Label-Free Segmentation of COVID-19 Lesions in Lung CT

Qingsong Yao, *Student Member, IEEE*, Li Xiao, *Member, IEEE*, Peihang Liu
and S. Kevin Zhou, *Fellow, IEEE*

***Abstract*—Scarcity of annotated images hampers the building of automated solution for reliable COVID-19 diagnosis and evaluation from CT. To alleviate the burden of data annotation, we herein present a label-free approach for segmenting COVID-19 lesions in CT via pixel-level anomaly modeling that mines out the relevant knowledge from normal CT lung scans. Our modeling is inspired by the observation that the parts of tracheae and vessels, which lay in the high-intensity range where lesions belong to, exhibit strong patterns. To facilitate the learning of such patterns at a pixel level, we synthesize 'lesions' using a set of surprisingly simple operations and insert the synthesized 'lesions' into normal CT lung scans to form training pairs, from which we learn a normalcy-converting network (NormNet) that turns an 'abnormal' image back to normal. Our experiments on two different datasets validate the effectiveness of NormNet, which conspicuously outperforms a variety of unsupervised anomaly detection (UAD) methods.**

***Index Terms*—COVID-19, label-free lesion segmentation, pixel-level anomaly modeling**

I. Introduction

THE world has been facing a global pandemic caused by a novel Coronavirus Disease (COVID-19) since December 2019 [1], [2]. According to the report from World Health Organization, COVID-19 has infected over 20 millions people including more than half a million deaths up to August 10 [3]. In clinics, real-time reverse-transcriptionpolymerase-chainreaction (RT-PCR) is the golden standard to make a definite diagnosis of COVID-19 infection [4]. However, due to the high false-negative rate [5], [6] and the shortage of equipment of RT-PCR, the radiological imaging techniques, e.g., x-ray and computed tomography (CT) still play a key role in COVID-19 diagnosis and evaluation [2], [7].

Compared to x-ray, CT screening is proved more effective due to its high spatial resolution and the unique relationship between CT density and lung air content [8]–[11]. For COVID-19 evaluation, segmentation of the infection lesions from CT scans is crucial for quantitative measurement and follow-up assessment [12], [13]. As it is time-consuming for experts to go through the 3D volumes slice by slice, automatic segmentation is highly desirable in clinical practice [2], [14]. Recently, deep learning based methods have been proposed for COVID-19 lesion screening [2] and some of them are proved successful for COVID-19 segmentation [12]–[14].

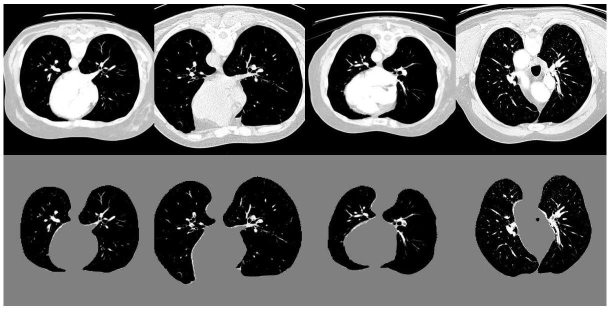

Fig. 1. Normal lung CT image (top) and its corresponding thorax area (bottom), clipped with an HU range of $[\mathbf{-800, 100}]$ and scaled to $[\mathbf{0, 1}]$.

Despite such success, they all rely on large-scale well-labeled datasets. However, obtaining such datasets is very difficult due to two related concerns. On the one hand, labeling a 3D CT volume is costly and time-consuming. Often it needs experienced radiologists, who are busy fighting the COVID-19 pandemic and hence lack time for lesion labeling. On the other hand, the COVID-19 lesions not only have a variety of complex appearances such as Ground-Glass Opacity (GGO), reticulation, and consolidation [15], but also have high variations in texture, size, and position. Those diversities raise a greater demand for rich annotated datasets. Accordingly, large-scale well-labeled COVID-19 datasets are scarce, which limits the use of Artificial Intelligent (AI) to help fight against COVID-19. As reported in Table I, most of the public COVID-19 datasets focus on diagnosis which only have classification information, while only a few of them provide semantic segmentation labels. While research attempts [16]–[18] have been made to address the challenges, these works, nevertheless, still need annotated images for training purpose. In this paper, we present a **label-free approach**, requiring no lesion annotation.

Although it is very difficult to build a large well-labeled COVID-19 dataset, collecting a large-scale normal CT volume dataset is much easier. It is also interesting to notice that the patterns of normal lungs are regular and easy to be modeled. The thorax of a normal person consists of large areas of air and a few tissues (such as tracheae and vessels [8]), which can be clearly distinguished by CT intensity [8]. As shown in Fig. 1, the air region is usually displayed as black background, with

Yao, Xiao and Zhou are with Institute of Computing Technology, Chinese Academy of Sciences. Zhou is corresponding author. Emails: yaoqingsong19@mails.ucas.edu.cn; xiaoli, zhoushaohua@ict.ac.cn.

Liu is with Beijing University of Posts and Telecommunications. Email: phliu@bupt.edu.cn.

TABLE I

A SUMMARY OF PUBLIC COVID-19 DATASETS. THE QUANTITY IS SPECIFIC TO THE CASES OF COVID-19.

Dataset	Modality	Quantity	Task
Chestxray [19]	X-rays	434	Diagnosis
COVID-CT [20]	CT image	342	Diagnosis
Patients Lungs [21]	X-rays	70	Diagnosis
Radiography [22]	X-rays	219	Diagnosis
SIRM-COVID [23]	2D CT image	340	Diagnosis
POCOVID-Net [24]	Ultrasound	37	Diagnosis
SIRM-Seg [23], [25]	CT image	110	Segmentation
Radiopedia [25], [26]	CT volume	9	Segmentation
Coronacase [27], [28]	CT volume	20	Segmentation
Mosmed [29]	CT volume	50	Diagnosis
BIMCV [30]	X-rays	10	Segmentation
BIMCV [30]	CT / X-rays	5381	Diagnosis

its Hounsfield unit (HU) value around -1000 [8]. Meanwhile, the tissue (with its HU > -500 [8]) has its intensity values similar to those of lesions, but it exhibits a regular pattern, which makes it amenable for modeling say by a deep network. This fact motivates us to formulate lesion segmentation as a **pixel-level anomaly modeling** problem. We hypothesize that if all the normal signals are captured at a pixel level, then the remaining abnormal pixels are localized automatically, which are grouped together as lesions.

To facilitate pixel-level anomaly modeling, we propose to synthesize ‘lesions’ and insert them into normal CT images, forming pairs of normal and ‘abnormal’ images for training. Surprisingly, such ‘lesion’ synthesis procedure constitutes a few simple operations, such as random shape generation, random noise generation within the shape and traditional filtering. Using these training pairs, we train a deep image-to-image network such as 3D U-Net [31] that converts an ‘abnormal’ image into normal. We call our network as a normalcy-converting network (NormNet). The NormNet essentially learns a decision boundary between normal tissues (particularly the tissues in a high intensity range) and synthetic ‘lesions’. We validate the effectiveness of NormNet on two different datasets. Empirically, it clearly outperforms various competing label-free approaches and its performances are even comparable to those of supervised method by some metrics.

