confidence are often ignored. To achieve a low error rate for monitoring pseudo labels, most of the true pseudo labels with low confidence must be discarded.
- Second, high confidence errors do exist. A new view of the inner conflict model emerged to overcome this limitation. In particular, there is only one model to overcome the error regardless of which pseudo label selection metric is used. The probability of different models confidently making the same error is lower. Most false labels will have less impact on learning. Fig. 3 is a chart to measure disagreement between models between different models.

The input is a dataset image with two labeling methods, namely doctor label and pseudo label, then different Initialization is carried out and processed in two models. The two models train data with different labels. One model provides pseudo-supervision for the other (pre-pseudo label) so that the pre-pseudo label can be more trusted in training. Noise-loss power (dynamic loss) was introduced at the semi-supervised

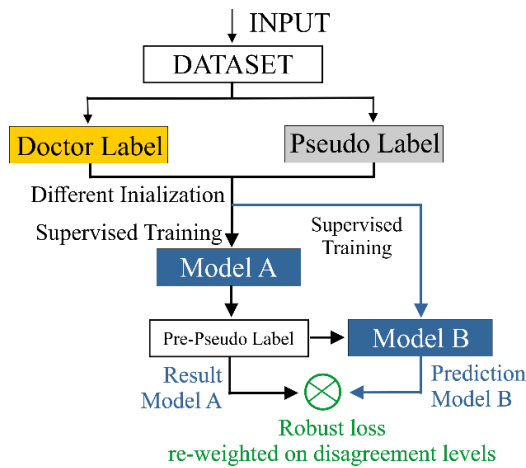


Figure. 3 Overview of training framework

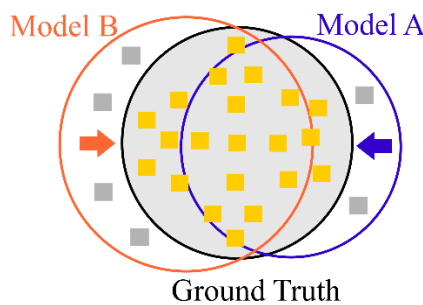


Figure. 4 Illustration of disagreement between models

Algorithm 1: DMT process code in segmentation

Input: The data set is labeled pseudo (Su) and labeled doctor sub (So)
Output: The last best model (F)

If: Starting from giving different pre-workout weights, then:

Initialization Model A (F_A^0) and Model B (F_B^0) with different pre-train weights

- Train F_A^0 on St
- Train F_B^0 on St

Else:

Initialization F_A^0 and F_B^0 with the same pre-train weight
 SA.SB = Maximized Sampling Difference (St)

- Train F_A^0 pada SA
- Train F_B^0 pada SB

$\alpha = \{20\%, 40\%, 60\%, 80\%, 100\%\}$

Foreach iteration i {1, 2, 3, 4, 5} do

- Predict and store the top i pixel of each class in Su with $F_A^{i-1} \rightarrow$ pseudo labeled set Sp
- Perfect F_A^i from F_A^{i-1} on the latest St and Sp with dynamic loss
- Predict and store the top i pixel of each class in Su with $F_B^{i-1} \rightarrow$ pseudo labeled set Sp
- Perfect F_B^i from F_B^{i-1} on the latest St and Sp with dynamic loss

F = The best (F_A^5, F_B^5)

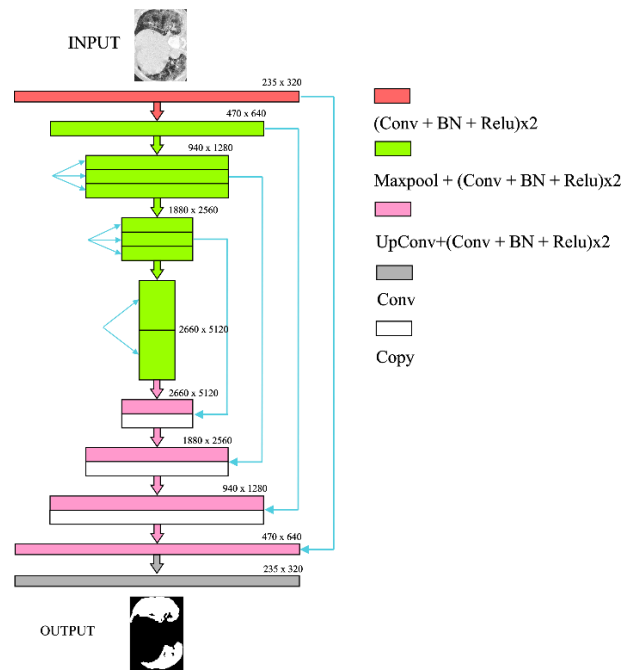


Figure. 5 Network architecture

training stage, utilizing an inter-difference model based on the predictions and beliefs of the two models. Furthermore, implementing DMT will gradually exploit pseudo data for better performance. Note that other disagreement-based semi-supervised learning methods use different models and learn by maximizing their agreement on quasi-data [20, 21]. In contrast, see disagreement [22] as a principle, i.e., the inter-disagreement model provides learning possibilities (Fig. 4).

