Dense Regression Activation Maps For Lesion Segmentation in CT scans of COVID-19 patients

Weiyi Xie Colin Jacobs Bram van Ginneken

***Abstract*—Automatic lesion segmentation on thoracic CT enables rapid quantitative analysis of lung involvement in COVID-19 infections. Obtaining voxel-level annotations for training segmentation networks is prohibitively expensive. Therefore we propose a weakly-supervised segmentation method based on dense regression activation maps (dRAM). Most advanced weakly-supervised segmentation approaches exploit class activation maps (CAMs) to localize objects generated from high-level semantic features at a coarse resolution. As a result, CAMs provide coarse outlines that do not align precisely with the object segmentations. Instead, we exploit dense features from a segmentation network to compute dense regression activation maps (dRAMs) for preserving local details. During training, dRAMs are pooled lobe-wise to regress the per-lobe lesion percentage. In such a way, the network achieves additional information regarding the lesion quantification in comparison with the classification approach. Furthermore, we refine dRAMs based on an attention module and dense conditional random field trained together with the main regression task. The refined dRAMs are served as the pseudo labels for training a final segmentation network. When evaluated on 69 CT scans, our method substantially improves the intersection over union from 0.335 in the CAM-based weakly-supervised segmentation method to 0.495.**

***Index Terms*—Weakly-supervised semantic segmentation, class activation map, dense regression activation map, COVID-19, computed tomography, medical imaging.**

I. Introduction

THE coronavirus disease 2019 (COVID-19) has been declared a global pandemic since March of 2020. The total number of infected cases has reached over 83 million worldwide, with 1.8 million deaths by 2020. Unfortunately, both numbers are still increasing. To reduce the fatality rate, effective diagnosis and treatment planning are essential. As COVID-19 mainly damages the lungs of infected subjects, chest Computed Tomography (CT) plays a critical role in rapid diagnosis and progression monitoring of the COVID-19 infections. Based on chest CT analysis, standardized CT scoring systems, such as the COVID-19 Reporting and Data System (CO-RADS) [1], were defined to quantify the degree of suspicion of COVID-19 according to CT findings into 1-5 scores with an increasing level of suspicion. Similarly, a CT severity scoring system [2] was designed to assess the extent of parenchymal involvement of the disease. These scoring systems may be applied more accurately and rapidly when the automatic segmentation of infected areas (lesions) would be available. Therefore, this work aims at developing an algorithm that can automatically segment lesions related to COVID-19 on chest CT scans.

(*Corresponding author: Weiyi Xie, e-mail: weiyi.xie@radboudumc.nl*)

This work was supported by the Dutch Lung Foundation under the project 5.1.17.171.

All authors are with the Radboud university medical center, Radboud Institute for Health Sciences, Department of Medical Imaging, Nijmegen, The Netherlands.

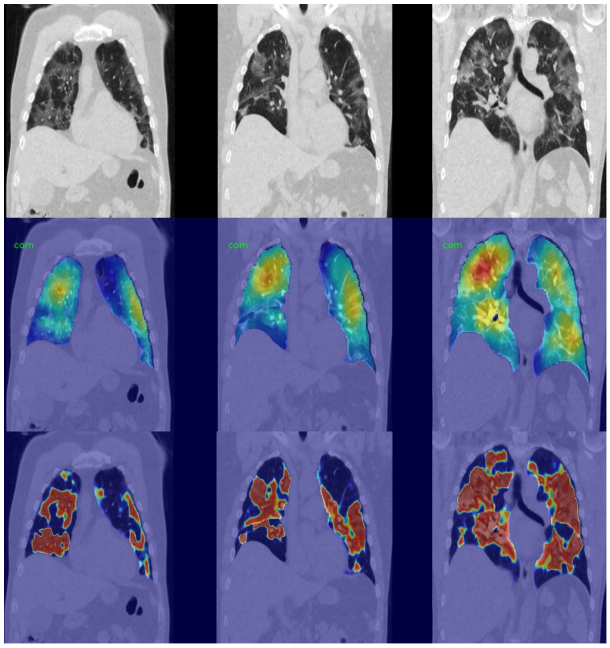

Fig. 1: Visualization of Class Activation Maps (CAMs) (2nd row) and dense regression activation Maps (dRAMs) (3rd row) in coronal views. CAMs and dRAMs were generated on the same subject (whose CT scan was shown at the first row) from our test set.

One of the major obstacles of semantic segmentation is the need to acquire a large amount of voxel-wise annotations for training the networks, which is particularly challenging when facing a new problem such as COVID-19. The lack of training data makes state-of-the-art supervised methods impractical. Therefore, in this work, we present a novel weakly-supervised segmentation method that only requires lobe-wise severity scores as the input reference for training and can produce dense and precise localized lesion maps that can be used as lesion segmentations.

Weakly-supervised semantic segmentation has been extensively studied in recent years. In the weakly-supervised setting, reference annotations can be provided using scribbles [3], or surface points [4]. Both these approaches seek a trade-off between annotation efforts and the amount of training information provided to the network regarding shapes and locations of target objects. However, because typical COVID-19 CT abnor-

malities often have bilateral lung involvement with a peripheral and diffuse distribution [5], manually annotating scribbles or extreme points could still be very demanding. To reduce the annotation cost further, it is favorable to use only image or region-level labels. Early weakly-supervised segmentation methods using image-level labels were based on multi-instance learning frameworks [6] and expectation-maximization algorithm [7]. The current state-of-the-art weakly supervised segmentation methods using image-level labels were based on class activation maps [8]–[11] (see recent results on PASCAL VOC2012 benchmark). CAMs correspond to the regions responsible for distinguishing image categories in a classification task. Because CAMs naturally only represent discriminative regions and may not fully cover or detect all objects, iterative approaches [11], [12] were proposed to erase already-found object maps in the previous iteration and force the network to discover new and complement regions at later iterations.

One major drawback of CAMs is that they are generated by taking high-level convolution features (at the bottom of the convolution neural network, usually before global pooling and linear layers) and multiplying them with class-specific weights in the linear layer. These high-level features contain rich semantic information but are generally at a low resolution compared with the input. The use of low-resolution features causes CAMs to lose local details, which is problematic since segmentation requires dense voxel-wise predictions. In addition, CAMs intrinsically reflect classification decisions, which are not necessarily aligned with the object segmentation task. Instead of using low-resolution features, BagNet [13] resorted to features in the earlier layers of convolution neural networks for extracting CAMs. Their method may indeed produce fine-resolution CAMs. However, low-level features do not suffice to represent complex objects without high-level semantics, leading to possibly very noisy CAMs. Another research direction is to use CAMs generated at a low resolution only as the initial seed regions. Extra steps were needed to refine CAMs for generating object segmentations. A seeded region growing module was proposed in [8] to expand CAMs towards the complete object boundaries in an iterative manner. AffinityNet [9] exploited local inter-pixel affinities as the transition probability matrix and applied random walks to revise CAMs. Many of these CAM refinement methods were implemented as post-processing steps. Therefore their hyper-parameters were tuned separately from the neural network training. For instance, random walks based on trained voxel-wise affinities were executed in separate post-processing steps to refine CAMs in AffinityNet [9].