It should be noted that our approach differs from a research line called unsupervised anomaly detection (UAD) [32]–[36], which aims to detect the out-of-distribution (OOD) data by memorizing and integrating anomaly-free training data and has been successfully applied in many image-level holistic classification scenarios. However, when applying the UAD methods for pixel-level image segmentation, their performances are rather limited [35], which we will confirm in our experiments. Further, our method differs from those methods in the inpainting [37] task, whose images in both training and testing sets are contaminated by the masks (noises) from the same domain. Finally, our method is different from synthetic data augmentation [38], which manually generates images according to the labeled lesion area. In contrast, we do not need any image with labeled COVID-19 lesions.

In summary, we make the following contributions:

- We propose the NormNet, a pixel-level anomaly modeling network, to distinguish the COVID-19 lesion from healthy tissues in the thorax area. This training procedure only needs a large-scale healthy CT lung dataset, without any labeled COVID-19 CT volume.
- We design an effective strategy for generating synthetic ‘lesions’ using surprisingly simple operations such as random shape, noise generation, and image filtering.
- The experiments show that our NormNet achieves better performances than various competing label-free methods on two different COVID-19 datasets.

II. RELATED WORK

A. COVID-19 screening and segmentation for chest CT

Deep learning based methods for chest CT greatly help COVID-19 diagnosis and evaluation [2], [7]. Wang et al. [39] propose a weakly-supervised framework for COVID-19 classification at the beginning of the pandemic, which achieves high performance. Wang et al. [40] exploit prior-attention residual learning for more discriminative COVID-19 diagnosis. Ouyang et al. [41] solve the imbalanced problem of COVID-19 diagnosis by a dual-sampling attention network. However, it is more difficult for the COVID-19 segmentation task due to the lack of well-labeled data [18], lesion diversities [15] and noisy labels [17]. Researchers have made attempts to address the above challenges. For example, to tackle the problem of labeled data scarcity, Ma et al. [28] annotate 20 CT volumes from coronacases [27] and radiopedia [26]. Fan et al. [18] propose a semi-supervised framework called Inf-Net. Zhou et al. [16] solve the same issue by fitting the dynamic change of real patients data measured at different time points. However, all of these models depend on data with semantic labels. In this work, we propose an unsupervised anomaly modeling method called NormNet, which achieves comparable performances, but with no need of labeled data.

B. Anomaly detection

Anomaly detection or outlier detection is a lasting yet active research area in machine learning [42], which is a key technique to overcome the data bottleneck [43]. A natural choice for handling this problem is one-class classification methods, such as OC-SVM [44], SVDD [45], Deep SVDD [46] and 1-NN. These methods detect anomaly by clustering a discriminate hyper-lane surrounding the normal samples in the embedding space.

However, these methods can only detect anomaly in image-level. In medical imaging analysis, it is also important to find the abnormal area [43], [47]. Recently, CNN-based generative models such as Generative Adversarial Networks (GAN) [48], and Variational Auto-encoders (VAE) [49] are proved essential for unsupervised anomaly segmentation [50]. These methods first capture the normal distribution by learning a mapping between the normal data and a low-dimensional latent space by reconstruction loss. They assume that if this process is only trained with normal distributions, a lesion area with abnormal shape and context can not be correctly mapped and reconstructed, resulting in high reconstruction error, which helps to localize the lesion area. The f-AnoGAN method [50],

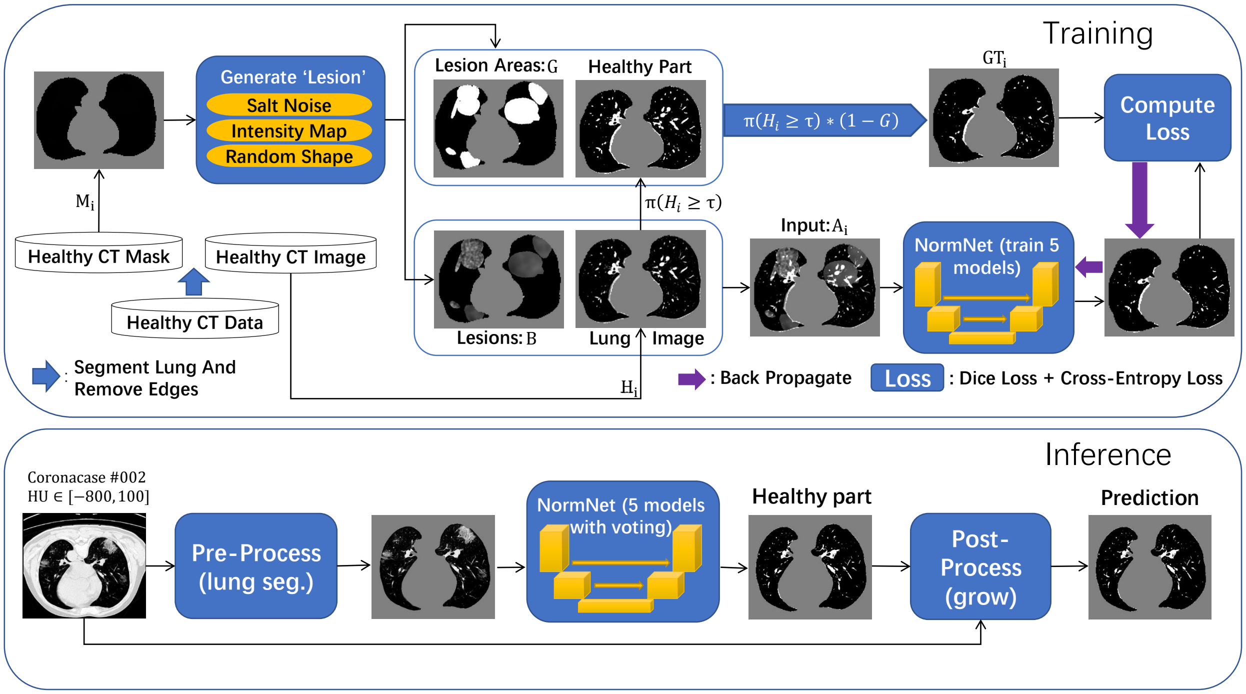

Fig. 2. The overall framework of proposed NormNet.

[51] learns the projection by solving an optimization problem, while VAE [49] tackles the same problem by penalizing the evidence lower bound (ELBO). Several extensions such as context encoder [52], constrained VAE [53], adversarial autoencoder [53], GMVAE [54], Bayesian VAE [55] and anoVAEGAN [56] improve the accuracy of the projection. Based on the pretrained projection, You et al. [54] restore the lesion area by involving an optimization on the latent manifold, while Zimmerer et al. [43] locate the anomaly with a term derived from the Kullback-Leibler (KL)-divergence.

Despite the success of these methods for the classification tasks [57], [58], their segmentation performances are insufficient [35]. The assumptions used by those reconstruction-based methods are shown to be problematic [34], [59]. Firstly, the calibrated likelihoods of the decoder may not be precise enough [60]. The out-of-distribution data have some possibilities to be successfully reconstructed [61], which raises false-negatives. Furthermore, the reconstruction is far from perfect [60], [62]. The decoder can not reconstruct all of the details of normal data precisely, which may cause false positives. As a result, these anomaly segmentation methods have limited segmentation performance, as indicated in the brain tumor segmentation task [35]. Moreover, specifically in lung CT, as some of the tissues are very small and appear irregularly, their information is easily lost during the down-sampling process of the encoder [63], which causes more segmentation errors.