Yellow and gray squares represent positive samples and negative samples. The result values of model A and model B have inter-model disagreements that allow for an increase in performance to the ground truth. Below is the algorithm 1 used:

The architecture used in each model (model A and model B) uses the same architecture and is shown in Fig. 5.

The architecture used in this experiment is similar to U-Net. This architecture is designed and widely applied for medical image segmentation, with maximum results compared to other architectures. As shown in Fig. 5, firstly, there is a contraction path (left side of the architecture) where the image is sampled downwards, and there is an expansive path (right side of the architecture) where the image is sampled upwards. They have a skip connection between these contraction and expansion pathways (grey arrow). Two 3x3 convolutions brought the

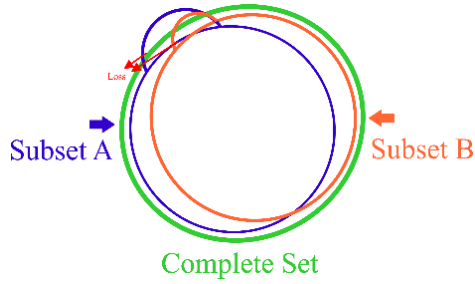


Figure. 6 The sampling difference is maximized. The complete set is shuffled at random initially, and subsets A and B are drawn the same size but with the least overlapping samples

number of feature channels to 64, followed by max-pooling kernel size 2 and step two responsible for down-sampling. The same approach of doubling the number of feature channels each time until the sampling finally starts. In the expansive path, up-sample the image using Conv2D Transpose. Connections are passed and combined along the channel dimensions when up-sampling the image, two 3x3 convolutions, with a feature count of 512 and then up-sampling. Unlike contraction paths, half the number of feature channels here each time the image is up-sampled. In the last layer, use one Conv. additional 1x1 to convert the number of channels into the number of classes owned.

Using two different models with different initialization weights, namely training on two different labeled subsets. After making various comparisons, the comparison of the set with the best results is 1:20 (for model A) and 1:8 (for model B). 1 is the value for the doctor label with 50 images and 1000/400 pseudo label images.

The dynamic loss. We use the dynamic loss, where the quantified inter-model disagreement is the dynamic loss weight. Dynamic weight loss ω_u is defined as:

$$\omega_u = \begin{cases} p_B^{\gamma 1}, & y_A = y_B \\ p_B^{\gamma 2}, & y_A \neq y_B, c_A \geq c_B \\ 0, & y_A \neq y_B, c_A < c_B \end{cases} \quad (1)$$

The dynamic loss on pseudo label samples Lu is then defined as:

$$Lu = \frac{1}{N} \sum_{u, y_A \in U} \omega_u CE(y_A, FB(u)), \quad (2)$$

Where $CE(\cdot)$ is the cross-entropy loss, intuitively, pseudo-labeled data exist in three different cases in training. The three cases are described below:

1. Agreement. FB agrees with the pseudo label.
2. Negative disagreement. FB disagrees with the pseudo label, but the confidence in FB 's decision is lower than the pseudo labels.
3. Positive disagreement. FB disagrees with the pseudo label and has higher confidence. Note that training uses the labeled subset along with the pseudo labeled data. The loss for data labeled LX remains unchanged i.e., the typical cross-entropy loss. Below is the formula for the loss of entropy LX :

$$Lx = \frac{1}{N} \sum_{u, gt \in x} CE(gt, FB(x)), \quad (3)$$

Where x and gt denote image and ground truth pairs. The combined loss L is defined as:

$$L = Lx + Lu \quad (4)$$

With regard to semantic segmentation, ω_u^{HxW} is a pixel-wise map (H for height and W for width), the re-weighting strategy remains the same and applies to each pixel

This experiment uses the loss function pixel-wise cross-entropy loss, where this loss can examine each pixel individually and compare the predicted pixel vector with the target pixel vector. After the data is trained using the DMT method, the training model is then tested on the validation data and the loss results.

