2021 IEEE Winter Conference on Applications of Computer Vision (WACV) | 978-1-6654-0477-8/20/$31.00 ©2021 IEEE | DOI: 10.1109/WACV48630.2021.00250

A Weakly Supervised Consistency-based Learning Method for COVID-19 Segmentation in CT Images

Issam Laradji[1,2,6], Pau Rodriguez[2], Oscar Mañas[2,7], Keegan Lensink[3,5], Marco Law[4,5], Lironne Kurzman[5], William Parker[4,5], David Vazquez[2], and Derek Nowrouzezahrai[6]

[1]issam.laradji@gmail.com, [2]Element AI, [3]Xtract AI, [4]SapienML, [5]University of British Columbia, [6]McGill University, [7]Universitat Politècnica de Catalunya

Abstract

Coronavirus Disease 2019 (COVID-19) has spread aggressively across the world causing an existential health crisis. Thus, having a system that automatically detects COVID-19 in tomography (CT) images can assist in quantifying the severity of the illness. Unfortunately, labelling chest CT scans requires significant domain expertise, time, and effort. We address these labelling challenges by only requiring point annotations, a single pixel for each infected region on a CT image. This labeling scheme allows annotators to label a pixel in a likely infected region, only taking 1-3 seconds, as opposed to 10-15 seconds to segment a region. Conventionally, segmentation models train on point-level annotations using the cross-entropy loss function on these labels. However, these models often suffer from low precision. Thus, we propose a consistency-based (CB) loss function that encourages the output predictions to be consistent with spatial transformations of the input images. The experiments on 3 open-source COVID-19 datasets show that this loss function yields significant improvement over conventional point-level loss functions and almost matches the performance of models trained with full supervision with much less human effort. Code is available at: `https://github.com/IssamLaradji/covid19_weak_supervision`.

1. Introduction

The severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) has quickly become a global pandemic and resulted in over 400,469 COVID-19 related deaths as of June 8th, 2020[1]. The virus comes from the same family as the SARS-CoV outbreak originated in 2003 and the MERS-CoV outbreak of 2012, and is projected to join other coronavirus strains as a seasonal disease. The disease can present itself in a variety of ways ranging from asymptomatic to acute respiratory distress syndrome (ARDS). However, the primary and most common presentation associated with morbidity and mortality is the presence of opacities and consolidation in a patient's lungs. As the disease spreads, healthcare centers around the world are becoming overwhelmed and facing shortages of the essential equipment necessary to manage the symptoms of the disease. Severe cases require admission to the intensive care unit (ICU) and need mechanical ventilation, some sources [10] citing at a rate of 5% of all infected. Thus, availability of ICU beds due to the overwhelming number of COVID-19 cases around the world is a large challenge. Rapid screening is necessary to diagnose the disease and slow the spread, making effective tools essential for prognostication in order to efficiently allocate intensive care services to those who need it most.

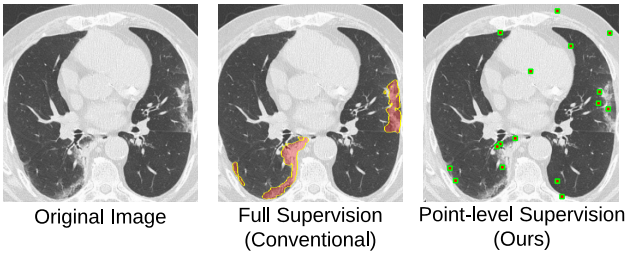

Figure 1: **Labeling Scheme.** We illustrate the difference between labels obtained using full supervision and point-level supervision. One point is placed on each infected region, and several on the background region.

Upon inhalation, the virus attacks and inhibits the alveoli of the lung, which are responsible for oxygen exchange [43]. In response, and as part of the inflammatory repair process, the alveoli fill with fluid, causing var-

[1]Source: World Health Organization.

978-1-6654-0477-8/21/$31.00 ©2021 IEEE
DOI 10.1109/WACV48630.2021.00250

ious forms of opacification within the lung when viewed on Computed Tomography (CT) scans. Due to the increased density, these areas present on CT scans as increased attenuation with preserved bronchial and vascular markings known as ground glass opacities (GGO). In addition, the accumulation of fluid progresses to obscure bronchial and vascular regions on CT scans is known as consolidation.

While reverse transcription polymerase chain reaction (RT-PCR) has been considered the gold standard for COVID-19 screening, the shortage of equipment and strict requirements for testing environments limit the utility of this test in all settings. Further, RT-PCR is also reported to suffer from high false negative rates due to its relatively low sensitivity yet high specificity [1]. CT scans are an important complement to RT-PCR tests which were shown to demonstrate effective diagnosis, including follow-up assessment and the evaluation of disease evolution [1, 52].

In addition, expert interpretation of CT scans can provide insight on the severity of the infection by identifying various patterns of opacification. The prevalence of these patterns, which are correlated with the severity of the infection, has been correlated to different stages of the disease [21, 42]. The quantification of the opacification composition enables efficient estimation of the stage of the disease and the patient outcome.

Deep learning-based methods have been widely applied in medical image analysis to combat COVID-19 [12, 15, 41]. They have been proposed to detect patients infected with COVID-19 via radiological imaging. For example, COVID-Net [39] was proposed to detect COVID-19 cases from chest radiography images. An anomaly detection model was designed to assist radiologists in analyzing the vast amounts of chest X-ray images [34]. For CT imaging, a location-attention oriented model was employed to calculate the infection probability of COVID-19 [5]. A weakly-supervised deep learning-based software system was developed in [48] using 3D CT volumes to detect COVID-19. A list of papers for COVID-19 imaging-based AI works can be found in Wang et al. [40]. Although plenty of AI systems have been proposed to provide assistance in diagnosing COVID-19 in clinical practice, there are only a few related works [9], and no significant impact has been shown using AI to improve clinical outcomes, as of yet.