The design of our method is to alleviate such issues. Firstly, we choose a 3D U-Net [31] as our encoder-decoder structure, and use the skip connection of U-Net to alleviate the loss of information. Next, to avoid inaccurate modeling, we generate a segmentation map from the original healthy CT and compute the loss based on it directly. At last, to encourage our NormNet to learn a decision boundary for healthy signals, we use synthetic lesions as anomalies.

III. Method

In this section, we firstly introduce the overall framework of our NormNet. Then we illustrate how to generate diverse 'lesions' in the given lung mask. Finally, we clarify how to post-process the lesion results predicted by our NormNet to obtain the final lesion mask for an unseen test image.

A. Overall framework

Let $\{R_1, R_2, \cdots, R_T\}$ be a set of T healthy lung CT images. We clip the raw image R_i with an HU range of $[-800, 100]$ and scale the clipped image to $[0, 1]$, obtaining R_i'. As shown in Fig. 2, our methods firstly use CNN-based lung segmentation method to obtain the lung masks $\{M_1', M_2', \cdots, M_T'\}$ and the thorax areas $\{H_1', H_2', \cdots, H_T'\}$ with $H_i' = R_i' \odot M_i'$, where $\odot$ stands for pixel-wise multiplication. It is worth noting that because no segmentation model can achieve 100% accuracy, and there are always some edges caused by segmentation errors left in the thorax area H_i', we introduce a simple pre-processing step (in Section III-B) to remove erroneous edges and generate a new lung mask M_i. Finally the thorax areas are updated to $\{H_1, H_2, \cdots, H_T\}$ with $H_i = H_i' \odot M_i$.

Then we use the synthetic 'lesion' generator described in Section III-C to synthesize various 'lesions' B within the lung masks M_i with diverse shapes G and textures, and inject them into the thorax area H_i to form the input A_i. Because the tissue patterns in the high-intensity range (say HU$\geq T$ with the threshold $T = -500$) in normal images are rather distinguishable from that of lesions, we concentrate on processing within this range and compute ground truth as

$$GT_i = \pi(H_i \geq \tau) \odot (1 - G), \tag{1}$$

where $\pi(.)$ is an indicator function that produces a binary mask. Note that the value of τ in H_i is equivalent to the HU threshold; for example, $T = -500$ means $\tau = 0.33$. Our NormNet is learned to predict the healthy part from A_i via encouraging it to be close to GT_i (aka minimizing Dice loss and cross-entropy loss). In this procedure, our NormNet learns to capture the context of healthy tissues quickly and precisely.

When our NormNet is applied to an unseen COVID-19 CT volume, it recognizes the healthy part of the volume with a high confidence and the lesion part of the volume with a low confidence. The confidence scores thus can be used as a decision boundary to predict the healthy parts and lesions. Because our training process is random, we learn the 5 models under the same setting to form an ensemble. A majority-vote for healthy parts is conducted as the final prediction. As our method is trained by the ground truth whose HU$\geq T$, a small number of lesion pixels whose HU$< T$ are not taken into consideration and might get missed. So, we grow the localized lesion areas to bring them back, following the post-processing step in Section III-D.

B. Removing erroneous edges

As mentioned above, this step is to separate the wrong edges caused by segmentation errors from lung mask M_i'. For a pair of inputs $\{M_i', H_i\}$, we select all the connected areas in thorax area H_i with most of the pixels lying on the edges of the lung segmentation mask M_i', and mark them as the wrong edges E_i. To avoid injecting noise into those edges, we use the lung mask without those edges, formulated as $M_i = M_i' - E_i$. Note that we only launch this process in the training phase, leveraging the fact that no lesion occurs inside a healthy volume.

C. Synthetic 'lesion' generator

As shown in Fig. 3, the generator constitutes a set of simple operations, following the two steps: (i) generating lesion-like shapes and (ii) generating lesion-like textures. Below, we elaborate each step.

1) Generating lesion-like shapes: Multiple COVID-19 lesions may exist in a CT scan and they have various shapes. To obtain multiple lesion-like shapes with a CT, we propose the following pipeline. Below, $U[a, b]$ denotes a uniform probability within the range $[a, b]$.

- For each lung mask M_i with a shape of size $[32, 512, 512]$, compute a factor $\lambda = \frac{M_i}{max_j M_j}$ as the fraction of the lung mask M_i comparing to the one with maximum volume. This factor controls the number of ellipsoids being generated with a larger λ likely yielding more ellipsoids.
- Create several ellipsoids as follows: (1) Sample a number $N_s \sim U[5\lambda, 10\lambda]$ and then generate N_s small-size ellipsoids with the radius of each ellipsoid randomly selected from $U[3, 10]$; (2) Sample a number $N_m \sim U[5\lambda, 10\lambda]$ and then generate N_m medium-size ellipsoids with the radius of each ellipsoid randomly selected from $U[10, 32]$; and (3) Generate a large size ellipsoid with a probability of 0.2λ and with its radius $\sim U[32, 64]$.
- For each generated ellipsoid, deform it using elastic transformation [64] with random parameters and rotate it to align with random axes, yielding a blob C. Then position this blob at a random center inside the lung H_i.

At this stage, we have a set of blobs $\{C_1, C_2, \ldots\}$. Then we merge connected blobs and obtain several non-adjacent blobs $\{G_1, G_2, \ldots\}$ with varying shapes. For each blob G_j, we synthesize a patch of lesion B_j by the following steps.

2) Generating lesion-like textures: The texture pattern of lesions varies; thus it is challenging to generate lesion-like textures. Below we outline our attempt of doing so using a set of simple operations. It should be noted that our method is far from prefect; nevertheless, we find it is empirically effective.

We follow a series of three steps, namely noise generation, filtering, and scaling/clipping operations, to generate the lesion-like textures.

- Noise generation. For each pixel denoted by x, generate salt noise $b_1(x)$

$$b_1(x) = \begin{cases} 1 & \text{with a probability } \ a(x); \\ 0 & \text{with a probability } 1 - a(x), \end{cases} \tag{2}$$

where the pixel-dependent probability function $a(x)$ will be defined later.
- Filtering. Filter the noise image $b_1(x)$ to obtain $b_2(x)$ using a Gaussian filter g with a standard deviation σ_b.

$$b_2(x) = g(x; \sigma_b) \otimes b_1(x), \tag{3}$$

where $\otimes$ is the standard image filtering operator. The standard deviation σ_b is randomly sampled as follows:

$$\sigma_b \sim \begin{cases} U[0.8, 2] & \text{with a probability of } 0.7; \\ U[\ 2, 5] & \text{with a probability of } 0.3. \end{cases} \tag{4}$$

.
- Scaling and clipping. This yields the lesion-like pattern $B_j(x)$.

$$B_j(x) = clip_{[0,1]}(\beta b_2(x)), \tag{5}$$

with β being the scaling factor that is obtained by

$$\beta = \mu_0 / mean_{0.2}(b_2(x)), \tag{6}$$

where $\mu_0 \sim U[0.4, 0.8]$ and $mean_t(f(x))$ is the mean intensity of the image $f(x)$ that passes the threshold t.

Now, we describe how to obtain the pixel-dependent probability function $a(x)$, again using a series of noise generation, filtering, and scaling operations.

- Noise generation. For each pixel x, independently sample the uniform probability $U[0, 1]$ to get a noise image $a_1(x) \sim U[0, 1]$.
- Filtering. Filter the noise image $a_1(x)$ to obtain $a_2(x)$ using a Gaussian filter g with a standard deviation σ_a.

$$a_2(x) = g(x; \sigma_a) \otimes a_1(x), \tag{7}$$

where the standard deviation $\sigma_a \sim U[2, 20]$.