Initialization diss-agreement. The main problem in exploiting the disagreement between the models is how to initiate the different model initialization. For tasks that require a pre-training load to function correctly, e.g., semantic segmentation. Different pre-training weights are hard to get, and extra time is needed. It is, therefore, necessary for new pre-training to be compared with existing assignments. Fig. 6 shows the illustration of a diss-agreement.

Pseudo-label noise. This function is already included in the DMT package, where the way it works is illustrated in Fig. 7.

The illustrative example above is implemented in model A. Because it may have a large pseudo label noise, the pseudo labeling is carried out by training on model A. Model A will produce three cases, namely prediction 1, prediction 2 and prediction 3. There are three possible cases in joint training and three appropriate loss weighting strategies based on the disagreement between the two models. The three possible cases will be compared with the

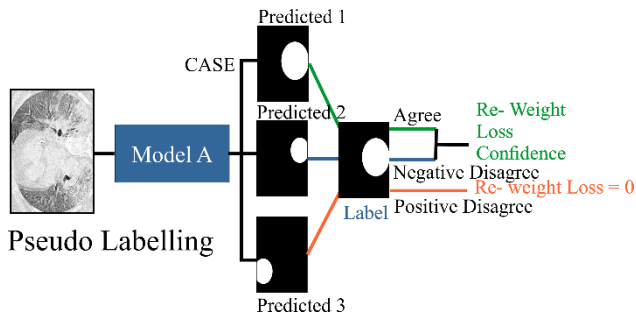


Figure. 7 Pseudo label noise diagram

label/ground truth results. The results will be three possible decisions: agree, negative disagree, and positive disagree. Agree and negative disagree decisions will be considered for training on model B, and positive disagree decisions will be ignored.

3.4 Testing

We tested the system using testing data on the dataset in the testing phase. We report the performance of the testing-set mean intersection over-union (mean IOU) test in the segmentation task, with the test data not using data already used in the training process

3.5 Evaluation

At the evaluation stage of the results, we present a Matrix Evaluation with various methods. The methods used are The Dice Similarity Coefficient (DSC), IOU, precision, sensitivity, and specification. The details of each method are described below:

The dice similarity coefficient (DSC). DSC calculates the area of overlap between two input and output images divided by the total number of pixels in both images. Covid-19 infection, cross-entropy loss evaluates the class predictions for each pixel vector one by one and then averages it for all pixels. This can be a problem if your various classes have an unbalanced representation in the image, as the most common classes can dominate training. Milletari et al. [23] have proven effective in equaling this problem by implicitly establishing a balance between foreground and background classes. So, for the same scenario used in 1 and 2, the following calculations will be performed:

$$\mathcal{L}_{Dice} = 1 - \frac{2 \sum_i^N p_i g_i}{\sum_i^N p_i^2 + \sum_i^N g_i^2} = \frac{\sum_i^N (p_i - g_i)^2}{\sum_i^N p_i^2 + \sum_i^N g_i^2} \quad (5)$$

L of the dice, that is, the weighting of each class is based on the reciprocal of the volume. The formula is shown in Eq. (6). Therefore, the L NR-dice is strong against noisy labels and foreground imbalance simultaneously. A more straightforward equation formula is as follows:

$$DSC = \frac{TP}{TP + FP + FN} \quad (6)$$

There are four conditions when the segmented image is compared with ground truth, namely:

- True positive (TP) is the number of pixels with a value of 1 in the image that corresponds to a dataset of 1.
- True negative (TN) is the number of pixels with a value of 0 in the image that corresponds to the dataset with a value of 0.
- False negative (FN) is the number of pixels with a value of 0 in the image that corresponds to a dataset of 1.
- False positive (FP) is the number of pixels with a value of 1 in the image corresponding to a dataset with a value of 0.

Comparison of definitions of TP, TN, FP, and FN based on the presence or absence of areas detected by COVID-19

The dice function is specially designed for segmentation tasks, dealing with imbalances between foreground and background pixels and noisy labels simultaneously. Experimental results [19] with CT images of 558 COVID-19 patients demonstrated the effectiveness of the dice function. The dice coefficient is very similar to IOU. They are positively correlated, meaning that if one says model A is better than model B at segmenting images, the other will say the same. Like IOU, they range from 0 to 1, with 1 indicating the most significant similarity between prediction and truth.