According to Ma et al. [26], it takes around 400 minutes to delineate one CT scan with 250 slices. That is an average of 1.6 minutes per slice. On the other hand, it takes around 3 seconds to point to a single region at the pixel level Papadopoulos et al. [30]. Thus, point-level annotations allow us to label many more slices quickly.

Point-level annotations are not as expressive as segmentation labels, making effective learning a challenge for segmentation models (Fig. 1). Conventionally, segmentation models train on point-level annotations using the cross-entropy on these labels. While this loss can yield good results in some real-life datasets [3], the resulting models usually suffer from low precision as they often predict big blobs. Such predictions are not suitable for imbalanced images where only few small regions are labeled as foreground. Thus, we propose a consistency-based (CB) loss function that encourages the model's output predictions to be consistent with spatial transformations of the input images. While consistency methods have been successfully deployed in semantic segmentation, the novel aspect of this work is the notion of consistency under weak supervision, which utilizes unlabeled pixels during training. We show that this regularization method yields significant improvement over conventional point-level loss functions was on 3 open-source COVID-19 datasets. We also show that this loss function results in a segmentation performance that almost matches that of the fully supervised model. To the best of our knowledge, this is the first time that self-supervision has been applied in conjunction with point-level supervision on a medical segmentation dataset.

We summarize our contributions and results on 3 publicly available CT Scans [2] as follows:

1. We propose a framework that trains using a consistency-based loss function on a medical segmentation dataset labeled with point-level supervision.
2. We present a trivial, yet cost-efficient point-level supervision setup where the annotator is only required to label a single point on each infected region and several points on the background.
3. We show that our consistency-based loss function yields significant improvement over conventional point-level loss functions and almost matches the performance of models trained with full supervision.

2. Related Work

In this section, we review semantic segmentation methods applied to CT scans for general medical problems, and for COVID-19. We also review methods for weakly supervised problem setups and self-supervision methods that were shown to help generalization performance. For all of our methods, we use an ImageNet-pretrained VGG16 FCN8 [24] backbone as our segmentation method.

Semantic segmentation for CT Scans has been widely used for diagnosing lung diseases. Diagnosis is often based on segmenting different organs and lesions from chest CT slices, which can provide essential information for doctors to identify lung diseases. Many methods exist that perform nodule segmentation of lungs. Early algorithms are based on image processing and SVMs to segment nodules [15].

[2] Found here: https://medicalsegmentation.com/covid19/

Then, algorithms based on deep learning emerged [12]. These methods include central focus CNNs [41] and GAN-synthesized data for nodule segmentation in CT scans [14]. A recent method uses multiple deep networks to segment lung tumors from CT slices with varying resolutions, and multi-task learning of joint classification and segmentation [13].

Semantic segmentation for COVID-19 While COVID-19 is a recent phenomenon, several methods have been proposed to analyze infected regions of COVID-19 in lungs. Fan et al. [9] proposed a semi-supervised learning algorithm for automatic COVID-19 lung infection segmentation from CT scans. Their algorithm leverages attention to enhance representations. Similarly, Zhou et al. [49] proposed to use spatial and channel attention to enhance representations, and Chen et al. [6] augment U-Net [32] with ResNeXt [47] blocks and attention. Instead of focusing on the architecture, Amyar et al. [2] proposed to improve the segmentation performance with a multi-task learning approach which includes a reconstruction loss. Although previous methods are accurate, their computational cost can be prohibitive. Thus, Qiu et al. [31] proposed Miniseg for efficient COVID-19 segmentation. Unfortunately, these methods require full supervision, which is costly to acquire compared to point-level supervision: our problem setup.

Weakly supervised semantic segmentation methods can vastly reduce the required annotation cost for collecting a training set. According to Bearman et al. [3], manually collecting image-level and point-level labels for the PASCAL VOC dataset [8] takes only 20.0 and 22.1 seconds per image, respectively. These annotation methods are an order of magnitude faster than acquiring full segmentation labels, which is 239.0 seconds on average. Other forms of weaker labels were explored as well, including bounding boxes [16] and image-level annotation [50]. Weak supervision was also explored in instance segmentation where the goal is to identify object instances as well as their class labels [19, 20, 51]. In this work, the labels are given as point-level annotations instead of the conventional per-pixel level labels and the task is to identify the class labels of the regions only.

Self-supervision for weakly supervised semantic segmentation is a relatively new research area that has strong potential in improving segmentation performance. The basic idea is to generate two perturbed versions of the input and apply consistency training to encourage the predictions to be similar [46]. For example, FixMatch [36] combined consistency regularization with pseudo-labeling to produce artificial image-level labels. In the case of dense predictions, the outputs need to be further transformed in order to compare them against a consistency loss, making the model's output equivariant against transformations. Self-supervision was recently applied in a weakly supervised setup where annotations are image-level [44]. The idea was to make the output consistent across scales, which led to new state-of-the-art results on PASCAL VOC dataset. Ouali et al. [29] proposed to apply cross-consistency training, where the perturbations are applied to the outputs of the encoder and the dense predictions are enforced to be invariant. These perturbations can also be used for data augmentation, which can be learnt automatically using methods based on reinforcement learning and bilevel optimization [7, 28]. For medical segmentation, self-supervision has been used along with semi-supervised learning [4, 22]. Bortsova et al. [4] made the outputs consistent across elastic transforms, while Li et al. [22] added a teacher-student paradigm for consistency training. In this work, we apply consistency loss on the novel setup of medical segmentation with point supervision.

3. Methodology

Problem Setup and Network Architecture. We define the problem setup as follows. Let X be a set of N training images with corresponding ground truth labels Y. Y_i is a $W \times H$ matrix with non-zero entries that indicate the locations of the object instances. The values of these entries indicate the class label that the point corresponds to.