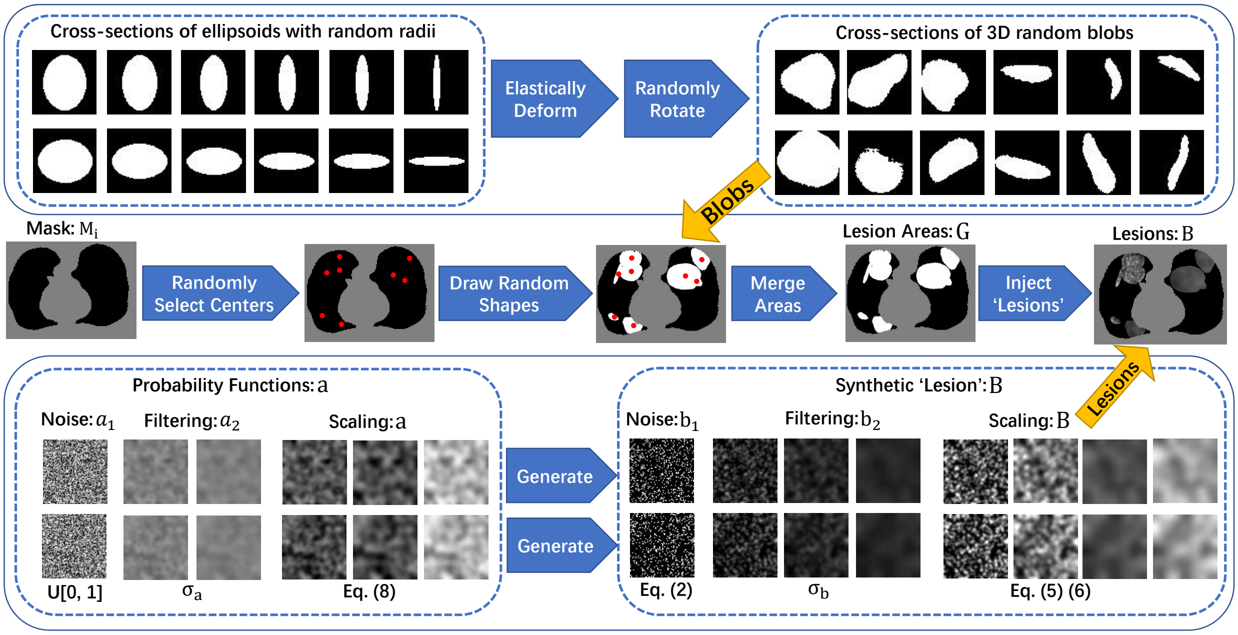

Fig. 3. The schematic diagram of the proposed noise generator. We generate several diverse shapes and fill the connected areas with various filtered and scaled salt noises.

- Scaling. This yields the desired function $a(x)$.

$$\begin{aligned} a(x) &= scale_{[a_L,a_U]}(a_2(x)) \\ &= (a_U - a_L) * \frac{a_2(x) - a_{2,min}}{a_{2,max} - a_{2,min}} + a_L, \end{aligned} \quad (8)$$

where $a_U \sim U[0, 0.3]$, $a_L \sim U[0, 0.3]$ and $a_U - a_L > 0.15$.

Finally, we inject the synthetic lesions B_j into the various blobs G_j, and place these blobs at random centers inside the lung area H_i. Mathematically, the image A_i with synthetic 'lesions' is generated by finding the maximum value of the lung area H_i and the synthetic lesions B_j at each pixel point:

$$A_i = \max(H_i, B_1, B_2, \cdots). \quad (9)$$

Our goal is to learn a network that takes A_i as input and outputs GT_i.

D. Post processing

A post processing procedure is designed to obtain the final lesion prediction based on difference between the original CT volume and predicted healthy areas. As illustrated in Fig. 4, the final prediction is obtained with the following steps:

- Compute the lung mask (Fig. 4(b)) and predict the healthy part by NormNet (Fig. 4(c));
- Compute the lesion region by subtracting the predicted healthy part from lung mask to get Fig. 4(d). Considering that only bright pixels $\geq \tau$ are in the lung mask, the full-pixel raw lesion areas (Fig. 4(f)) is calculated, aiming to 'recover' less bright lesions;
- Mean filtering F with kernel size k is then applied to Figs. 4(d) and 4(f) to smooth the lesion region and then remove the background noise via thresholding, which obtains the results in Fig. 4(e) and 4(g), respectively;

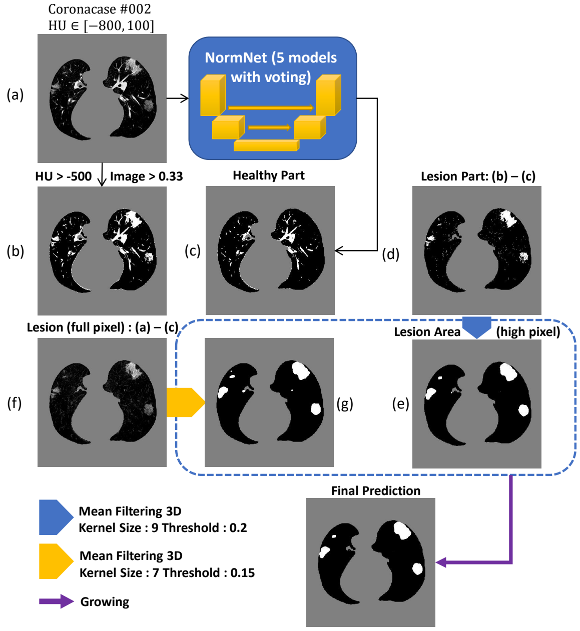

Fig. 4. The illustration of the post-processing process.This step removes the healthy part from the COVID-19 CT volume and generate final prediction by mean filtering and growing.

- A region growing algorithm is applied to obtain the final predicted regions, which firstly expands the lesion regions of Fig. 4(f), and then removes the pixels out of the full pixel lesion regions defined by Fig. 4(g).

IV. Experiments

Below we firstly provide a brief description of the various CT lung datasets used in our experiments. Then we present our experimental settings and the baseline approaches we implement and compare. Finally, we show our main experimental results and an ablation study.

A. Datasets

One distinguishing feature of the paper lies in unleashing the power embedded in existing datasets. Rather than using a single dataset, we seamlessly integrate multiple CT lung datasets for two different tasks of healthy lung modeling, COVID-19 lesion segmentation, and general-purpose lung segmentation into one working solution.

1) CT datasets for healthy lung modeling: LUNA16 [65] is a grand-challenge on lung nodule analysis. The images are collected from The Lung Image Database Consortium image collection (LIDC-IDRI) [66], [67], [69], and each image is labeled by 4 experienced radiologists. As half of the images are healthy and clean except for those contain nodule areas, we select 453 CT volumes from LUNA16 and remove the slices with nodules to formulate our healthy lung CT dataset.

2) CT datasets for COVID-19 lesion segmentation : To measure the performance of our methods towards COVID-19 segmentation, we choose two public COVID-19 CT segmentation datasets in the Table I with semantic labels. It is worth noting that our method segments the COVID-19 lesions under the unsupervised setting, and thus the labeled datasets are only used for testing. All of the CT slices have been resized to 512×512.