Intersection over union (IOU). A method to measure the percent overlap between ground truth and prediction output. The IOU metric works similarly to the DSC, which is often used as a loss function during training in segmentation cases. The IOU metric measures the difference in the number of pixels between the ground truth and the prediction and then is divided by the total number of pixels present in both.

$$IoU = \frac{Target \cap Prediction}{Target \cup Prediction} \quad (7)$$

The intersection ($A \cap B$) is obtained from the pixels in the prediction results and ground truth, while the

union (AUB) only consists of all pixels in the prediction or target mask. IOU calculates the overlap between the predicted segmentation and the underlying truth divided by the combined area between the predicted and the underlying truth. This metric ranges from 0–1 (0–100%), with 0 indicating no overlap and 1 indicating perfectly overlapping segmentation.

Precision. It is used to display positive detection results against ground truth. Of all the predicted objects in a particular drawing, how many of those objects have a matching basic truth annotation. The formula for finding the precision value is:

$$Precision = \frac{TP}{TP + FP} \quad (8)$$

It is also used to calculate the ratio of the correct positive predictions compared to the overall positive predictive results. In the research conducted, the precision value is accumulated in the form of a percent with a range of 1-100%. For example, a precision value of 80% means that eight out of every ten labels detect the number of correct pixels (a value of 1 in the image corresponding to the data set 1). The remaining 20% means that two out of every ten labels detect the number of incorrect pixels (a value of 0 in the image according to the data select 0).

Sensitivity. It is used to describe the completeness of positive predictions concerning the ground truth. A number shows how many people are correct (true positives) compared to people detected correctly by the diagnostic tool (true positives + false negatives). The fewer false negatives, the higher the sensitivity. The formula for finding the sensitivity value is:

$$Sensitivity = \frac{TP}{TP + FN} \quad (9)$$

It is used to calculate the ratio of true positive predictions compared to the overall true positive data. In the research conducted, the precision value is accumulated in the form of percent with a 1-100% range. For example, the sensitivity value is 70%, meaning that seven out of every ten labels detect the correct number of pixels (a value of 1 in the image corresponding to data set 1). The remaining 20% means that three out of every ten labels detect a true false number of pixels (the value 0 in the image corresponds to the data select 1).

Specificity. A number shows how well the system is in assessing whether it fits, calculated by the number

of correct and correct detected (true negatives) divided by all that we predicted correctly (true negatives + false positives). The fewer false positives, the higher the specificity. The formula for finding the specificity value is:

$$Specificity = \frac{TN}{TN + FP} \quad (10)$$

Specificity is used to calculate the correctness of negative predictions compared to negative data. In the research conducted, the precision value is accumulated in the form of percent with a 1-100% range. For example, a specificity value of 60% means that an average of six out of every ten labels detects the correct number of pixels (a value of 0 in the image corresponding to a data set of 0). The remaining 40% means that four out of every ten labels detect the wrong number of pixels (a value of 1 in the image corresponds to a select 0 data).

4. Experiment

Below we first describe the details behind the various COVID-19 CT lung datasets used in our experiments. Then we present the experimental setup and the basic approach we applied. Finally, we show the main experimental results and the evaluation metrics used in the experiment.

4.1 Covid-19 segmentation dataset

The data used in this experiment is secondary data. Secondary data is taken from an open-source website [23] that provides COVID-19 CT-scan images and the results of COVID-19 CT-scan image segmentation results. CT-scan data of the lungs were obtained from various patients affected by the COVID-19 virus. The CT-scan of the lungs for COVID-19 was in .jpg and .png formats. The details of the data obtained are described in Table 2.

There are two types of datasets from the datasets obtained, namely training datasets and testing datasets. The training dataset has two types of label data: doctor labeling, in which doctors label using specific tools, and pseudo labeling, namely using a labeled data model to predict labels for unlabeled data [24] and using tools [25]. Each data has a total of 50 images and 1600 images for each type of CT-scan image (Covid-19 CT scan image), mask (segmentation result from COVID-19 CT scan), and edge (segmentation result in the form of edge detection).