Instead of relying on early layer features or refine CAMs in post-processing steps, we propose to train a segmentation network directly to generate high-resolution dense regression activation maps (dRAMs). We present a network trained for regressing the per-lobe lesion percentage. We used implied lesion percentage information from the lobe-wise severity scores, as typically provided by radiologists. When annotating lobe-wise severity scores, radiologists measure the lesion percentage per lobe and assign a corresponding score if the ratio falls in a specified range (Table I (b)). This lobe-level supervision limits lesion searching to lobes, which is considerably easier than that using scan-level labels in [14]. Meanwhile, the per-lobe lesion percentage is a richer type of information regarding the lesion volume, and such an approach was not used in previous CAMs-based classification approaches based on categorical labels. Because the per-lobe lesion percentage was defined as an interval given a lobe-wise severity score, we propose an interval regression loss to enforce the predicted percentage to fall in a particular range. Furthermore, we introduce an attention module for revising dRAMs, trained together with the regression task. The refinement of dRAMs is necessary because the regression target does not provide information regarding the object boundary. Inspired by AffinityNet, we intended to capture local voxel-wise affinities in the attention module, which enriches object semantics using neighboring information in revising dRAMs.

Our key contributions are as follows: 1) we propose a lesion segmentation framework that produces fine-resolution segmentation maps using only lobe-wise labels in training; 2) we convert the lesion segmentation problem to regression of the per-lobe lesion percentage defined by the lobe-wise severity score. The regression problem is solved using a proposed interval regression loss. These ideas are generic and can be extended to other weakly-supervised segmentation problems if specific statistics of the object segmented are available as the regression target; 3) we refine the dense regression activation maps using an attention neural network module and dense conditional random field trained together with the main regression target.

A. Related works

There are recent works on COVID-19 lesion segmentation that attempted to reduce the demand for voxel-level supervision in training. Fan et al. [15] proposed a semi-supervised training strategy that requires a few labeled images to train the initial segmentation model and leverages primarily unlabeled data to fine-tune the model progressively. Laradji et al. [16] proposed to use point-level labels in an active learning schema to generate lesion segmentation maps. Yao et al. [17] superimposed synthesized lesions on healthy CT scans for their network to learn to separate high-intensity structures such as vessels from possible COVID-19 lesions. Xu et al. [18] proposed a generative adversarial learning framework to segment COVID-19 lesions, which primarily relied on scan-level labels and used a small amount of voxel-level labeled data to initialize training. Wang et al. [14] proposed to train a binary CNN classifier based on the presence of COVID-19 on CT scans and used the classifier to generate CAMs for lesion localization. However, without the refinement of CAMs, their approach was limited to lesion localization rather than segmentation.

Our approach is closely related to CAMs-based weakly supervised segmentation approach. We differ in three general building blocks of these methods: 1) the generation of CAMs by training a convolution neural network, often in a classification task. 2) regularization, as an ill-defined problem, weakly-supervised semantic segmentation based on image labels requires networks to localize objects, whereas

such information is not given in the supervision. The loss of location information may cause training to converge to a trivial solution. Therefore, regularization is necessary. 3) CAM refinement is needed because CAMs do not necessarily reflect object shapes and align with the object boundaries.

In these three aspects, our approach generates dense class regression maps (dRAMs) by training a segmentation network towards a regression target, which is the major novelty of this paper. In terms of regularization, we use entropy loss to ensure the dRAMs differentiate foreground and background as confident as possible. Meanwhile, we use equivariant regularization proposed in [10] to improve training consistency in a self-supervised learning schema. The idea is to introduce an inherent constrain that dRAM produced by an affine-transformed input should be similar to affine-transformed dRAM produced by the original input. In terms of the refinement, our method is motivated by the use of local voxel affinities in [9], where authors proposed a network to predict semantic affinities among local pixel pairs. Then CAMs were refined by iteratively running random-walks in which the probability transitional matrix was derived from semantic affinities. A similar but end-to-end solution can be found in [10] where an attention module was designed for capturing global pixel affinities, which can be trained together with the main classification task. The dRAMs refinement process in our method is closely related to [10] and [9], while the major difference is that we capture voxel affinities via attention maps inspired by [10], but computed within a local neighborhood similar to that in [9]. The use of local affinities is because we intended to rely on local details in revising dRAMs.

II. Data

In this study, we used CT scans from patients who presented at the emergency wards of the Radboud University Medical Center, the Netherlands, from March to September 2020. Patients were referred for CT imaging because of suspicion of moderate to severe COVID-19 pneumonia. The ethical review board approved the retrospective and anonymous collection of this data (Radboudumc CMO2016-3045, Project 20027). All CT scans were obtained with a low-dose thin slice protocol without administration of contrast. Further details can be found elsewhere [19].

Following the guidelines of the Dutch Radiollgical Society [1], the radiology report for each scan contained CO-RADS and lobe-wise severity scores. CO-RADS 1 is defined as a scan that is normal or has non-infectious etiologies, and thus a very low level of suspicion for COVID-19. CO-RADS 2 indicates the CT-scan has features typical for infections other than COVID-19. CO-RADS 3 indicates equivocal findings: features compatible with COVID-19 but also with other diseases. CO-RADS 4 and 5 indicate a high and very high level of COVID-19 suspicion, respectively. CO-RADS 6 was given to scans from patients that were already known to be positive for COVID-19 with reverse transcription-polymerase chain reaction (RT-PCR) tests at the time of reporting. Lobe-wise severity scores indicate the extent of lobar involvement of the COVID-19 infection. A score from 0 to 5 is assigned to each lobe according to the visually assessed lesion percentage of that lobe. The total CT severity score is the summation of lobe-wise severity scores. The mapping between lobar severity score and lesion percentage per lobe can be found in Table I(b). We used lobe-wise severity scores as the weak labels in training our models.

TABLE I: The distribution of CO-RADS scores and lobe-wise severity scores across the training and test sets in the primary data collection. CO-RADS score 1-6 indicates the level of suspicion for COVID-19 positive disease, ranging from very low, low, equivocal, high, very high, and confirmed PCR positive, respectively. Lobe-wise Severity scores indicate the extent of lobe-wise involvement due to COVID-19 infection.

(a) CO-RADS scores

CO-RADS	#subjects for training	#subjects for testing
1	10	1
2	17	1
3	116	14
4	61	15
5	90	26
6	28	12
Total	322	69

(b) Lobe-wise severity scores

severity scores (percentage per lobe)	#training lobes	#testing lobes
0 (0%)	410	52
1 (1-5%)	401	64
2 (5-25%)	401	114
3 (25-50%)	226	69
4 (50-75%)	131	34
5 (75-100%)	41	12
Total Lobes	1610	345
Total Scans	322	69

1) Data Selection and Partitioning: For this study, we selected 391 subjects (randomly split into 322 for training and 69 for testing). This selection included all subjects that were available when this project started. A single scan was used for each subject. Thirty subjects in the training set were used as the validation set during model development to prevent overfitting. The distribution of CO-RADS and lobe-wise severity scores is provided in Table I (a and b).

In addition to this primary data set, we later randomly selected another 435 subjects not included in the primary data collection. Their baseline CT scans were all reported with a total severity score of 0 and CO-RADS 1. These 435 CT scans were used as an auxiliary data collection for training our vessel segmentation network (see Sect. III-A3).

2) Reference Standard: For evaluating our method, lesion segmentation references on 69 test scans in the primary collection were obtained from Thirona (Nijmegen, the Netherlands), a medical image analysis service company specializing in chest CT analysis.