We use a standard fully-convolutional neural network that takes as input an image of size $W \times H$ and outputs a $W \times H \times C$ per-pixel map where C is the set of object classes of interest. The output map is converted to a per-pixel probability matrix S_i by applying the softmax function across classes. These probabilities indicate how likely each pixel belongs to the infected region of a class $c \in C$.

Proposed Loss Function. Our weakly supervised method uses a loss function that consists of a supervised point-level loss and an unsupervised consistency loss. Given a network f_θ that outputs a probability map S_i given an image X_i, we optimize its parameters θ using the following loss function,

$$\mathcal{L}(X, Y) = \sum_{i=1}^{N} \underbrace{\mathcal{L}_P(X_i, Y_i)}_{\text{Point-level}} + \lambda \underbrace{\mathcal{L}_C(X_i)}_{\text{Consistency}} , \qquad (1)$$

where λ is used to weigh between the two loss terms.

Point-level loss. We apply the standard cross-entropy function against point annotations, which is defined as follows,

$$\mathcal{L}_P(X_i, Y_i) = -\sum_{j \in \mathcal{I}_i} \log(f_\theta(X_i)_{jY_j}) , \qquad (2)$$

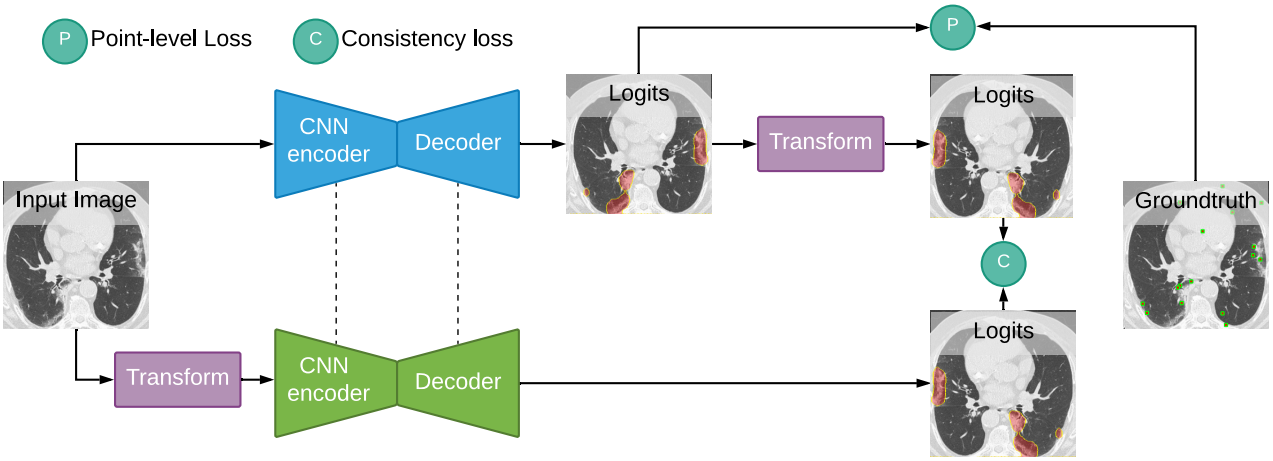

Figure 2: **Model Training.** Our model has two branches with shared weights. The first branch encodes the original input x while the second branch encodes the transformed input $t(x)$. The point-level loss compares the outputs $f(x)$ and $f(t(x))$ with the corresponding weak labels y and $t(y)$. In addition, an unsupervised consistency loss is used to make the outputs $t(f(x))$ and $f(t(x))$ consistent.

where $f_\theta(X_i)_{jY_j}$ is the output corresponding to class Y_j for pixel j, and $\mathcal{I}_i$ is the set of labeled pixels for image X_i.

Consistency loss. We first define a set of geometric transformations $T = \{t_1, t_2, ..., t_n\}$. An example of t_k is horizontal flipping, which can be used to transform an image X_i and its corresponding label Y_i collectively to their flipped version. The goal of this loss function is to make the model's output consistent with respect to these transformations on the input image. The loss function is defined as follows,

$$\mathcal{L}_C(X_i) = \sum_{j \in \mathcal{P}_i} |t_k(f_\theta(X_i))_j - f_\theta(t_k(X_i))_j|, \quad (3)$$

where $\mathcal{P}_i$ is the set of pixels for image X_i. This unsupervised loss function helps the network learn equivariant semantic representations that go beyond the translation equivariance that underlies convolutional neural networks, serving as an additional form of supervision.

Model Training. The overview of the model training is shown in Fig. 2 and Alg. 1. The model has two branches with shared weights θ. At each training step k, we sample an image X_i and a transform function $t_k \in T$. The model's first branch takes as input the original image X_i and the second branch takes as input the transformed image $t_k(X_i)$. The transformed output of the first branch, $y_1 := t_k(f_\theta(X_i))$, is aligned with the prediction of the second branch $y_2 := f_\theta(t_k(X_i))$ for pixel-wise comparison by the consistency loss function 3.

In addition to the consistency loss, the point-level loss $\mathcal{L}_P$ is applied to both input X_i and $t_k(X_i)$, *i.e.* $\mathcal{L}_P(t_k(X_i), t_k(Y_i))$, where $t_k(Y_i)$ is a pseudo ground-truth mask for $t_k(X_i)$ generated by applying the same geometric transformation t_k to the ground-truth mask Y_i. In this case, the network is forced to update the prediction for $t_k(X_i)$ to be more similar to $t_k(Y_i)$.

In this work, we use geometric transformations which allow us to infer the true label of images that undergo these transformations. For instance, the segmentation mask of the flipped version of an image is the flipped version of the original segmentation mask. Thus, we include the following transformations: 0, 90, 180 and 270 degree rotation and a horizontal flip. At test time, the trained model can then be directly used to segment infected regions on unseen images with no additional human input.

4. Experiments

4.1. Experimental Setup

Here we describe the details behind the datasets, methods, and evaluation metrics used in our experiments.