Total CT-scan data of the lungs for COVID-19 are 1700 images containing patient data CT-scan of

Table 2. Dataset details

Dataset Type	Information		Number of datasets	Format
Trainin g Data	Doctor- labeling	CT- Scan	50	.JPG
		Mask	50	.PNG
		Edge	50	.PNG
	Pseudo- labeling	CT- Scan	1600	.JPG
		Mask	1600	.PNG
Edge		1600	.PNG	
Testing Data	CT- Scan		50	.JPG
	Mask		50	.PNG
Total Data	CT- Scan		1700	.JPG
	Mask		1700	.PNG
	Edge		1650	.PNG

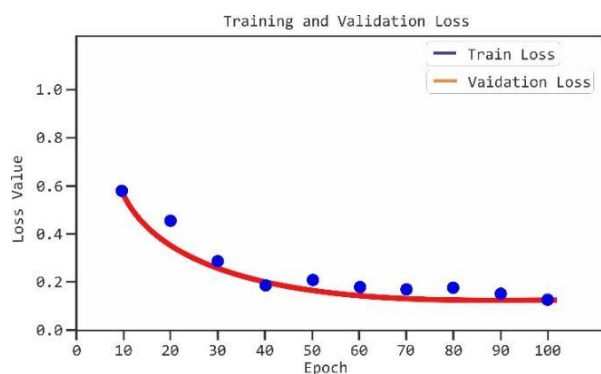


Figure. 9 Results of training and validation loss

Table 3. Details of adding a dataset

Dataset Type	Information	Number of datasets	Format
Testing Data	Ct- scan	350	.JPG
	Mask	350	.PNG

Table 4. Results % IOU

Method	1/10	1/32	Oracle
DMT	67.37 (-5.13)	46.66 (25.84)	72.50
	71.4 (-0.29)	53.79 (17.32)	71.11
	72.70 (-2.05)	83.04(-11.71)	74.75
	84.80 (-3.36)	73.80 (-14.36)	86.16

the lungs for COVID-19, 1700 images with CT-masks and labels, and 1650 images with edge detection results for CT-scan half COVID-19 lungs. Due to the lack of data for testing, data collection was carried out on other open sources [26]. The results were obtained as shown in Table 3.

There are two types of data for testing data, namely CT-scan and mask, each of which amounts to 350 images in .JPG and .PNG formats. Then the total data that has been obtained is tested.

Table 5. Experiments using several augmentation methods at the consistency regulation stage

Method	1/8	1/20	Oracle	Augmentation
DMT	84.80 (-3.36)	73.80 (-4.36)	86.16	-
DMT	86.87	85.42	87.31	Random Scale
DMT	80.01	85.21	88.90	Random Horizontal Flip
DMT	82.34	84.78	89.89	Random Crop
DMT	90.12	88.56	91.32	Random Augmentation

4.2 Segmentation result

Before discussing the result segmentation section, we first show the training process results on our system. The training results are illustrated in a graph that is shown in Fig. 9.

Based on the graph, it can be analyzed that the loss value gets good results because the graph gradually decreases towards 0. The loss validation graph follows the training loss, which means the system is very good at computing (the orange line indicates training loss and the blue dot indicates loss validation). The larger the epoch, the smaller the loss value (the better), and the best loss value of 0.19 using 100 epochs.

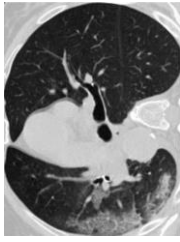



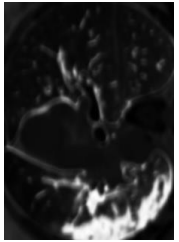
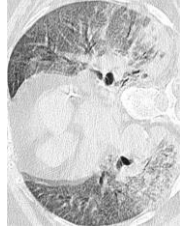


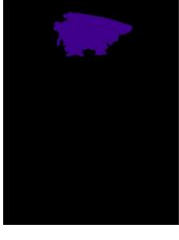
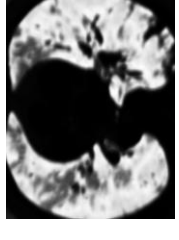
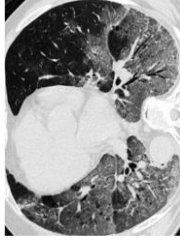



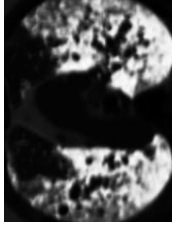
After getting the best computational value, testing and comparisons are carried out on various labeled datasets at different data ratios [19]. Then the segmentation results are obtained from various scenarios and get the best value. The segmentation results are displayed using the IOU percentages shown in Table 4.

Based on the data above, we tested with various data ratios. We use four data ratio scenarios, namely 1/10, 1/5, 1/12, and 1/8 for the ratio in model A and the data ratio is 1/32, 1/15, 1/26, and 1/20 for the ratio in model B. The results show that the best comparisons are made on the 1/8 model A and 1/20 model B data sets (in the table with blue bold font). The value of the model ratio means 50 data is equivalent to a scale of 1, 1 for doctor labels and 8/20 for pseudo labels with the percentage of oracle values (level of disagreement) reaching 86.16%.