First, lung parenchyma regions with a higher attenuation were identified by thresholding and morphologic operations. Automatic methods were used to suppress vessels and airways. Following the radiology report, lesion candidates in lobes

not affected by COVID-19 were then removed. A certified image analyst with at least one year of experience reviewed the remaining lesion candidates and corrected segmentations where necessary.

The analysts also labeled segmented lesions into ground-glass, consolidation, and mixed to evaluate segmentation performance for different lesion subtypes. During the annotation process, the analyst could consult a radiologist in cases of doubt.

III. Methods

A. Weakly-supervised segmentation framework

The overview of the proposed weakly-supervised lesion segmentation framework is shown in Fig. 2. We first trained a regression network to predict the lesion percentage per lobe and in the process we generate the dense regression activation maps (dRAMs). Since the lobe-wise severity scores (Table I (b)) represent a lesion percentage per lobe in a range (a severity score 1 indicates the percentage of lesion involvement in the lobe is within the range of 1%-5%, e.g.), we propose an interval regression loss for training the regression network. In addition to the regression, dRAMs are refined in an auxiliary training task that employs a dense conditional random field and an attention mechanism. Also, an independently trained vessel segmentation model was used in the refinement process to suppress false detected vessels. Moreover, regularization techniques were used to stabilize the regression training. Finally, we used the refined dRAMs as pseudo segmentation references for training a segmentation network from scratch. The following subsections elaborate on each of these steps.

1) Generation of dense regression activation maps: The dense regression activation map is generated by training a regression network for predicting the lesion percentage per lobe. We used standard 3D U-Net [20] as the regression network (detailed parameters in Fig. 2 (a)) because of its simplicity and robustness in various medical segmentation tasks. The 3D U-Net has three down-sampling layers in the encoding path, and each layer consists of two convolutions and a max-pooling operation. Following the down-sampling path, two more convolutions are used to double the number of convolution filters. In the up-sampling path, three layers are used to reconstruct the resolution, and each contains one tri-linear interpolation, followed by two convolutions to reduce the interpolation artifacts. Before the final one, convolution kernels have $3 \times 3 \times 3$ kernel size, a stride of 1 voxel, and zero-padding. The last convolution is a $1 \times 1 \times 1$ convolution to squeeze features to have a single channel before applying sigmoid activation. The network takes an $80 \times 80 \times 80$ chunked image as the input, which is cropped around each segmented lobe and resized. The segmentation of pulmonary lobes was done using a publicly-available algorithm [21]. The input and the output size of the 3D U-Net are identical as we used zero-padding. For each lobe chunk input, the region outside the lobe of interest was set to zero. The output of this network is referred to as the dense regression activation map (dRAM).

The lobe chunk image as the input allowed us to compute the lesion percentage for the given lobe by simply averaging the dRAM over all voxels within the lobe (lobe-wise mean pooling). The reference for regression training was the lobe-wise severity score reflecting only a particular interval of the per-lobe lesion percentage (see Table I (b) for the mapping between lobe-wise severity scores and lesion percentage per lobe). Therefore, we propose an interval regression loss that enforces the predicted percentage fall into a corresponding predefined range. Denoting the predicted lesion percentage as $\hat{y}$, the lower bound and the upper bound of the percentage range defined by the severity score as r_l and r_u, we defined the interval regression loss function to minimize as

$$\begin{aligned} &max(0, (\hat{y}\text{–}0.5 * (r_l + r_u))^2 - K), \\ &\text{where } K = (0.5 * (r_l\text{–}r_u))^2. \end{aligned} \tag{1}$$

This can also be interpreted as the quadratic version of the piecewise linear loss function that minimizes $|\hat{y}\text{–}r_l| + |\hat{y} - r_u| - |r_u - r_l|$. The interval regression loss is weighted on each instance (a lobe image chunk) by the reverse frequency of the corresponding severity score in the training set.

2) Regularization techniques for regression training: Training on weak labels may converge to a trivial solution because information on the location of objects or, in our case, abnormal regions, is not available. Therefore, regularization techniques are commonly used to stabilize training. Wang et al. [10] introduced an implicit equivariant constraint for training their weakly supervised segmentation networks based on class activation maps (CAMs). Their basic idea was to enforce CAMs produced by an affine-transformed input be similar to the affine-transformed CAMs produced by the original input.

Denote the 3D U-Net for regression training as $F(\cdot)$, a predefined spatial affine transformation as $A(\cdot)$, and an input image to the network as I. Then $F(I)$ represents the dRAM. Equivariant regularization can be formulated as

$$R_{ER} = ||F(A(I)) - A(F(I))||_1. \tag{2}$$

Equivariant regularization can also be interpreted as a way to introduce the self-supervising correspondences between affine-transformed objects. The affine transformations used in this paper are resizing and rotation at a random scale or angle.

Additionally, we introduce entropy regularization to reduce uncertainty in the generated dRAM. Given dRAM as $F(I)$, which is already rescaled into a probability distribution by a sigmoid activation before lobe-wise average pooling, we introduce an entropy regularization term that minimizes

$$R_E = -F(I) * \log(F(I)) - (1.0 - F(I)) * \log(1.0 - F(I)). \tag{3}$$

3) Vessel segmentation: During our initial experiments, we observed that raw dRAMs may erroneously include vessels, possibly caused by the interval regression target since this only defines a range of acceptable per-lobe lesion percentages. To suppress vessels in our framework, we trained a separate 3D U-Net segmentation network (using the same architecture as our regression network) on the auxiliary data collection of 435 CT scans without COVID-19 CT signs (details in II-1) for segmenting vessels. Here the training references were generated by applying Otsu's threshold [22] on vessel maps generated by a Frangi filter [23]. Using this model, we generated vessel segmentations for all 322 training scans in the