- *Coronacases:* There are 10 public CT volumes in the [27] uploaded from the patients diagnosed with COVID-19. These volumes are firstly delineated by junior annotators[1], and then refined by two radiologists with 5 years experience, and finally, all the annotations are verified and refined by a senior radiologist with more than 10 years experience in chest radiology diagnosis [28].
- *Radiopedia:* Another 8 axial volumetric CTs are released from Radiopaedia [26] and have been evaluated by a radiologist as positive and segmented [25].

3) CT datasets for general purpose lung segmentation : To obtain the accurate lung area in the CT volume, we choose nnU-Net [68] as our lung segmentation method, which is proved to be state-of-the-art segmentation framework in medical imaging analysis. We use two lung CT datasets with semantic labels for the lung region:

- *NSCLC left and right lung segmentation:* This dataset consists of lung volume segmentation collected on 402 CT scans from The Cancer Imaging Archive NSCLC Radiomics [69]–[71].
- *StructSeg lung organ segmentation:* This dataset consists of 50 lung cancer patient CT scans with lung organ segmentation. The dataset served as a segmentation challenge during MICCAI 2019 [72].
- *MSD Lung tumor segmentation* This dataset consists of 63 labelled CT scans, which served as a segmentation challenge during MICCAI 2018 [73]. The lung regions are labeled by Ma et al. [28].

We choose 2D U-Net as the backbone. The model is trained by nnU-Net [68] in 5-fold cross-validation, which segments the lung region very precisely with Dice scores larger than 0.98 in both Coronacases and Radiopedia datasets.

B. Experimental settings

1) Evaluation metrics: We use several metrics widely used to measure the performance of segmentation models in medical imaging analysis, including precision score (PSC), sensitivity (SEN) and Dice coefficient (DSC), which are formulated as follows:

$$PSC = \frac{tp}{tp+fp}; SEN = \frac{tp}{tp+fn}; DSC = \frac{2tp}{2tp+fn+fp},$$

where tp, fp and fn refer to the true positive, false positive and false negative respectively.

2) Pre-processing: All of the images in the training and testing sets are segmented for the lung region at first. Then we unify their spacing to $0.8 \times 0.8 \times 1mm^3$, as well as orientation. Next, all of the images are clipped with window range $[-800, 100]$ and normalized to $[0, 1]$. Finally, the lung regions are centralized and padded to 512×512 with 0.

3) Training and inference details: We choose 3D U-Net [31] as backbone for NormNet, implemented by MONAI[2]. As all of the volumes in both training and testing phases are well aligned, no more augmentation is needed. The NormNet is trained on a TITAN RTX GPU and optimized by the Adam optimizer with default settings. We train our network for 3500 iterations with a batch size of 8, and set the learning rate to 3e-4. For the testing phase, as the contexts of healthy signals are precisely captured by our NormNet, these signals are predicted with high probability. Therefore, we select those pixels with probability > 0.95 as healthy parts in the COVID-19 CT volume. For the mean filtering in the post processing, we set kernel sizes f to 9, 7 and thresholds to 0.2, 0.15 for lesion parts with bright pixels (Fig. 4d) and full pixels (Fig. 4f), respectively. We obtain these values according to the hyperparameter search, which are fixed to all of two COVID-19 datasets.

C. Baselines

We compare our methods with existing deep learning based methods in medical imaging analysis for unsupervised anomaly detection (UAD) methods to evaluate the effectiveness of our approach. To eliminate the influence of irrelevant factors, we use the images with only lung regions as training and testing sets for all of the experiments (expect for **VAE Original**). These encoder-decoder based methods are trained with a learning rate of 3e-4 and a batch size of 16 for 6000 iterations. To obtain the best performance for each method, we perform a greedy search up to two decimals to get the threshold with best Dice score for each COVID-19 dataset.

- **AE:** An Autoencoder with a dense bottleneck $z \in \mathbb{R}^{128}$.
- **VAE [49]:** As the reconstruction is more difficult for lung CT images, so we set α for KL loss as 1e-6 to make the reconstruction easily.

[1] Ma et al. provide 20 well-labeled CT volumes, in addition to the 10 volumes of coronacases, the other 10 volumes have been clipped to [0 – 255] without any information about HU, which is not applicable based on our methods.

[2] https://github.com/Project-MONAI/MONAI

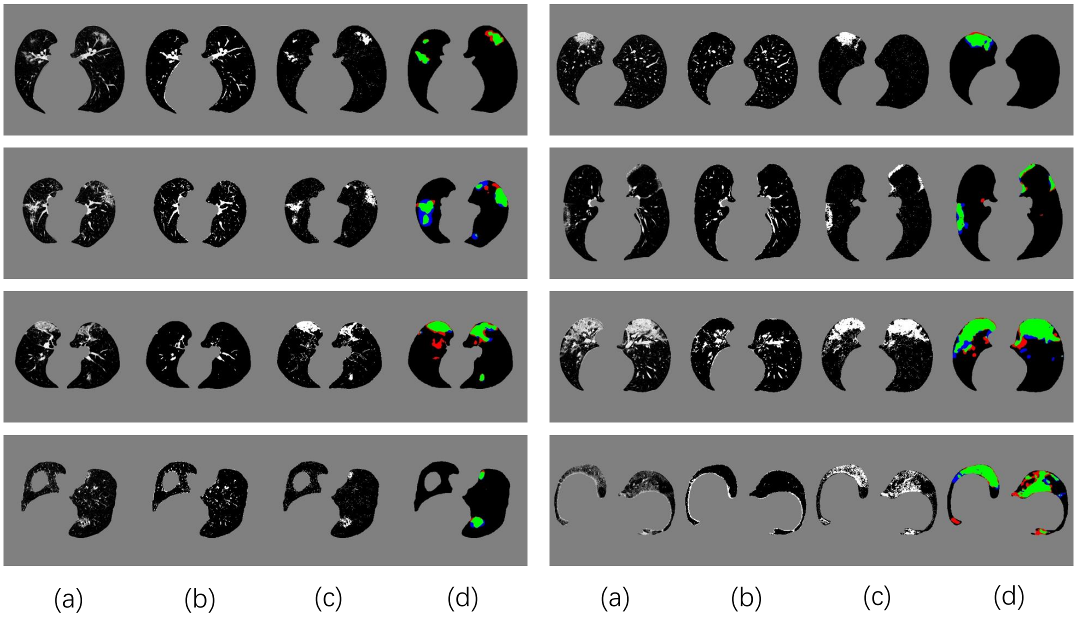

Fig. 5. Visual results of our NormNet for COVID-19 segmentation. (a), (b), (c) and (d) represents input, healthy tissues (predicted from our NormNet), lesion parts, and final segmentation, respectively. The green, blue, and red areas in (d) refer to true positive, false negative, and false positive, respectively.

- **VAE Spatial [56]:** A Variational Autoencoder with a spatial bottleneck $z \in \mathbb{R}^{8\times 8\times 128}$.
- **VAE Original:** A Variational Autoencoder trained with the raw lung CT images without lung segmentation.
- **Context VAE [52]:** Force the encoder to capture more information by inpainting cropped input. We set crop size to 32.
- **Constrained VAE [53]:** Map the reconstructed image to the same point as the input in the latent space.
- **GMVAE [54]:** Replace the mono-modal prior of the VAE with a Gaussian mixture [35].
- **Bayesian VAE [55]:** Aggregate a consensus reconstruction by Monte-Carlo dropout. The dropout rate is 0.2.
- **KL Grad [43]:** Use the gradient map for KL loss to segment anomalies.
- **VAE restoration [54]:** Restore the abnormal input to decrease the evidence lower bound (ELBO). The restoration part is marked as the detected abnormal area.
- **f-AnoGAN [50]:** To keep the training process of f-Anogan stable, we resize the lung image to $[64, 64]$ after center crop.