4.1.1 Datasets

We evaluate our weakly supervised learning system on three separate open source medical segmentation datasets (referred to as COVID-19-A/B/C). For each dataset, a point-level label is obtained for a segmentation mask by taking the pixel with the largest distance transform as the centroid.

Thus, we generate a single supervised point for each disjoint infected region on the training images. For the background region, we randomly sample several pixels as the ground-truth points (Figure 1). We show the dataset statistics in Table 1 and describe them in the next sections.

Algorithm 1: Model Training

Input : $X = \{X_1, X_2, ..., X_n\}$ images, $Y = \{Y_1, Y_2, ..., Y_n\}$ point-level masks.
Output : Trained parameters θ^*
Parameters: A weight coefficient λ, A set of transformation functions T, A model forward function f_θ.

for *each batch* B **do**
  $\mathcal{L} \leftarrow 0$
  for *each* $(X_i, Y_i) \in B$ **do**
    Compute Point Loss
    $\mathcal{L}_P \leftarrow -\sum_{j \in \mathcal{I}_i} \log(f_\theta(X_i)_{jY_j})$
    Uniformly sample a transform function
    $t_k \sim T$
    Compute Consistency Loss
    $\mathcal{L}_C \leftarrow \sum_{j \in \mathcal{P}_i} |t_k(f_\theta(X_i))_j - f_\theta(t_k(X_i))_j|$
    $\mathcal{L} \leftarrow \mathcal{L} + \mathcal{L}_P + \lambda \mathcal{L}_C$
  end
  Update θ by backpropagating w.r.t. $\mathcal{L}$
end

COVID-19-A [9, 27] consists of 100 axial lung CT JPEG images obtained from 60 COVID-19 lung CTs provided by the Italian Society of Medical and Interventional Radiology. Each image was labeled for ground-glass, consolidation, and pleural effusion by a radiologist. We discarded two images without areas of infection from this dataset due to their low resolution. Images were resized to a fixed dimension of 352×352 pixels and normalized using ImageNet statistics [33]. The final dataset consisted of 98 images separated into a training set ($n = 50$), validation set ($n = 5$), and a test set ($n = 48$).

COVID-19-B [27] consists of 9 volumetric COVID-19 chest CTs in DICOM format containing a total of 829 axial slices. Images were first converted from Houndsfield units to unsigned 8-bit integers, then resized to 352×352 pixels and normalized using ImageNet statistics [33].

We use COVID-19-B to evaluate the consistency loss on two splits of the dataset: separate and mixed. In the separate split (COVID-19-B-Separate), the slices in the training, validation, and test set come from different scans. The goal is to have a trained model that can generalize to scans of new patients. In this setup, the first 5 scans are defined as the training set, the sixth scan as validation, and the remaining scans as the test set.

For the mixed split (COVID-19-B-Mixed), the slices in the training, validation, and test set come from the same scans. The idea is to have a trained model that can infer the masks in the remaining slices of a scan when the annotator only labels few of the slices in that scan. In this setup, the first 5 scans are defined as the training set, the sixth scan as validation, and the remaining scans as the test set. For each scan, the first 45% slices of the scan are defined as the training set, the next 5% as the validation set, and the remaining slices as the test set.

COVID-19-C [25] consists of 20 CT volumes. Lungs and areas of infection were labeled by two radiologists and verified by an experienced radiologist. Each three-dimensional CT volume was converted from Houndsfield units to unsigned 8-bit integers and normalized using ImageNet statistics [33].

As with COVID-19-B, we also split the dataset into *separate* and *mixed* versions to evaluate our model's efficacy. For the separate split (COVID-19-B-Sep), we assign 15 scans to the training set, 1 scan to the validation set, and 4 scans to the test set. For the mixed split (COVID-19-C-Mixed), we separate the slices from each scan in the same manner as in COVID-19-B, training on the first 45% axial slices, validating on the next 5% of slices, and testing on the remaining 50% of slices.

4.1.2 Evaluation Metrics

As common practice [35], we evaluate our models against the following metrics for semantic segmentation:

Intersection over Union (IoU) measures the overlap between the prediction and the ground truth: $IoU = \frac{TP}{TP+FP+FN}$, where TP, FP, and FN is the number of true positive, false positive and false negative pixels across all images in the test set.

Dice Coefficient (F1 Score) is similar to IoU but gives more weight to the intersection between the prediction and the ground truth: $F1 = \frac{2*TP}{2*TP+FP+FN}$.

PPV (Positive Predicted Value) measures the fraction of positive samples that were correctly predicted, which is also known as precision: $PPV = \frac{TP}{TP+FP}$.

Sensitivity (recall) measures the fraction of real positive samples that were predicted correctly: $Sens. = \frac{TP}{TP+FN}$.

Specificity (true negative rate) measures the fraction of real negative samples that were predicted correctly: $Spec. = \frac{TN}{FP+TN}$.

4.2. Methods and baselines

We provide experiments with three weakly supervised loss functions based on point-level annotations and a fully-supervised upper bound method:

Table 1: Statistics of open-source COVID-19 datasets.

Name	# Cases	# Slices	# Slices with Infections (%)	# Infected Regions
COVID-19-A	60	98	98 (100.0%)	776
COVID-19-B	9	829	372 (44.9%)	1488
COVID-19-C	20	3520	1841 (52.3%)	5608

- *Point loss (PL)*. It is defined in Eq. 2 in Bearman et al. [3]. The loss function encourages all pixel predictions to be background for background images and applies cross-entropy against the provided point-level annotations, ignoring the rest of the pixels.
- *CB(Flip) + PL*. It is defined in Eq. 1 in Section 3, which combines the point loss and the horizontal flip transformation for the consistency loss.
- *CB(Flip, Rot) + PL*. It is the same as *CB(Flip) + PL* except that the transformation used for the consistency loss also includes the 0, 90, 180, and 270 degree rotation transformation uniformly sampled for each image.
- *Fully supervised*. This loss function combines weighted cross-entropy and IoU loss as defined in Eq. (3) and (5) from Wei et al. [45], respectively. It is an efficient method for ground truth segmentation masks that are imbalanced. Since this loss function requires full supervision, it serves as an upper bound performance in our experimental results.