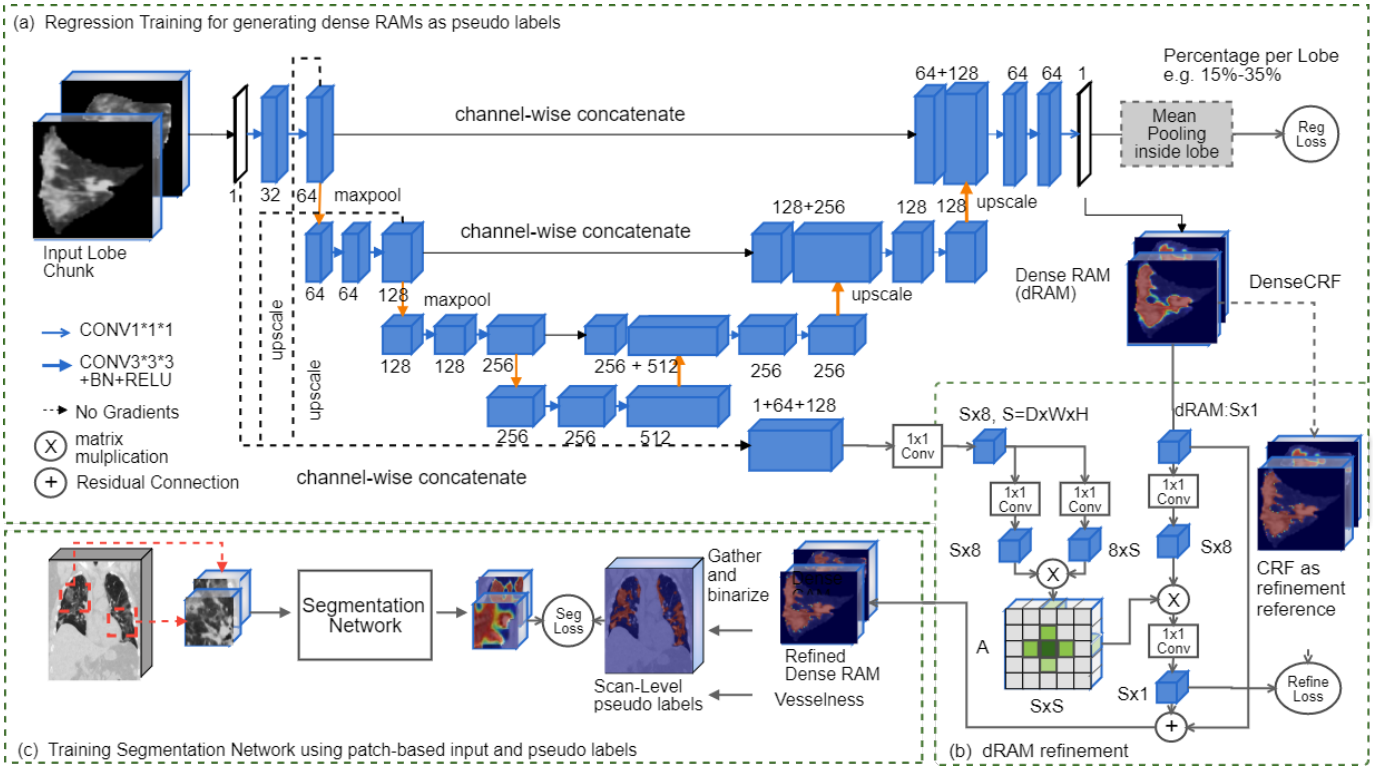

Fig. 2: Overview of the proposed weakly-supervised segmentation framework. (a) Our architecture is similar to a 3D U-net, but not trained with segmentation masks, only performing a regression task, i.e., predicting the percentage of lesions per lobe. At the end of the network, a dense regression activation map (dRAM) is generated, corresponding to the lesion segmentation. (b) Depicts the attention module for dRAM refinement, according to a synthetic refinement reference generated by applying dense conditional random field on the raw dRAM and vessel suppression. (c) We train an independent segmentation network using scan-level pseudo labels gathered from the lobe-wise refined dRAMs and vessel segmentation.

primary data collection, taking voxels with confidences above 0.7 as vessels from the model prediction. The predicted vessels were used as object cues in our framework to suppress false lesions. Such low-level features were commonly exploited in weakly-supervised semantic segmentation. As an example, Wei et al. [24] used saliency maps based on low-level image features to train convolutional segmentation networks in the weakly supervised settings in a progressive manner.

4) Refinement of dense regression activation maps: As dRAMs were obtained by training on a regression target and in this process, no voxel-wise supervision was provided, raw dRAMs may not suffice to delineate lesions accurately. To alleviate this issue, we proposed to refine dRAMs using an auxiliary training objective. We post-processed dRAMs by suppressing vessels using vessel segmentations (III-A3) and applying dense conditional random field (denseCRF) on dRAMs. The post-processed dRAMs were used as the training target to provide voxel-wise supervision in the refinement step. As shown in Fig. 3, vessels were suppressed in the dRAM, and dense conditional random field helped to refine the lesion borders in the refinement target (the 4th row). We used a bootstrapping loss [25] for the refinement training because both dRAMs and vessel segmentations were generated by automatic methods and may contain noise. The bootstrapping loss function minimizes

$$\begin{aligned} &\textstyle\sum_{k=0}^{L}[\beta t_k + (1-\beta)z_k]log(q_k) \\ &z_k = 1[k = argmax(q_i)], i = 0, 1, .., L \end{aligned} \tag{4}$$

where L is the number of classes (3 in our case, including background, vessel, and lesion), for the class label k, t_k is one-hot encoding pseudo reference and z_k is the bootstrapping reference. q_k is the softmax probability of assigning a voxel into the class k. β is set to 0.8. Note that we detached the computation of the bootstrapping reference z_k in the gradient back-propagation. The idea of this loss is to leverage the knowledge learned by the network during training to provide hints for the true labels.

To further improve the dRAM refinement, we added an attention module on top of the generated dRAMs from the regression training. This module calculated local affinities using low-level convolution features from the regression network and image intensities. First, the input image was concatenated with convolution features before pooling at the first and second layers of the regression network (up-scaled to the same size as the input image). This concatenation is for capturing the low-level information, including raw voxel intensities. We detached these concatenated features out of the back-propagation computation. We then reduced the concatenated features to have eight channels via a $1 \times 1 \times 1$ convolution filter to save computational memory.

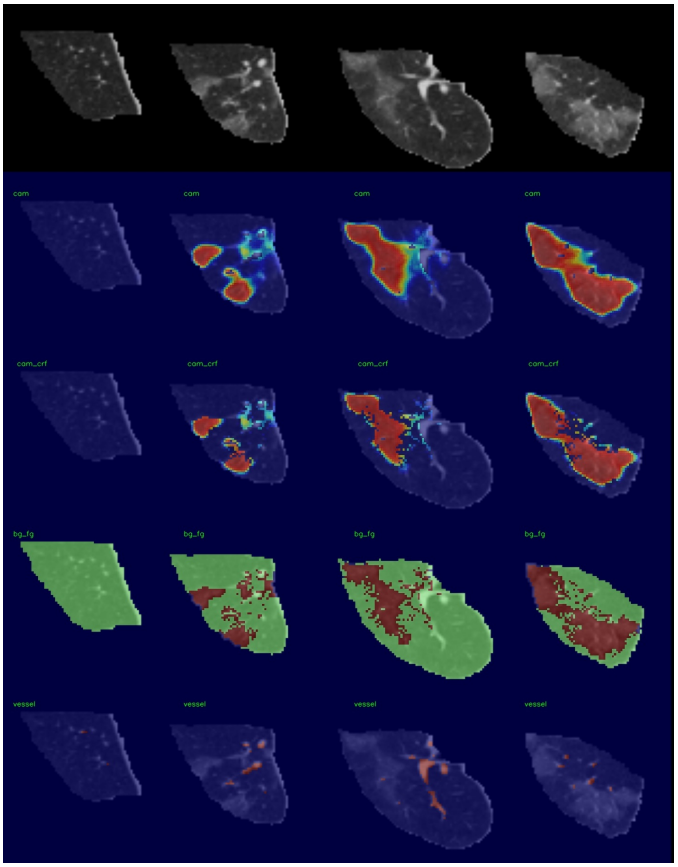

Fig. 3: The generation of the refinement target. The top row shows a lobe chunk image. The second and third-row visualize the dRAM and dense CRF result after the suppression of vessels, respectively. The fourth row shows the synthesised target labels for the refinement step, including foreground ■ and background ■. The bottom row shows the vessel segmentation.