In order to reveal the top-line for each dataset, we train **nnU-Net** [68] in 5-fold cross-validation. Furthermore, to test the performance of the supervised model when inferring unseen datasets, we train nnU-Net on two COVID-19 datasets and test on the left one, called **nnU-Net-Unseen**.

D. Segmentation results

Our NormNet firstly votes for the healthy tissues from the CT volumes with COVID-19 lesions. To test the performance of our NormNet, we collect all bright pixels with $\tau \geq 0.33$ of the CT volumes. As in Table II, our method successfully distinguishes the COVID-19 lesion parts and healthy parts with AUC larger than 85%. When we choose the prediction threshold as 0.95, the high specificity ensures that most of the lesions are treated as anomaly. Then, the post-processing procedure grows the lesion area to contain more lesions with less bright pixels ($\tau < 0.33$). We also use mean filtering in the post-processing to remove the isolated healthy pixels that are segmented as anomaly, as shown in Fig. 5c. Therefore, our method reaches the Dice scores of 68.7%, 59.4%[3] and 69.7% (shown in Table III) in the two different COVID-19 datasets respectively, which are significantly ahead of other unsupervised anomaly detection methods. The visual results shown in Fig. 5 reveal that most of the COVID-19 lesions are successfully (green area) segmented by our NormNet.

TABLE II
THE RESULTS OF SEGMENTATION PERFORMANCES OF BRIGHT PIXELS.

Dataset	Precision	Sensitivity	Specificity	AUC
Coronacases	90.6	78.8	80.2	87.8
Radiopedia	93.5	70.7	87.9	89.5

On the contrary, the other unsupervised anomaly detection methods have limited power to segment COVID-19 lesion. As shown in Fig. 6, due to the inaccurate reconstructions, the reconstruction-based methods such as VAE [49] and f-AnoGAN [50] can not reconstruct the tissues precisely. Such a reconstruction error greatly affects the segmentation performance of COVID-19 lesions. On the other hand, the encoder can not make sure to treat the COVID-19 lesion as anomaly, and suppress the lesion in the reconstruction results. Thus the KL-grad [43] and restoration [54] have less effect either. These two serious shortcomings result in low COVID-19

[3] we remove CT volume #6 from the Radiopedia dataset as it has only about 70 positive pixels in 42 slices.

TABLE III

THE QUANTITATIVE RESULTS OF OUR METHOD COMPARED TO OTHER UAD METHODS AND NNU-NET. FOR EACH COLUMN, THE **TOP**, SECOND AND THIRD VALUES ARE HIGHLIGHTED.

Methods	Coronacases DSC (%)	 PSC (%)	 SEN (%)	Radiopedia DSC (%)	 PSC (%)	 SEN (%)
nnU-Net [68]	80.1±6.73	80.2±12.4	82.3±9.30	76.7±5.81	77.1±14.0	80.5±13.11
nnU-Net-Unseen	77.1±10.2	81.1±11.0	75.9±15.9	73.9±9.45	66.9±13.4	85.3±9.85
AE	28.3±15.5	21.5±15.3	52.1±11.3	30.3±17.7	24.4±19.0	58.9±6.2
VAE [49]	26.4±14.5	19.8±14.0	50.1±9.8	28.1±17.5	21.6±17.6	62.3±5.7
VAE Spatial [56]	27.4±16.5	21.0±16.4	49.9±11.9	30.7±19.8	24.8±20.7	59.2±8.0
VAE Original	10.9±8.0	6.9±6.1	41.3±8.2	12.3±10.5	8.5±8.9	44.9±4.9
Context VAE [52]	29.7±16.0	21.8±15.6	61.0±9.8	32.3±21.3	24.3±20.6	72.2±6.0
Constrained VAE [53]	27.9±14.8	21.0±14.7	53.2±10.5	29.2±17.7	22.9±18.3	61.3±5.6
GMVAE [54]	25.7±16.4	20.2±14.4	51.0±12.6	28.6±17.7	22.3±19.5	**63.3±7.2**
Bayesian VAE [55]	27.5±15.0	20.8±14.7	50.9±11.4	29.6±16.8	23.5±17.6	58.2±6.8
KL Grad [43]	9.5±8.2	5.5±5.2	**65.5±19.7**	10.2±14.2	6.7±10.3	39.1±20.3
VAE Restoration [54]	12.8±4.5	16.3±10.1	12.1±2.5	9.1±3.7	16.5±16.0	8.8±1.6
f-AnoGAN [50]	15.4±12.6	10.8±10.8	38.3±13.2	19.7±17.3	14.2±14.9	55.2±8.9
Proposed w/o growing	65.5±17.9	**88.1±5.23**	56.2±21.7	54.2±17.5	**60.8±21.2**	51.3±16.9
Proposed	**68.7±15.8**	85.1±6.97	62.1±22.8	**59.4±17.4**	60.4±19.7	61.8±18.4

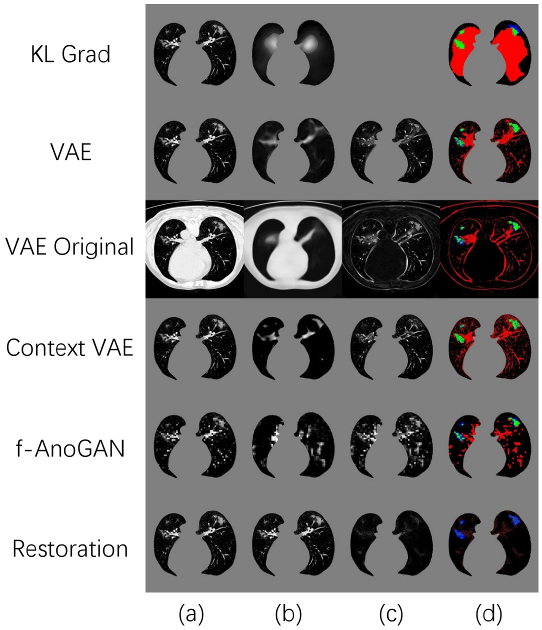

Fig. 6. Visual results of various UAD methods. (a), and (d) refer to input, and final results, respectively. The image (b) in the ”KL Grad” method means the gradient map of KL loss, while it in the other methods means reconstruction or restoration results. The image (c) of the methods expect of 'KL Grad' means difference map.

TABLE IV

THE DICE SCORES OF FIVE MODELS AND VOTING PERFORMANCE WITH DIFFERENT NUMBER OF ITERATIONS..

Iterations	2000	2500	3000	3500	4000	4500
voting	67.6	70.2	66.9	68.7	70.4	67.4
model$_1$	62.1	68.8	67.4	69.6	68.8	59.2
model$_2$	65.1	64.5	66.0	68.0	53.5	66.7
model$_3$	66.3	69.5	69.1	70.6	68.8	58.4
model$_4$	71.1	69.9	50.7	64.6	66.6	70.4
model$_5$	64.2	64.0	66.1	45.8	69.8	70.7

TABLE V

THE DICE SCORES OF CORONACASES OBTAINED BY THE NORMNET TRAINED WITH THE MODULES OF SHAPES, PROBABILITY MAPS AND SALT NOISES SWITCHED ON AND OFF.