Implementation Details All methods use an Imagenet-pretrained VGG16 FCN8 network [24]. Models are trained with a batch size of 8 for 100 epochs with ADAM [17] and a learning rate of 10^{-4}. We also achieved similar results with optimizers that do not require a learning rate [23, 37, 38]. The reported scores are on the test set which were obtained with early stopping on the validation set. Point annotations were obtained by uniformly sampling one pixel from each annotated mask. The same amount of points are uniformly sampled from the background.

4.3. Segmentation Results

Here we evaluate the loss functions on three covid datasets and discuss their results.

4.3.1 COVID-19-A

Table 2 shows that with only point supervision, our method was able to perform competitively compared to full supervision. In terms of sensitivity, it can be observed that the point loss outperformed the fully-supervised baseline by 0.11 points. For the other metrics, we were able to obtain competitive performance when using the consistency-based (CB) loss. The gap between fully supervised and point-based loss is reduced when using flips and rotations (Flip, Rot) instead of simple horizontal flips (Flip). Moreover, with (Flip, Rot), our method surpasses the fully-supervised sensitivity by 0.12 points. Figure 3 contains qualitative results comparing the ground truth with the point-level loss and the effect of the CB loss. It can be observed how using the CB loss produces masks that are more contained in the region of interest (the lungs). COVID-19-A is a small and easy dataset compared to COVID-19-B and COVID-19-C. Thus, in the next sections, we show that with bigger datasets, *CB point loss* obtains even better performance on the rest of the metrics with weak supervision.

Table 2: COVID-19-A Segmentation Results

Loss Function	Dice	IoU	PPV	Sens.	Spec.
Fully Supervised	0.65	0.48	0.52	0.85	0.85
Point Loss (PL)	0.54	0.37	0.39	**0.88**	0.73
CB(Flip) + PL (Ours)	0.58	0.41	0.46	0.80	0.82
CB(Flip, Rot) + PL (Ours)	**0.73**	**0.57**	**0.65**	0.82	**0.92**

4.3.2 COVID-19-B

As seen in Table 3 and 4, the CB method is more robust against different splits of the data. In both COVID-19-B-Sep and COVID-19-B-Mixed, the CB method achieves similar results, whereas there is more variance in the results with *Point Loss* and *W-CE* metrics. While the W-CE baseline has an average gap of 0.37 between *sep* and *mixed* over all metrics, the CB Point loss only has a difference of 0.07 with (Flip) and 0.08 with (Flip, Rot). Remarkably, on *sep*, our weakly supervised method with (Rot, Flip) improved by 0.48, 0.42, and 0.56, on the Dice, IoU, and Sensitivity metrics, with respect to the W-CE baseline. On PPV and Specificity, our method was able to retain a competitive performance, with a difference of 0.16 and 0.02 respectively. Except the for Sensitivity in COVID-19-B-Sep, the CB loss (Rot, Flip) yields better results than the point loss.

4.3.3 COVID-19-C

As seen in Tables 5 and 6, the fully supervised method performs better on COVID-19-C than in the other two datasets and the performance gap between mixed and sep is smaller.