Denote this reshaped feature map as x, $x \in R^{D\times W\times H\times 8}$. The affinity $a(x_i, x_j)$ between two locations i and j in x can be computed in an embedded Gaussian function $a(x_i, x_j) = e^{(W_\theta x_i)^T (W_\phi x_j)}$, where W_θ and W_ϕ are linear transformations without changing dimensions of the input. Local affinities for a location i were measured by computing pairwise $a(x_i, x_j)$ between i and surrounding locations j in a $3 \times 3 \times 3$ spatial window taking i as the center with a connectivity of 2, resulting in total 19 neighboring pairs including the self-connection. Computing local affinities for all locations in x resulted in a attention map A, $A \in R^{D\times W\times H\times 19}$. Local affinities for location i were normalized over neighboring locations Ωj of i. The normalizing factor $\zeta(x)$ was simply a summation $\zeta(x) = \sum_{j\in\Omega_j} a(x_i, x_j)$. In matrix form, this normalization is equivalent as applying softmax function over the last dimension on A. To revise dRAM y, we first project dRAM into a subspace by $g(y)$ implemented as the linear projection $W_g^T y$, and apply matrix multiplication between the projected dRAM and the attention map A. This matrix multiplication can be seen as each location in dRAM selectively collecting information from its local neighbors. The impact from local neighbors in updating a dRAM location was determined by their pairwise affinities. This is very similar to propagate messages from local neighbors using random-walks for refining CAMS in [9].

Since the updated dRAM is in a subspace, we projected it back using a linear transformation $r(\cdot)$ or W_r in the form of matrix multiplication. The whole process of using local affinities to refine dRAM y in the location i can be formulated as follows

$$\hat{y_i} = r(\frac{1}{\zeta(x)} \sum_{\Omega j} a(x_i, x_j) g(y_j)) + y_i. \quad (5)$$

The use of residual connections allows gradients to flow through a network directly if the dRAM already provided a good segmentation. Note that dRAM refinement branch in Fig. 2 (b) and the segmentation branch in Fig. 2 (a) were trained simultaneously along with the regularization loss terms. The total loss is the weighted average of the regression term, regularization terms, and the refinement term. The weight for the main regression target was set to 2.0, and the rest of the learning objectives were weighted by 1.0.

5) Context aggregation: In this step, a lesion segmentation network was trained from scratch using pseudo lesion and vessel labels. This segmentation network is a standard 3D U-Net, the same as the regression network. This step is necessary because the regression network may overlook features across lobes due to the use of lobe chunk images as the network input. Therefore, this step constructs a scan-level dRAM by filling lobe-wise refined dRAMs back to the scan from which the lobe chunks were cropped. The scan-level dRAM was used as the pseudo lesion label. We re-sampled scans and pseudo labels into an isotropic spacing of 1.5 millimeters to set the receptive field of the network to 132 millimeters (88 voxels in the 3D U-Net). The input to the network is a mini-batch of two $132 \times 132 \times 132$ 3D image chunks, randomly cropped from the scan during training, and the corresponding output chunks in size of $44 \times 44 \times 44$ due to valid-padding in convolutions (this is the only difference compared to what we use in the regression network).

The pseudo vessel and lesion labels were stacked together channel-wise, where 0 indicates background, 1 denotes the vessels, and 2 indicates lesions. Due to the possible noise in the pseudo labels, the final segmentation training adopted the bootstrapping loss function (Eq. 4). Because the final segmentation network was trained in a patch-based fashion, the softmax probabilities of all 3D output chunks are tiled together by sliding over the entire scan without overlap to build up a scan-level probability map. The lesion prediction was assigned if a maximum probability was found on the 2nd channel in the softmax probability map.

6) Lesion segmentation in different subcategories: Given the lesion segmentation results, we further labeled each connected component in the segmentation into one of three subcategories: ground-glass, consolidation, and mixed. We adopted a non-parametric approach based on Kullback–Leibler divergence (KLD) for the similarity measurements. We had analysts manually segment and label COVID-19 lesions in six selected CT scans from the validation set in our main data

TABLE II: Segmentation results of the proposed weakly supervised segmentation framework (dRAM) in comparison with the baseline approach (CAM) and the fully supervised segmentation framework (nn-UNet) on the test set in the main data collection. Weakly-supervised methods are denoted by W, fully supervised methods as F. The best weakly supervised method is shown in bold, evaluated by a Wilcoxon signed-rank test (p <0.01 with Bonferroni correction)

a) Segmentation performance

Method	IoU			
	Overall	Consolidation	GGO	Mixture
CAM (W)	0.335	0.527	0.049	0.165
dRAM (W)	0.495	0.710	0.182	0.273
nnU-Net (F)	0.619	0.782	0.276	0.409

b) Ablation study

Method	Ablation	IoU
nnU-Net (F)	-	0.619
CAM (W)	raw	0.222
	+refine	0.336
dRAM (W)	raw	0.395
	+regularizer	0.435
	+refine	0.452
	+attention	0.469
	+vessel suppression	0.475
	+context	**0.495**

collection, including lesions with all three subcategories. We performed connected component analysis for these reference scans using the segmentation references and computed the mean and standard deviation for all the components. For each test scan, the same connected component analysis was applied to the segmentation map of our method. Moreover, we looped through all test components to compute the mean and standard deviation. Assuming that intensity values were Gaussian distributed for each component, we can compute pairwise KLDs between a test component and all components in the labeled six segmentation references to measure the similarity between components. Due to the impact of the component size in computing the statistics, we first shortlisted K components in the references with the smallest differences in lesion volumes (K=10) to a test component. Among shortlisted components in the references, components with the smallest N KLD were selected for weighted label voting (N=5). Weights were determined by their rankings in their KLD similarities to the test component. KLD for two Gaussians is defined as:

$$KLD(p,q) = log(\frac{\sigma_2}{\sigma_1}) + \frac{\sigma_1^2 + (\mu_1 - \mu_2)^2}{2\sigma^2} - \frac{1}{2}, \quad (6)$$

where q and p are distributed under $N(\mu_1, \sigma_1)$ and $N(\mu_2, \sigma_2)$, respectively.

IV. Results

A. Training and testing details

Training, validation, and testing of each experiment were carried out on a machine with an NVidia TitanX GPU with 12 GB memory. The methods were implemented using Python 3.6 and the Pytorch 1.1.0 library [26]. The trainable parameters of each method were initialized using Kaiming He initialization [27] and were optimized using stochastic gradient descent with a momentum of 0.9, and the initial learning rate is set to 10^{-5}. Dense conditional random field was implemented using the Pydensecrf python library [28][1].

During training and testing, CT scans were standardized by clamping intensity values to the $[-1200 \sim 300]$ range before re-scaling into $[0 \sim 1]$. We segmented lobes using an automated algorithm proposed in [21] on all the CT scans, used for masking out regions outside the lobes during the training and the testing.

We applied random flip, resampling, and contrast stretching as data augmentation methods during all methods training.

We resampled input scans into an isotropic spacing of 1.2 millimeters (with a small random jittering) by tri-linear interpolation for training regression networks. Input chunk images were rescaled using tri-linear interpolation by a factor of 0.8, 1.0, 1.2, and 1.5 when running the regression network at a test time to generate multi-scale dRAMs. These dRAMs were merged by averaging.

For training the final segmentation network, we resampled the scans and pseudo labels into a fixed isotropic spacing of 1.5 millimeters by tri-linear interpolation for both training and testing. Not using multi-scale input images and test ensembles guarantied the runtime efficiency of our final model.