Shapes	Probability Maps	Salt Noises	Dice
✓	✓	✓	68.7
×	✓	✓	51.5
✓	×	✓	62.0
✓	✓	×	25.5

segmentation performances, reported in Table III.

E. Ablation study

1) Voting: To explore the effects of randomness in the training process, we evaluate the performances of the five models and their voting results with different number of iterations. As shown in Table IV, the performances of the five models oscillate as the iteration increases, while the NormNet greatly alleviates this problem through the voting mechanism of 5 models.

2) Modules of synthetic 'lesion' generator: The steps of synthetic 'lesion' generator can be roughly divided into three parts: Generate shapes (G_j in Section III-C.1), probability maps (a_i in (8)), and salt noises (B_i in (6)). To investigate the influence of each part, we train a new NormNet without the corresponding diversity. To eliminate the diversity of shapes, we generate 5 ellipsoids with radius = 12 for any lung area H_i without any deformation. For probability maps, we set probability = 2. At last, we set $\sigma_b = 2$ and $\mu_0 = 150$ for synthetic salt noises with the same texture. As shown in Table V, the loss of diversities affects the accuracy of the decision boundary and the segmentation performance. Especially, the salt noises filtered and scaled by fixed parameters have limited contexts, which are easily learned by the NormNet, resulting in extremely inaccurate decision boundary. Thus, our various synthetic 'lesions' can force the NormNet to learn a decision boundary for healthy tissues, which can be further used to segment COVID-19 lesions.

TABLE VI
THE DICE SCORES OF CORONACASES UNDER DIFFERENT HU THRESHOLDS.

HU threshold T	-700	-600	-500	-400	-300
Dice	55.1	67.0	68.7	64.3	61.1

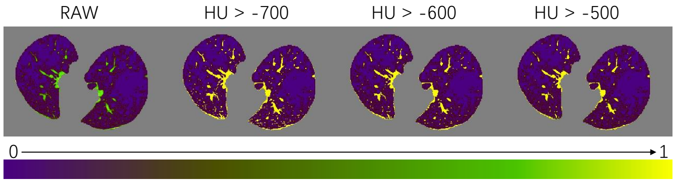

Fig. 7. The visualization of masks under different HU thresholds. Many noisy pixels with complex contexts occur when setting the threshold as $T = -700$. We use a colormap for better visualization of the nuances.

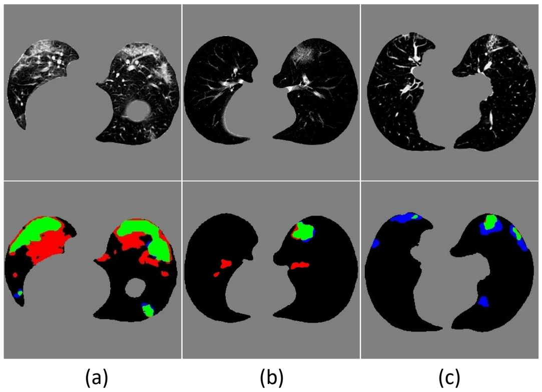

Fig. 8. Samples of failure predictions to show the limitation of our method. The red area means false positive while the blue area indicates false negative.

3) Hyparameter analysis: The threshold of HU is important in our method, since it filters the background noises while trying to keep the pattern complexity at a level that can be effectively managed by the network. On the one hand, if the threshold is too high, our NormNet only segments healthy pixels in a small-scale set, which causes more abnormal pixels missing. On the other hand, if the threshold is too small, some noisy pixels with complex contexts (as shown in Fig. 7) are segmented by the NormNet. It raises the difficulty and turns the NormNet to capture the features of synthetic 'lesions' instead of healthy tissues, as we can not make sure the contexts of synthetic 'lesions' are the same to COVID-19 lesions, the NormNet overfits the synthetic 'lesions' and can not segment COVID-19 successfully. As in Table VI, the performance drops rapidly when the HU threshold $T = -700$.

V. CONCLUSIONS AND DISCUSSIONS

In this paper, we propose the NormNet, a pixel-level anomaly modeling network to turn an 'abnormal' volume back to normal. A decision boundary for normal parts of the NormNet is learned by segmenting healthy tissues from the diverse synthetic 'lesions', which can be further used to segment COVID-19 lesions, without training on any labeled data. The experiments on two different COVID-19 datasets validate the effectiveness of the NormNet.

Despite the improvement compared to existing unsupervised anomaly detection methods, there is still a gap between our methods and supervised methods such as nnU-Net [68]. After exploring the failure predictions of our methods, we find that they are divided into three categories:

1) Some anomalies such as pulmonary fibrosis (the first row shown in Fig. 8) are treated as COVID-19 lesions.
2) Gaps between datasets: for example, most of the layer thicknesses in Luna16 dataset are around 1mm. However, in Radiopedia dataset slices are padded together, which generate different contexts. The unseen contexts are treated as anomalies by our NormNet, which results in the most of false-positives in Radiopedia dataset.
3) Our method is only sensitive to the pixels with values larger than τ. Although most of lesions can be successfully detected, a small part of lesions with pixels smaller than τ are still missed (as shown in the right column of Fig. 8). These small lesions also serve a difficult problem for both supervised methods [17] and anomaly detection.

In future, we plan to extend our method to address the above limitations and explore the possibility of applying the 'lesion' generator for segmentation in non-thoracic regions.

REFERENCES

[1] C. Wang, P. W. Horby, F. G. Hayden, and G. F. Gao, "A novel coronavirus outbreak of global health concern," *The Lancet*, vol. 395, no. 10223, pp. 470473, feb 2020.
[2] F. Shi *et al.*, "Review of Artificial Intelligence Techniques in Imaging Data Acquisition, Segmentation and Diagnosis for COVID-19," *IEEE Rev. Biomed. Eng.*, vol. 3333, no. c, pp. 113, 2020
[3] WHO, "Coronavirus disease (COVID-19) Situation Report 164", 2020. [Online]. Available:https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200702-covid-19-sitrep-164.pdf?sfvrsn=ac074f58_2
[4] Z. Y. Zu *et al.*, "Coronavirus disease 2019 (covid-19): A perspective from china, *Radiology*, p. 200490, 2020.
[5] Y. Fang *et al.*, "Sensitivity of chest CT for COVID-19: Comparison to RT-PCR," *Radiology*, p. 200432, 2020
[6] M.-Y. Ng *et al.*, "Imaging profile of the COVID19 infection: Radiologic findings and literature review," *Radiology: Cardiothoracic Imaging*, vol. 2, no. 1, p. e200034, 2020.
[7] G. D. Rubin *et al.*, "The role of chest imaging in patient management during the COVID-19 pandemic: A multinational consensus statement from the fleischner society," *Radiology*, p. 201365, apr 2020.
[8] B. A. Simon, G. E. Christensen, D. A. Low, and J. M. Reinhardt, "Computed Tomography Studies of Lung Mechanics," *Proc Am Thorac Soc*, vol. 2(6), pp. 507-517, Dec. 2005, DOI: 10.1513/pats.200507-076DS.
[9] S. K. Zhou, H Greenspan, D. Shen (Eds.), "Deep learning for medical image analysis," *Academic Press*, 2017.
[10] SK Zhou (Ed.), " Medical Image Recognition, Segmentation and Parsing: Machine Learning and Multiple Object Approaches,"*Academic Press*, 2015.
[11] Sofka et al. "Multi-stage learning for robust lung segmentation in challenging CT volumes," in *MICCAI*, 2011.
[12] L. Huang *et al.*, "Serial quantitative chest CT assessment of COVID-19: Deep-learning approach, *Radiol. Cardiothorac. Imaging*, vol. 2, p. e200075, 2020.
[13] F. Shan *et al.*, "Lung infection quantification of COVID-19 in CT images with deep learning," *arxiv:2003.04655*, 2020.
[14] Y. Cao *et al.*, "Longitudinal assessment of COVID-19 using a deep learningbased quantitative CT pipeline: Illustration of two cases," *Radiol. Cardiothorac. Imaging*, vol. 2, no. 2, p. e200082, 2020.
[15] M.-Y. Ng *et al.*, "Imaging profile of the COVID-19 infection: Radiologic findings and literature review," *Radiol. Cardiothorac. Imaging*, vol. 2, no. 1, p. e200034, 2020.