After getting the best comparison of the data without regulatory consistency, an experiment was carried out with the addition of preprocessing, namely regulatory consistency, to test whether the data could increase the IOU value. Regulatory consistency was carried out using several data augmentation methods shown in Table 5.

Several augmentation methods used in the test are random scale, random horizontal flip, random crop,

Table 6. Comparison of Segmentation Results with other Methods

CT Image	Ground Truth	DMT	Mask R-CNN	U-NET
				
				
				

and random augmentation. Each augmentation method has a good impact on the segmentation results because it can test their consistency. Based on testing the best results when using random augmentation (in the table marked in blue bold font), the data will be more diverse and produce the best value compared to other augmentation methods, achieving 91.32% results.

We used the 1600 pseudo label images and 50 doctor label images, the pseudo-supervised DMT method we proposed as the core idea managed to get the best value. Using the principle of distrust between two different models by calculating the dynamic loss function. DMT will do multi-training on pseudo labels with doctor's labels so that they can be trusted in segmented areas. This study emphasizes the use of pseudo labels as much as possible to help research because of the lack of labeled data obtained. We get good results by adding preprocessing and data augmentation methods and using multiple data ratios.

4.3 Comparison

Furthermore, the segmentation results using the DMT method and regulatory consistency are compared with other recent segmentation methods to prove that DMT has the best segmentation results. In this research, we use two comparison methods,

namely Marsk RCNN and UNET. Mask RCNN is very popular when it comes to object detection and segmentation, as it has a conceptually simple, flexible, and generic framework for object instance segmentation. This approach efficiently detects objects in the image while generating a high-quality segmentation mask for each instance. The R-CNN mask is also easy to generalize to other tasks [27]. Furthermore, UNET is also popularly used in medical image segmentation. The U-Net architecture is symmetrical and consists of two main parts: paths contract (encoder) and expansive path (decoder). The first section extracts the features associated with the categories that pixels belong to. The second section uses convolution sequences and high-resolution sequences. The first part attributes generate a mask containing local information and categorization as output [28]. The two newest methods that proved reliable in performing our segmentation task were compared with our DMT model. The results of the comparison between DMT, R-CNN, and UNET are shown in Table 6.

Table 6 shows the visualization of the CT covid image, ground truth, the segmentation results from the comparison method used, namely the RCNN and UNET masks, as well as the segmentation results from the method used, namely DMT. It can be seen

Table 7. The results of the evaluation matrix with several other segmentation methods.

Method	Param.	Dice	IOU	Prec.	Sen.	Spec.
DMT (our)	12.18 M	0.723	0.781	0.843	0.753	0.845
Mask R-CNN	7.85 M	0.441	0.431	0.329	0.421	0.278
U-Net	9.65 M	0.708	0.719	0.678	0.836	0.665

that mask RCNN had poor results in segmenting COVID-19 in this experiment, followed by UNET. DMT outperforms other methods of segmenting COVID-19 by displaying visuals that resemble ground truth. We also present quantitative data to obtain details of the comparison values obtained in the tests [10]. Table 7 shows the value of the evaluation matrix.

Table 7 shows the results of our matrix evaluation using dice, IOU, precision, sensitivity, and Specificity calculations performed on the DMT, mask R-CNN, and UNET segmentation methods. mask R-CNN produces a dice value of 0.441, and U-Net produces a dice value of 0.723. By having 12.18 million parameters, DMT gets the best results and outperforms each evaluation matrix (in the table marked in blue bold font) with a dice value of 0.732, IOU 0.781, precision 0.843, sensitivity 0.735, and specificity 0.845 when compared to other segmentation methods.

5. Conclusion

This experiment focuses on countering pseudo surveillance label noise with a re-weighted loss function based on model disagreement. Our research contributes to trying out a new model for developing CNN architecture that considers physician intervention in validating the results. DMT will conduct multi-training on pseudo labels with doctor's labels to be trusted in area segmentation. Furthermore, we have adapted DMT to an iterative framework for better performance in image semantic segmentation. With the lack of labeled data, DMT is very effective, which is not considered in other segmentation methods such as mask RCNN and UNET. DMT is flexible and easy to implement. This experiment demonstrated the effectiveness of the proposed DMT. The system settings used, such as the addition of the consistency setting method, had a tremendous impact on the final result of 91.32%, which shows the latest segmentation results.