B. Evaluation Metrics

The Intersection over Union (IoU), also known as the Jaccard index, between predictions and segmentation references, was used to evaluate segmentation performance. The IOU between two binary masks X, Y is defined as:

$$IoU(X,Y) = \frac{|X \cap Y|}{|X \cup Y|}. \quad (7)$$

C. Methods in comparison

1) The fully-supervised method: To evaluate the segmentation performance of a model trained on voxel-wise labels, analysts annotated lesions on 108 scans from our training set in the main data collection using the same protocol as defined in II-2. We trained a 3D U-Net based on the nn-UNet [29] framework, which has shown superior performance in many medical image segmentation challenges. The framework itself is an implementation of U-Net but took advantage of model ensembles (2D, 3D, and Cascading U-Nets), rich data augmentation techniques, and the combination of state-of-the-art segmentation loss functions. In their framework, all of these configurations are automatically adapted during training to data at hand. We resampled scans into 1.5 millimeters isotropic spacing by tri-linear interpolation for training and testing the nn-UNet.

2) Weakly supervised methods: We used the standard CAM-based weakly-supervised segmentation as the baseline approach, where the network is only the down-sampling path of the 3D U-Net. During training, CAMs were generated by applying the fully-connected layer on the lowest resolution features before the global average pooling. The network was

[1]https://github.com/lucasb-eyer/pydensecrf

trained to classify each lobe chunk image into two classes: positive if lesions are present and negative if not. Training and testing hyperparameters (data augmentation, resampling of input images, test ensembles and model initialization, e.g.) were the same as those used in the proposed method. During testing, lobe-wise CAMs were tiled back to become a scan-level CAM, which was rescaled into $[0 \sim 1]$ by subtracting the minimum and dividing the maximum. The lesion segmentation was obtained by binarizing the rescaled scan-level CAM using Otsu's threshold. Because of using features at a low resolution, results from raw CAMs were far from the segmentation expectation.

To obtain a reasonable performance, we applied Otsu's segmentation method within the lung area to obtain high-attenuation in the lung as the lesion proposal after excluding vessels predicted by the trained vessel segmentation network (see Sect. III-A3). Then CAMs were refined by excluding regions outside the lesion proposal. The result from the refined CAMs was denoted as CAM in Table IV (a). Results from both raw and refined CAMs can be found in the ablation study in Table IV (b). The baseline result refers to the result using refined CAMs.

To evaluate the effectiveness of each component in the proposed weakly supervised segmentation framework based on dRAMs, we conducted the following ablation study. Raw dRAMs (denoted as raw in Table IV(b)) were trained using only the regression loss. This demonstrates the benefits of using fine resolution features from a segmentation network and training towards the regression target rather than the classification training using CAMs. The importance of proper regularization was measured by adding entropy and equivariant losses (denoted as regularizer) to raw dRAMs training. On top of using regularizers, refinement steps with or without attention module were both evaluated. Additional advantages of using vessel suppression in the refinement step were also assessed on top of the contribution of the attention module. Finally, we reported the performance of the proposed method, which was the segmentation model in context aggregation step trained using pseudo lesion and vessel labels (denoted as context in Table IV (b) and dRAM in Table IV (a)). A Wilcoxon signed-rank test was employed to assess whether the performance difference was statistically significant ($p < 0.01$ with Bonferroni correction). The best weakly-supervised approach was significantly better than all ablated alternatives and is shown in bold in Table IV (b).

D. Quantitative Results

From the results in Table IV, the proposed method reaches 0.495 overall IoU compared with 0.336 IoU achieved by the baseline method in weakly-supervised settings trained using only lobe-wise severity scores. This performance improvement can also be seen in the segmentation of all lesion sub-categories. Meanwhile, both weakly-supervised segmentation methods are outperformed by the nnU-Net method trained on voxel-wise labels because of the rich semantic information embedded in voxel labels. The ablation study demonstrates that adding regularizers improves the performance dramatically from 0.395 to 0.435, benefiting from self-supervised training and entropy minimization. The refinement auxiliary training task further improves the performance to 0.452, which is caused by the refinement of lesion borders under the guidance of dense conditional random field. The attention module in the refinement process further boosts the performance to 0.469 by learning voxel-wise affinities. Vessel suppression provided additional improvement by reducing false lesions, forcing the regression training to discover other regions associated with the per-lobe lesion percentage. Finally, the context aggregation step recollects contextual features across lobes in a patch-based training, resulting in an IoU of 0.495.

E. Prediction of the CT Severity Score

Based on the segmentation prediction of abnormal regions in each lobe, the algorithm can output the lesion percentage per lobe, which can be translated into lobe-wise severity scores via the mapping in Table II-1(b). We computed linear weighted kappa scores for the baseline method, the proposed method, and the nnU-Net method trained with voxel-wise labels against the lobe-wise severity scores assigned by the radiologist. The kappa scores were categorized as slight, fair, moderate, good, or excellent based on k values of 0.20 or less, 0.21–0.40, 0.41–0.60, 0.61–0.80, and 0.81 or higher, respectively, following [30]. The k value was 0.392 (95% CI: 0.334, 0.449) for the baseline method, 0.461 (95% CI: 0.399, 0.524) for the proposed, and 0.514 (95% CI: 0.459, 0.570) for the nnU-Net method. There was a moderate agreement between the predicted scores by the proposed method and manual scores, but a fair agreement between the scores predicted by the baseline method and manual scores.

F. Qualitative Results

As shown in Fig.4, the result from raw CAMs often exhibited from substantial over-segmentation (1st row). This is because CAMs were computed from the low-resolution features and resampled to the original resolution by interpolation. The baseline method (2nd row) using refined CAMs often missed lesions (4th and fifth columns) due to the difficulties of finding optimal thresholds in post-processing CAM refinement steps. The proposed method (4th row) generally performed well on lesion segmentation. Compared with the nn-UNet results, the proposed method produced more false positives in regions near vessels with low attenuation. One reason is that the lobe-wise severity scores only represent an interval of the per-lobe lesion percentage, and not the precise percentage, which potentially allows the network to tolerate certain mistakes. This can also cause the method to be less precise for small ground glass lesions (see the subpleural region of the left upper lung in the 5th column). On the other hand, the regions near vessels appeared in a similar intensity range as the ground-glass opacities and were possibly related to inflammation caused by COVID-19. In general, ground-glass opacities may create challenges in visual recognition. This challenge may also cause measurement errors for radiologists in labeling severity scores, further contributing to confusion regarding ground-glass opacities in our method.