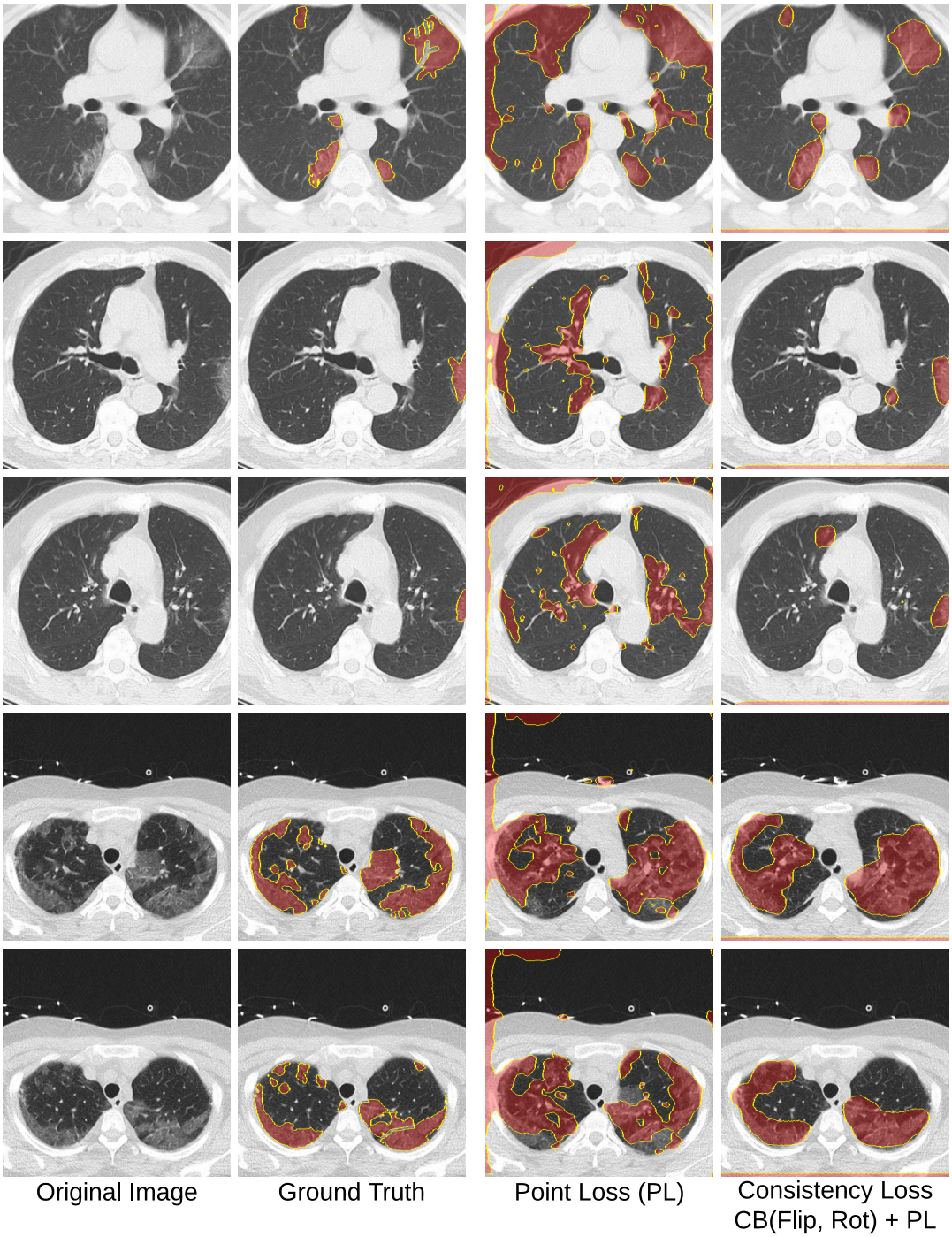

Figure 3: **Qualitative results.** We show the predictions obtained from training the model with the point-level loss in Bearman et al. [3] and our consistency-based (CB) loss. With the CB loss the predictions are much closer to the ground-truth labels.

This can be attributed to the larger size of COVID-19-C. The average gap in performance of the fully supervised baseline between the *mixed* and *sep* versions is 0.06 for COVID-19-C. The weakly supervised CB loss yields a gap of 0.05 in performance between *mixed* and *sep*. Similar to COVID-19-B, except for Sensitivity, the CB point loss

Table 3: COVID-19-B-Mixed Segmentation Results

Loss Function	Dice	IoU	PPV	Sens.	Spec.
Fully Supervised	0.84	0.73	0.90	0.80	1.00
Point Loss (PL)	0.33	0.20	0.20	0.91	0.94
CB(Flip) + PL (Ours)	0.73	0.57	**0.64**	0.85	**0.99**
CB(Flip, Rot) + PL (Ours)	**0.75**	**0.60**	0.63	**0.92**	**0.99**

Table 4: COVID-19-B-Sep Segmentation Results

Loss Function	Dice	IoU	PPV	Sens.	Spec.
Fully Supervised	0.24	0.14	0.89	0.14	1.00
Point Loss (PL)	0.57	0.40	0.44	**0.82**	0.94
CB(Flip) + PL (Ours)	0.69	0.53	0.72	0.66	**0.99**
CB(Flip, Rot) + PL (Ours)	**0.72**	**0.56**	**0.73**	0.70	0.98

Table 5: COVID-19-C-Mixed Segmentation Results

Loss Function	Dice	IoU	PPV	Sens.	Spec.
Fully Supervised	0.78	0.64	0.79	0.77	1.00
Point Loss (PL)	0.12	0.07	0.07	**0.95**	0.82
CB(Flip) + PL (Ours)	0.66	0.49	**0.56**	0.80	**0.99**
CB(Flip, Rot) + PL (Ours)	**0.68**	**0.51**	**0.56**	0.85	**0.99**

Table 6: COVID-19-C-Sep Segmentation Results

Loss Function	Dice	IoU	PPV	Sens.	Spec.
Fully Supervised	0.71	0.55	0.78	0.65	0.99
Point Loss (PL)	0.37	0.23	0.23	**0.97**	0.76
CB(Flip) + PL (Ours)	0.69	0.53	0.62	0.79	0.96
CB(Flip, Rot) + PL (Ours)	**0.75**	**0.59**	**0.66**	0.86	**0.97**

yields substantially better results than the point loss. We also observed better results when adding rotations. In fact, with (Flip, Rot), our weakly supervised method improves over the fully supervised baseline by 0.04, 0.04, and 0.21 on Dice, IoU and *Sensitivity* on the *sep* split.

4.4. Counting and Localization Results

In this setup we consider the task of counting and localizing COVID-19 infected regions in CT Scan images. Radiologists strive to identify all regions that might have relevance to COVID-19, which is a very challenging task, especially for small infected regions. Thus, having a model that can localize these regions can help improve radiologist performance in the identification of infected regions.

We consider the COVID-19-B and COVID-19-C datasets to evaluate 3 types of loss functions: point loss (Eq.2 from Bearman et al. [3]), LCFCN loss (Eq. 1 from Laradji et al. [18]), and consistency-based LCFCN loss that we propose in this section.

The consistency based LCFCN (CB LCFN Loss) loss extends the LCFCN loss with the CB loss proposed in Eq. 1 using the horizontal flip transformation. To evaluate these 3 loss functions, we consider each connected infected region as a unique region. The goal is to identify whether these regions can be counted and localized. We use the mean absolute error (MAE) and grid average mean absolute error (GAME) [11] to measure how well the methods can count and localize infected regions. We provide results for $GAME(L = 4)$ which divides the image using a grid of 4^L non-overlapping regions, and the error is computed as the sum of the MAE in each of these subregions.

Table 7: COVID-19-B-Mixed Counting and Localization

Loss Function	MAE	GAME
Point Loss	5.97	7.24
LCFCN Loss	1.15	2.09
CB LCFCN (Ours) Loss	**0.66**	**1.74**

Table 8: COVID-19-C-Mixed Counting and Localization

Loss Function	MAE	GAME
Point Loss	9.63	11.76
LCFCN Loss	1.01	1.70
CB LCFCN Loss (Ours)	**0.82**	**1.42**

Table 7 and 8 shows that the consistency loss helps LCFCN achieve superior results in counting and localizing infected regions in the CT image. It is expected that the *Point Loss* achieves poor performance as it predicts big blobs that can encapsulate several regions together. On the other hand, the consistency loss helped LCFCN improve its results suggesting the model learns more informative semantic features for the task with such self-supervision.