Conflicts of interest

All authors has no conflicts of interest

Author contributions

Ersa resita contributed to the experiment by processing experimental data, compiling manuscripts, and writing papers with input from all authors. Tita Karlita contributed to compiling the research and was responsible for overall direction and planning, and led the writing of the manuscript. Riyanto sigit contributed in designing the research, ideas, main concept, and outline of the evidence, contributed to the interpretation of the results and understanding of the ideas presented. Eko Mulyanto Yuniarno contributed to designing and directing the project to derive models and analyze data and perform analytical calculations. I Ketut Eddy Purnama contributed in designing the study and verifying analytical methods. Mauridhi Hery Purnomo contributed in proposing the experiment involved in planning and monitoring. All authors provide critical feedback and help shape research, analysis and manuscripts. All authors discuss the results and comment script.

Acknowledgments

We would like to thank the Ministry of Research, Technology, and Higher Education of the Republic of Indonesia Government through the Penelitian Dasar Unggulan Perguruan Tinggi program. Thanks to Penelitian Tesis Magister, Direktorat Jenderal Pendidikan Vokasi Kementerian Pendidikan, Kebudayaan, Riset dan Teknologi. We would like to extend our sincere thanks to the researchers at the Health Engineering Laboratory and the Knowledge Engineering Laboratory, Electronics Engineering Polytechnic Institute of Surabaya.

Reference

- [1] T. Ai, Z. Yang, H. Hou, C. Zhan, C. Chen, W. Lv, Q. Tao, Z. Sun, L. Xia, "Correlation of Chest CT and RT-PCR Testing for Coronavirus Disease 2019 (COVID-19) in China: A Report of 1014 Cases", *Radiology*, Vol. 296, 2020.
- [2] Z. Ye, Y. Zhang, Y. Wang, Z. Huang, B. Song, "Chest CT manifestations of new coronavirus disease 2019 (COVID-19) a pictorial review", *European Society of Radiology*, 2020.