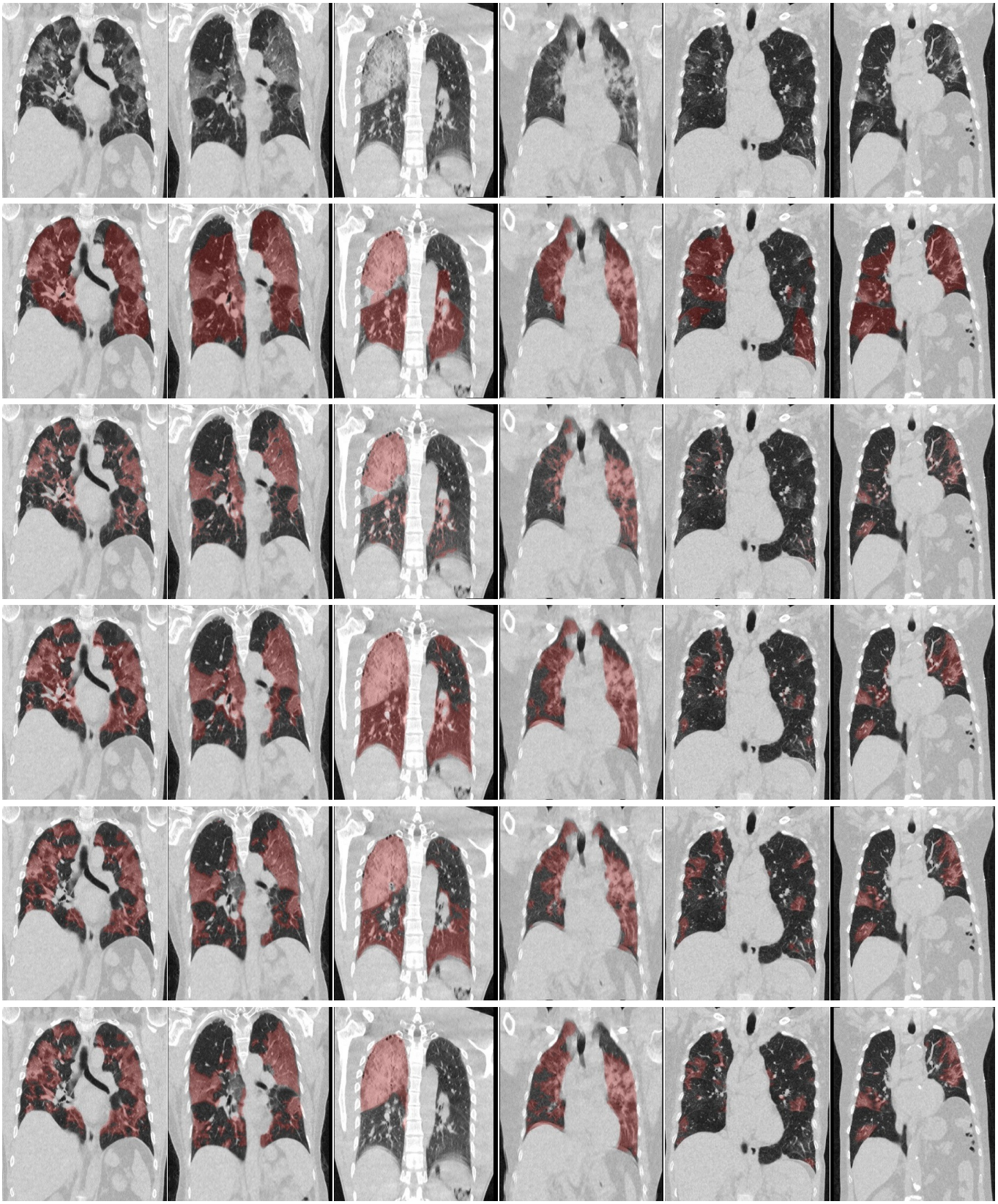

Fig. 4: Segmentation results for six representative test cases represented in columns. The 1st row shows six scans represented in a coronal slice. At the same slice, the 2nd, 3rd, 4th, and 5th rows show the segmentation results of the raw CAM, the baseline (refined CAM), the proposed and nnU-Net method, respectively. The last row shows the segmentation reference.

V. Discussion and Conclusion

We proposed a novel weakly-supervised segmentation method. This method is able to train a segmentation network using only severity scores provided for individual lobes, where these scores correspond to a range of percentages of affected regions in these lobes. The use of such visually assessed percentages of affected regions is common in radiological scoring systems. From only these lobe scores, the network is able to generate dense regression activation maps (dRAMs). These dRAMs were refined by aligning with the outputs from dense conditional random fields. We also proposed an attention module that enriches the semantic representation at each voxel based on its local neighbors (affinities). Pseudo labels were generated based on refined dRAMs and an additional step to remove false responses on vessels. The final segmentation was trained from scratch based on the pseudo labels using a bootstrapping loss to handle possible noise in the pseudo labels.

The proposed method achieved significant improvements in segmentation performance compared with the baseline approach. In terms of the prediction of lobe-wise severity scores, the proposed method reached a moderate agreement with the scores assigned by the radiologist, while the baseline method only reached a fair agreement. As we showed in our results, the proposed model sometimes produced false positives in the regions near vessels and may miss small ground glass lesions. However, as weak labels are cheap to collect, more advanced approaches can be built upon our model using our methods as the initial seed for interactive (e.g., adaptive learning scenarios) or iterative refinement (e.g., knowledge distillation).

The proposed method is generic and can be easily adapted to other weakly-supervised segmentation problems if specific object statistics are given and can be used as the regression target. We believe that the proposed weakly-supervised segmentation framework can be used for many segmentation problems in medical imaging, where automatic segmentation is often used for quantification analysis. In these scenarios, visually assessed quantification results from radiological scoring systems can be directly used in our framework as the regression target.

References

[1] M. Prokop, W. van Everdingen, T. van Rees Vellinga, J. Quarles van Ufford, L. Stoger, L. Beenen, B. Geurts, H. Gietema, J. Krdzalic, C. Schaefer-Prokop, B. van Ginneken, M. Brink, and the COVID-19 Standardized Reporting Working Group of the Dutch Radiological Society, "CO-RADS - a categorical CT assessment scheme for patients with suspected COVID-19: definition and evaluation," *Radiology*, vol. 296, no. 2, pp. E97–E104, 2020.

[2] K. Li, J. Wu, F. Wu, D. Guo, L. Chen, Z. Fang, and C. Li, "The clinical and chest CT features associated with severe and critical COVID-19 pneumonia," *Investigative Radiology*, vol. 55, no. 6, pp. 327–331, 2020.

[3] Z. Ji, Y. Shen, C. Ma, and M. Gao, "Scribble-based hierarchical weakly supervised learning for brain tumor segmentation," in *Medical Image Computing and Computer-Assisted Intervention*, ser. Lecture Notes in Computer Science, vol. 11766, 2019, pp. 175–183.

[4] H. Roth, L. Zhang, D. Yang, F. Milletari, Z. Xu, X. Wang, and D. Xu, "Weakly supervised segmentation from extreme points," in *Medical Image Computing and Computer-Assisted Intervention*, ser. Lecture Notes in Computer Science, vol. 11851, 2019, pp. 42–50.

[5] H. Shi, X. Han, N. Jiang, Y. Cao, O. Alwalid, J. Gu, Y. Fan, and C. Zheng, "Radiological findings from 81 patients with COVID-19 pneumonia in Wuhan, China: a descriptive study," *Lancet Infectious Diseases*, vol. 20, no. 4, pp. 425–434, 2020.

[6] P. O. Pinheiro and R. Collobert, "From image-level to pixel-level labeling with convolutional networks," in *Computer Vision and Pattern Recognition*, 2015, pp. 1713–1721.

[7] G. Papandreou, L.-C. Chen, K. P. Murphy, and A. L. Yuille, "Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation," in *International Conference on Computer Vision*, 2015, pp. 1742–1750.

[8] Z. Huang, X. Wang, J. Wang, W. Liu, and J. Wang, "Weakly-supervised semantic segmentation network with deep seeded region growing," in *Computer Vision and Pattern Recognition*, 2018, pp. 7014–7023.