5. Conclusion

Machine learning has the potential to solve a number challenges associated with COVID-19. One example is the identification of high-risk patients by segmenting infected regions in CT scans. However, conventional annotations methods rely on per-pixel labels which are costly to collect for CT scans. In this work, we have proposed an efficient method that can learn from point-level annotations, which are much cheaper to acquire than per-pixel labels. Our method uses a consistency-based loss that significantly improves the segmentation performance compared to conventional point-level loss on 3 COVID-19 open-source datasets. Further, our method obtained results that almost match the performance of the fully supervised methods and they are more robust against different splits of the data.

References

[1] T. Ai, Z. Yang, H. Hou, C. Zhan, C. Chen, W. Lv, Q. Tao, Z. Sun, and L. Xia. Correlation of chest ct and rt-pcr testing in coronavirus disease 2019 (covid-19) in china: a report of 1014 cases. *Radiology*, page 200642, 2020.

[2] A. Amyar, R. Modzelewski, and S. Ruan. Multi-task deep learning based ct imaging analysis for covid-19: Classification and segmentation. *medRxiv*, 2020.

[3] A. Bearman, O. Russakovsky, V. Ferrari, and L. Fei-Fei. What's the point: Semantic segmentation with point supervision. *ECCV*, 2016.

[4] G. Bortsova, F. Dubost, L. Hogeweg, I. Katramados, and M. de Bruijne. Semi-supervised medical image segmentation via learning consistency under transformations. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 810–818. Springer, 2019.

[5] C. Butt, J. Gill, D. Chun, and B. A. Babu. Deep learning system to screen coronavirus disease 2019 pneumonia. *Applied Intelligence*, page 1, 2020.

[6] X. Chen, L. Yao, and Y. Zhang. Residual attention u-net for automated multi-class segmentation of covid-19 chest ct images. *arXiv preprint arXiv:2004.05645*, 2020.

[7] E. D. Cubuk, B. Zoph, D. Mane, V. Vasudevan, and Q. V. Le. Autoaugment: Learning augmentation policies from data. *arXiv preprint arXiv:1805.09501*, 2018.

[8] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 2010.

[9] 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 Transactions on Medical Imaging*, 2020.

[10] W.-j. Guan, Z.-y. Ni, Y. Hu, W.-h. Liang, C.-q. Ou, J.-x. He, L. Liu, H. Shan, C.-l. Lei, D. S. Hui, B. Du, L.-j. Li, G. Zeng, K.-Y. Yuen, R.-c. Chen, C.-l. Tang, T. Wang, P.-y. Chen, J. Xiang, S.-y. Li, J.-l. Wang, Z.-j. Liang, Y.-x. Peng, L. Wei, Y. Liu, Y.-h. Hu, P. Peng, J.-m. Wang, J.-y. Liu, Z. Chen, G. Li, Z.-j. Zheng, S.-q. Qiu, J. Luo, C.-j. Ye, S.-y. Zhu, and N.-s. Zhong. Clinical characteristics of coronavirus disease 2019 in china. *New England Journal of Medicine*, 382(18): 1708–1720, 2020.

[11] R. Guerrero, B. Torre, R. Lopez, S. Maldonado, and D. Onoro. Extremely overlapping vehicle counting. *IbPRIA*, 2015.

[12] M. H. Hesamian, W. Jia, X. He, and P. Kennedy. Deep learning techniques for medical image segmentation: Achievements and challenges. *Journal of digital imaging*, 32(4): 582–596, 2019.

[13] J. Jiang, Y.-C. Hu, C.-J. Liu, D. Halpenny, M. D. Hellmann, J. O. Deasy, G. Mageras, and H. Veeraraghavan. Multiple resolution residually connected feature streams for automatic lung tumor segmentation from ct images. *IEEE transactions on medical imaging*, 38(1):134–144, 2018.

[14] D. Jin, Z. Xu, Y. Tang, A. P. Harrison, and D. J. Mollura. Ct-realistic lung nodule simulation from 3d conditional generative adversarial networks for robust lung segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 732–740. Springer, 2018.

[15] M. Keshani, Z. Azimifar, F. Tajeripour, and R. Boostani. Lung nodule segmentation and recognition using svm classifier and active contour modeling: A complete intelligent system. *Computers in biology and medicine*, 43(4):287–300, 2013.

[16] A. Khoreva, R. Benenson, J. H. Hosang, M. Hein, and B. Schiele. Simple does it: Weakly supervised instance and semantic segmentation. *CVPR*, 2017.

[17] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[18] I. H. Laradji, N. Rostamzadeh, P. O. Pinheiro, D. Vazquez, and M. Schmidt. Where are the blobs: Counting by localization with point supervision. *ECCV*, 2018.

[19] I. H. Laradji, N. Rostamzadeh, P. O. Pinheiro, D. Vazquez, and M. Schmidt. Instance segmentation with point supervision. *arXiv preprint arXiv:1906.06392*, 2019.

[20] I. H. Laradji, D. Vazquez, and M. Schmidt. Where are the masks: Instance segmentation with image-level supervision. In *BMVC*, 2019.

[21] M. Li, P. Lei, B. Zeng, Z. Li, P. Yu, B. Fan, C. Wang, Z. Li, J. Zhou, S. Hu, and H. Liu. Coronavirus disease (COVID-19): Spectrum of CT findings and temporal progression of the disease. *Academic Radiology*, 27(5):603–608, May 2020. doi: 10.1016/j.acra.2020.03.003. URL `https://doi.org/10.1016/j.acra.2020.03.003`.

[22] X. Li, L. Yu, H. Chen, C.-W. Fu, L. Xing, and P.-A. Heng. Transformation-consistent self-ensembling model for semisupervised medical image segmentation. *IEEE Transactions on Neural Networks and Learning Systems*, page 1–12, 2020. ISSN 2162-2388. doi: 10.1109/tnnls.2020.2995319. URL `http://dx.doi.org/10.1109/TNNLS.2020.2995319`.