- [3] V. Rajinikanth, N. Dey, A. N. J. Raj, A. E. Hassanien, K. C. Santosh, and N. S. M. Raja, "Harmony-Search and Otsu based System for Coronavirus Disease (COVID-19) Detection using Lung CT Scan Images", *Applied Sciences*, 2020.
- [4] F. Shi, J. Wang, J. Shi, Z. Wu, Q. Wang, Z. Tang, Y. Shi, and D. Shen, "Review of Artificial Intelligence Techniques in making Data Acquisition, Segmentation and Diagnosis for COVID-19", *IEEE Engineering in Medicine & Biology Society*, 2021.
- [5] M. Roxer, "Coronavirus Pandemic (COVID-19) – The Data", *The Global Change Data Lab, A Non-Profit Organization Based in The United Kingdom*, 2019. [Online]. Available: <https://ourworldindata.org/coronavirus-data>. [Accessed September 20 2020].
- [6] S. Walvekar and S. Shinder, "Efficient Medical Image Segmentation Of COVID19 Chest CT Images Based on Deep Learning Techniques", In: *Proc. of International Conf. on Emerging Smart Computing and Informatics (ESCI)*, Pune, India, 2021.
- [7] Y. Yang, J. Chen, R. Wang, T. Ma, L. Wang, J. Chen, W. Zheng, and T. Zhang, "Toward Unbiased COVID-19 Lesion Localisation And Segmentation Via Weakly Supervised Learning", In: *Proc. of IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, Nice, France, 2021.
- [8] N. Hasanzadeh, S. S. Paima, A. Bashirgonbadi, M. Naghibi, and H. S. Zadeh, "Segmentation of COVID-19 Infections on CT: Comparison of Four UNet-Based Networks", In: *Proc. of International Iranian Conference on Biomedical Engineering (ICBME)*, Tehran, Iran, 2020.
- [9] V. Vasilescu, A. Neacsu, E. Chouzenoux, J. C. Pesquet, and C. Burileanu, "A Deep Learning Approach For Improved Segmentation Of Lesion Related To COVID-19 Chest CT Scans", In: *Proc. of IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, Nice, France, 2021.
- [10] D. P. Fan, T. Zhou, G. P. Ji, Y. Zhou, G. Chen, H. Fu, J. Shen, and L. Shao, "Inf-Net: Automatic COVID-19 Lung Infection Segmentation from CT Images", *IEEE Engineering in Medicine & Biology Society*, 2020.
- [11] I. Laradji, P. Rodriguez, O. Manas, K. Lensink, M. Law, L. Kurzman, W. Parker, D. Vazquez, and D. Nowrouzezahrai, "A Weakly Supervised Consistency-based Learning Method for COVID-19 Segmentation in CT Images", In: *Proc. of IEEE Winter Conf. on Applications of Computer Vision (WACV)*, 2021.
- [12] T. Mahmud, M. A. Rahman, S. A. Fattah, and S. Y. Kung, "CovSegNet: A Multi Encoder-Decoder Architecture for Improved Lesion Segmentation of COVID-19 Chest CT Scans", In: *Proc. of IEEE Winter Conf. on Applications of Computer Vision*, 2021.
- [13] N. Paluru, A. Dayal, H. B. Jensen, T. Sakinis, L. R. Cenkeramaddi, J. Prakash, and P. K. Yalavarthy, "Anam-Net: Anamorphic Depth Embedding-Based Lightweight CNN for Segmentation of Anomalies in COVID-19 Chest CT Images", *IEEE Transaction on Neural Networks and Learning Systems*, Vol. 32, No. 3, 2021.
- [14] Y. Jiang, H. Chen, M. Loew, and H. Ko, "COVID-19 CT Image Synthesis With a Conditional Generative Adversarial Network", *IEEE Journal of Biomedical and Health Informatics*, Vol. 25, No. 2, 2021.
- [15] Y. H. Wu, S. H. Gao, J. Mei, J. Xu, D. P. Fan, R. G. Zhang, and M. M. Cheng, "JCS: An Explainable COVID-19 Diagnosis System by Joint Classification and Segmentation", *IEEE Transaction on Image Processing*, Vol. 30, 2021.
- [16] D. Hendrycks, N. Mu, E. D. Cubuk, B. Zoph, J. Gilmer, B. Lakshminarayanan, "AUGMIX: A simple data processing method to improve robustness and uncertainty", In: *Proc. of International Conf. of Learning Representation*, 2020.
- [17] E. Englesson and H. Azizpour, "Consistency Regularization Can Improve Robustness to Label Noise", In: *Proc. of International Conf. of Machine Learning*, 2021.
- [18] H. Zhang, Z. Zhang, A. Odena, and H. Lee, "Consistency Regularization For Generative Adversarial Networks", In: *Proc. of International Conf. Paper at ICLR*, 2020.
- [19] Z. Feng, Q. Zhou, Q. Gu, X. Tan, G. Cheng, X. Lu, J. Shi, and L. Ma, "DMT: Dynamic Mutual Training for Semi-Supervised Learning", *Pattern Recognition*, Vol. 130, 2021.
- [20] S. Qiao, W. Shen, Z. Zhang, B. Wang, and A. Yuille, "Deep Co-Training For Semi-Supervised Image Recognition", In: *Proc. of European Conference on Computer Vision*, 2021.
- [21] J. Peng, G. Estrada, M. Pedersoli, and C. Desrosiers, "Deep Co-Training For The Semi-Supervised Image Segmentation", *Pattern Recognition*, Vol. 107, 2019.
- [22] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks For Biomedical Image Segmentation", In: *Proc. of International Conference on Medical Image Computing and*

Computer-Assisted Intervention, 2015.

- [23] G. Wang, X. Liu, C. Li, Z. Xu, J. Ruan, H. Zhu, T. Meng, K. Li, N. Huang, and S. Zhang, "A Noise-robust Framework for Automatic Segmentation of COVID-19 Pneumonia Lesions from CT Images", *IEEE Transaction on Medical Imaging*, Vol. 39, 2020.
- [24] J. Bhattacharyya, "Pseudo Labelling – A Guide To Semi-Supervised Learning", *Intimate Events, Capturing the Essence of AI Industry*, 2020.
- [25] Radiologists, "COVID-19 CT segmentation dataset", *Artificial Intelligence AS*, 2020. [Online]. Available: <http://htmlsegmentation.s3.eu-north-1.amazonaws.com/index.html>. [Accessed 10 September 2019].
- [26] Radiologist, 2020. [Online]. Available: <http://medicalsegmentation.com/covid19/>. [Accessed 4 December 2019].
- [27] A. Kundu, "Performing Instance Segmentation on X-Ray Images with Mask R-CNN", 2020. (<https://avikkundu.medium.com/performing-instance-segmentation-on-x-ray-images-with-mask-r-cnn-761dbca23511>)
N. Hasanzadeh, S. S. Paima, M. Naghibi, H. S. Zadeh, A. Bashirgonbadi, "Segmentation of COVID-19 Infections on CT: Comparison of Four UNet-Based Networks", In: *Proc. of International Iranian Conference on Biomedical Engineering (ICBME 2020)*, Tehran, Iran, 2020.