[9] J. Ahn and S. Kwak, "Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation," in *Computer Vision and Pattern Recognition*, 2018, pp. 4981–4990.

[10] Y. Wang, J. Zhang, M. Kan, S. Shan, and X. Chen, "Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation," in *Computer Vision and Pattern Recognition*, 2020, pp. 12 275–12 284.

[11] Y. Wei, J. Feng, X. Liang, M.-M. Cheng, Y. Zhao, and S. Yan, "Object region mining with adversarial erasing: A simple classification to semantic segmentation approach," in *Computer Vision and Pattern Recognition*, 2017, pp. 1568–1576.

[12] C. González-Gonzalo, B. Liefers, B. van Ginneken, and C. I. Sánchez, "Iterative augmentation of visual evidence for weakly-supervised lesion localization in deep interpretability frameworks: application to color fundus images," *IEEE Transactions on Medical Imaging*, vol. 39, no. 11, pp. 3499–3511, 2020.

[13] W. Brendel and M. Bethge, "Approximating CNNs with bag-of-local-features models works surprisingly well on imagenet," *arXiv:1904.00760*, 2019.

[14] X. Wang, X. Deng, Q. Fu, Q. Zhou, J. Feng, H. Ma, W. Liu, and C. Zheng, "A weakly-supervised framework for COVID-19 classification and lesion localization from chest CT," *IEEE Transactions on Medical Imaging*, vol. 39, no. 8, pp. 2615–2625, 2020.

[15] 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*, vol. 39, pp. 2626–2637, 2020.

[16] I. Laradji, P. Rodriguez, F. Branchaud Charron, K. Lensink, P. Atighehchian, W. Parker, D. Vazquez, and D. Nowrouzezahrai, "A weakly supervised region-based active learning method for COVID-19 segmentation in CT images," *arXiv:2007.07012*, 2020.

[17] Q. Yao, L. Xiao, P. Liu, and S. K. Zhou, "Label-free segmentation of COVID-19 lesions in lung CT," *arXiv:2009.06456*, 2020.

[18] Z. Xu, Y. Cao, C. Jin, G. Shao, X. Liu, J. Zhou, H. Shi, and J. Feng, "GASNet: Weakly-supervised framework for COVID-19 lesion segmentation," *arXiv:2010.09456*, 2020.

[19] N. Lessmann, C. I. Sánchez, L. Beenen, L. H. Boulogne, M. Brink, E. Calli, J.-P. Charbonnier, T. Dofferhoff, W. M. van Everdingen, P. K. Gerke, B. Geurts, H. A. Gietema, M. Groeneveld, L. van Harten, N. Hendrix, W. Hendrix, H. J. Huisman, I. Isgum, C. Jacobs, R. Kluge, M. Kok, J. Krdzalic, B. Lassen-Schmidt, K. van Leeuwen, J. Meakin, M. Overkamp, T. van Rees Vellinga, E. M. van Rikxoort, R. Samperna, C. Schaefer-Prokop, S. Schalekamp, E. T. Scholten, C. Sital, L. Stöger, J. Teuwen, K. Vaidhya Venkadesh, C. de Vente, M. Vermaat, W. Xie, B. de Wilde, M. Prokop, and B. van Ginneken, "Automated assessment of COVID-19 reporting and data system and chest CT severity scores in patients suspected of having COVID-19 using artificial intelligence," *Radiology*, vol. 298, no. 1, pp. E18–E28, 2021.

[20] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3D U-Net: Learning dense volumetric segmentation from sparse annotation," in *Medical Image Computing and Computer-Assisted Intervention*, 2016, pp. 424–432.

[21] W. Xie, C. Jacobs, J.-P. Charbonnier, and B. van Ginneken, "Relational modeling for robust and efficient pulmonary lobe segmentation in CT scans," *IEEE Transactions on Medical Imaging*, vol. 39, pp. 2664–2675, 2020.

[22] N. Otsu, "A threshold selection method from gray level histograms," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 9, no. 1, pp. 62–66, 1979.

[23] A. F. Frangi, W. J. Niessen, K. L. Vincken, and M. A. Viergever, "Multiscale vessel enhancement filtering," in *Medical Image Computing and Computer-Assisted Intervention*, ser. Lecture Notes in Computer Science, vol. 1496, 1998, pp. 130–137.

[24] Y. Wei, X. Liang, Y. Chen, X. Shen, M.-M. Cheng, J. Feng, Y. Zhao, and S. Yan, "STC: A simple to complex framework for weakly-supervised semantic segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 11, pp. 2314–2320, 2017.
[25] S. Reed, H. Lee, D. Anguelov, C. Szegedy, D. Erhan, and A. Rabinovich, "Training deep neural networks on noisy labels with bootstrapping," *arXiv:1412.6596*, 2014.
[26] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems*, 2019, pp. 8024–8035.
[27] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *International Conference on Computer Vision*, 2015, pp. 1026–1034.
[28] P. Krähenbühl and V. Koltun, "Efficient inference in fully connected crfs with gaussian edge potentials," in *Advances in Neural Information Processing Systems*, 2011, pp. 109–117.
[29] F. Isensee, P. F. Jaeger, S. A. A. Kohl, J. Petersen, and K. H. Maier-Hein, "nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation," *Nature methods*, vol. 18, no. 2, pp. 203–211, 2020.
[30] H. L. Kundel and M. Polansky, "Measurement of observer agreement," *Radiology*, vol. 228, no. 2, pp. 303–308, 2003.

Weiyi Xie studied signal processing during his M.S. degree in Tampere University of Technology, Finland. Since 2018, he is a Ph.D. candidate at the Diagnostic Image Analysis Group, Radboud university medical center, Radboud Institute for Health Sciences, Department of Radiology and Nuclear Medicine, Nijmegen, the Netherlands. His current research focuses on deep learning techniques for medical image analysis on CT imaging for chronic obstructive pulmonary diseases (COPD).

Colin Jacobs received his B.S. and M.S. degrees in biomedical engineering at the Eindhoven University of Technology, the Netherlands, in 2008 and 2010, respectively. He obtained his Ph.D. at the Diagnostic Image Analysis Group, Radboud university medical center, Radboud Institute for Health Sciences, Department of Radiology and Nuclear Medicine, Nijmegen. His Ph.D. research focused on the automatic detection and characterization of pulmonary nodules in thoracic CT scans. Since 2017, he is Assistant Professor at the Diagnostic Image Analysis Group, Radboud university medical center, Radboud Institute for Health Sciences, Department of Radiology and Nuclear Medicine, Nijmegen, the Netherlands. There, he leads the research line on lung cancer image analysis. From 2010 to 2013 he worked as a Biomedical Engineer for Fraunhofer MEVIS, Germany.

Bram van Ginneken studied physics at the Eindhoven University of Technology, the Netherlands and received his M.S. in physics at the Utrecht University, the Netherlands, in 1995. In 2001 he obtained his Ph.D. at the Utrecht University, the Netherlands.

Since 2012, he is full professor of Medical Image Analysis at Radboud University Medical Center and chairs the Diagnostic Image Analysis Group. He also works for Fraunhofer MEVIS in Bremen, Germany, and is a founder of Thirona, a company that develops software and provides services for medical image analysis. He is member of the Editorial Board of Medical Image Analysis. He pioneered the concept of challenges in medical image analysis.