[23] N. Loizou, S. Vaswani, I. Laradji, and S. Lacoste-Julien. Stochastic polyak step-size for sgd: An adaptive learning rate for fast convergence. *arXiv preprint arXiv:2002.10542*, 2020.

[24] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. *CVPR*, 2015.

[25] J. Ma, C. Ge, Y. Wang, X. An, J. Gao, Z. Yu, and J. He. Covid-19 ct lung and infection segmentation dataset (version verson 1.0), 2020. URL `http://doi.org/10.5281/zenodo.375747`.

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

[27] MedSeg. Covid-19 ct segmentation dataset, 2020. URL `https://medicalsegmentation.com/covid19/`.

[28] S. Mounsaveng, I. Laradji, I. B. Ayed, D. Vazquez, and M. Pedersoli. Learning data augmentation with online bilevel optimization for image classification. *arXiv preprint arXiv:2006.14699*, 2020.

[29] Y. Ouali, C. Hudelot, and M. Tami. Semi-supervised semantic segmentation with cross-consistency training, 2020.

[30] D. P. Papadopoulos, J. R. Uijlings, F. Keller, and V. Ferrari. Training object class detectors with click supervision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6374–6383, 2017.

[31] Y. Qiu, Y. Liu, and J. Xu. Miniseg: An extremely minimum network for efficient covid-19 segmentation. *arXiv preprint arXiv:2004.09750*, 2020.

[32] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. *MICCAI*, 2015.

[33] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3), 2015.

[34] T. Schlegl, P. Seeböck, S. M. Waldstein, U. Schmidt-Erfurth, and G. Langs. Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In *International conference on information processing in medical imaging*, pages 146–157. Springer, 2017.

[35] F. Shan, Y. Gao, J. Wang, W. Shi, N. Shi, M. Han, Z. Xue, and Y. Shi. Lung infection quantification of covid-19 in ct images with deep learning. *arXiv preprint arXiv:2003.04655*, 2020.

[36] K. Sohn, D. Berthelot, C.-L. Li, Z. Zhang, N. Carlini, E. D. Cubuk, A. Kurakin, H. Zhang, and C. Raffel. Fixmatch: Simplifying semi-supervised learning with consistency and confidence, 2020.

[37] S. Vaswani, A. Mishkin, I. Laradji, M. Schmidt, G. Gidel, and S. Lacoste-Julien. Painless stochastic gradient: Interpolation, line-search, and convergence rates. In *Advances in Neural Information Processing Systems*, pages 3732–3745, 2019.

[38] S. Vaswani, F. Kunstner, I. Laradji, S. Y. Meng, M. Schmidt, and S. Lacoste-Julien. Adaptive gradient methods converge faster with over-parameterization (and you can do a line-search). *arXiv preprint arXiv:2006.06835*, 2020.

[39] L. Wang and A. Wong. Covid-net: A tailored deep convolutional neural network design for detection of covid-19 cases from chest x-ray images. *arXiv preprint arXiv:2003.09871*, 2020.

[40] L. L. Wang, K. Lo, Y. Chandrasekhar, R. Reas, J. Yang, D. Eide, K. Funk, R. Kinney, Z. Liu, W. Merrill, et al. Cord-19: The covid-19 open research dataset. *arXiv preprint arXiv:2004.10706*, 2020.

[41] S. Wang, M. Zhou, Z. Liu, Z. Liu, D. Gu, Y. Zang, D. Dong, O. Gevaert, and J. Tian. Central focused convolutional neural networks: Developing a data-driven model for lung nodule segmentation. *Medical image analysis*, 40:172–183, 2017.

[42] Y. Wang, C. Dong, Y. Hu, C. Li, Q. Ren, X. Zhang, H. Shi, and M. Zhou. Temporal changes of CT findings in 90 patients with COVID-19 pneumonia: A longitudinal study. *Radiology*, page 200843, Mar. 2020. doi: 10.1148/radiol.2020200843. URL `https://doi.org/10.1148/radiol.2020200843`.

[43] Y. Wang, C. Dong, Y. Hu, C. Li, Q. Ren, X. Zhang, H. Shi, and M. Zhou. Temporal changes of ct findings in 90 patients with covid-19 pneumonia: A longitudinal study. *Radiology*, 2020.

[44] Y. Wang, J. Zhang, M. Kan, S. Shan, and X. Chen. Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation, 2020.

[45] J. Wei, S. Wang, and Q. Huang. F3net: Fusion, feedback and focus for salient object detection. *arXiv preprint arXiv:1911.11445*, 2019.

[46] Q. Xie, Z. Dai, E. Hovy, M.-T. Luong, and Q. V. Le. Unsupervised data augmentation for consistency training, 2019.

[47] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500, 2017.

[48] C. Zheng, X. Deng, Q. Fu, Q. Zhou, J. Feng, H. Ma, W. Liu, and X. Wang. Deep learning-based detection for covid-19 from chest ct using weak label. *medRxiv*, 2020.

[49] T. Zhou, S. Canu, and S. Ruan. An automatic covid-19 ct segmentation based on u-net with attention mechanism. *arXiv preprint arXiv:2004.06673*, 2020.

[50] Y. Zhou, Y. Zhu, Q. Ye, Q. Qiu, and J. Jiao. Weakly supervised instance segmentation using class peak response. *CVPR*, 2018.

[51] Y. Zhou, Y. Zhu, Q. Ye, Q. Qiu, and J. Jiao. Weakly supervised instance segmentation using class peak response. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3791–3800, 2018.

[52] Z. Y. Zu, M. D. Jiang, P. P. Xu, W. Chen, Q. Q. Ni, G. M. Lu, and L. J. Zhang. Coronavirus disease 2019 (covid-19): a perspective from china. *Radiology*, page 200490, 2020.