[16] L. Zhou *et al.,* "A Rapid, Accurate and Machine-agnostic Segmentation and Quantification Method for CT-based COVID-19 Diagnosis," in *IEEE Transactions on Medical Imaging*, DOI:10.1109/TMI.2020.3001810.

[17] G. Wang *et al.,* "A Noise-robust Framework for Automatic Segmentation of COVID-19 Pneumonia Lesions from CT Images," in *IEEE Transactions on Medical Imaging*, DOI:10.1109/TMI.2020.3000314.

[18] D. Fan *et al.*, "Inf-Net: Automatic COVID-19 Lung Infection Segmentation from CT Images," in *IEEE Transactions on Medical Imaging*, DOI: 10.1109/TMI.2020.2996645.

[19] J. P. Cohen, P. Morrison, and L. Dao, "COVID-19 image data collection," *arxiv:2003.11597* 2020.

[20] J. Zhao, Y. Zhang, X. He, and P. Xie, "COVID-CT-Dataset: a CT scan dataset about COVID-19," *arXiv:2003.13865* 2020.

[21] "COVID-19 Patients Lungs X Ray Images 10000," https://www.kaggle.com/nabeelsajid917/covid-19-x-ray-10000-images, Accessed: 2020-04-11.

[22] M. E. H. Chowdhury, T. Rahman et al., "Can AI help in screening Viral and COVID-19 pneumonia?" *arXiv:2003.13145*, 2020.

[23] "Italian Society of Medical and Interventional Radiology COVID-19 dataset", SIRM, https://www.sirm.org/category/senza-categoria/covid-19, Accessed: 2020-05-28.

[24] J. Born *et al.*, "POCOVID-Net: Automatic Detection of COVID-19 From a New Lung Ultrasound Imaging Dataset (POCUS)," *arxiv:2004.12084* 2020.

[25] "COVID-19 CT segmentation dataset," https://medicalsegmentation.com/covid19/, Accessed: 2020-04-11.

[26] "Radiopedia" https://radiopaedia.org/articles/covid-19-4.

[27] "Coronacases" https://coronacases.org/.

[28] J. Ma *et al.*, "Towards Efficient COVID-19 CT Annotation: A Benchmark for Lung and Infection Segmentation," *arxiv:2004.12537* 2020.

[29] "Mosmed" https://mosmed.ai/en/.

[30] M. I. Vay *et al.*, "BIMCV COVID-19+: a large annotated dataset of RX and CT images from COVID-19 patients," *arxiv:2006.01174* 2020.

[31] . iek, A. Abdulkadir, S.S. Lienkamp, T. Brox, O. Ronneberger, "3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation," In: *MICCAI*, 2016, Springer, pp. 424-432.

[32] C. Varun, B.Arindam, and K. Vipin, "Anomaly detection: A survey," *ACM computing surveys (CSUR)*, vol. 41, 2009, num. 3, pp. 1-58.

[33] T. Crispi *et al.*, "Anomaly Detection in Medical Image Analysis," *Handbook of Research on Advanced Techniques in Diagnostic Imaging and Biomedical Applications*, IGI Global, pp. 426-44.

[34] B. Zong *et al.*, "Deep autoencoding gaussian mixture model for unsupervised anomaly detection," in *ICLR*, 2018

[35] C. Baur, S. Denner, B. Wiestler, S. Albarqouni, and N. Navab, "Autoencoders for Unsupervised Anomaly Segmentation in Brain MR Images: A Comparative Study," *arxiv:2004.03271* 2020.

[36] M. Astarak, I. Toma-Dasu, . Smedby and C. Wang, "Normal Appearance Autoencoder for Lung Cancer Detection and Segmentation," in *MICCAI*, Springer, 2019, pp. 249-256.

[37] P. Deepak, K. Philipp, D. Jeff, D. Trevor and E. A. A, "Context Encoders: Feature Learning by Inpainting," in *CVPR*, 2016.

[38] M. Frid-Adar, E. Klang, M. Amitai, J. Goldberger and H. Greenspan, "nthetic data augmentation using GAN for improved liver lesion classification," in *ISBI*, 2018, Washington, DC, pp. 289-293.

[39] X. Wang *et al.*, "A Weakly-supervised Framework for COVID-19 Classification and Lesion Localization from Chest CT," in *IEEE Transactions on Medical Imaging*, DOI: 10.1109/TMI.2020.2995965.

[40] J. Wang *et al.*, "Prior-Attention Residual Learning for More Discriminative COVID-19 Screening in CT Images," in *IEEE Transactions on Medical Imaging*, DOI: 10.1109/TMI.2020.2994908.

[41] X. Ouyang *et al.*, "Dual-Sampling Attention Network for Diagnosis of COVID-19 from Community Acquired Pneumonia," in *IEEE Transactions on Medical Imaging*, DOI: 10.1109/TMI.2020.2995508.

[42] G. Pang, C. Shen, L. Cao and A. Hengel, "Deep Learning for Anomaly Detection: A Review," *arxiv:2007.02500* 2020.pang

[43] D. Zimmerer, F. Isensee, J. Petersen, S. Kohl, and K. Maier-Hein, Unsupervised anomaly localization using variational auto-encoders, in *MICCAI* Springer, 2019, pp. 289297.

[44] Y. Chen, X. S. Zhou, and T. S Huang, "One-class svm for learning in image retrieval," In *The IEEE International Conference on Image Processing*, volume 1, pages 3437. IEEE, 2001.

[45] D. M. J. Tax and R. P. W. Duin, "Support Vector Data Description," *Mach. Learn.*, 2004, vol. 54, pp. 4566. DOI:https://doi.org/10.1023/B:MACH.0000008084.60811.49

[46] L. Ruff *et al.* "Deep One-Class Classification," in *ICML*, 2018, PMLR, vol. 80, pp. 393-4402.

[47] P. Seebck et al., "Exploiting Epistemic Uncertainty of Anatomy Segmentation for Anomaly Detection in Retinal OCT," in IEEE Transactions on Medical Imaging, vol. 39, no. 1, pp. 87-98, Jan. 2020, DOI: 10.1109/TMI.2019.2919951.

[48] I. J. Goodfellow *et al.*, "Generative Adversarial Networks", in *NIPS* Curran Associates, Inc., 2014, pp. 26722680.

[49] D. P. Kingma , and M. Welling, "Auto-Encoding Variational Bayes," In *ICLR*, 2014.

[50] T. Schlegl, P. Seebock, S. M. Waldstein, G. Langs, and U. Schmidt Erfurth, "f-AnoGAN: Fast unsupervised anomaly detection with generative adversarial networks," in *Medical Imaging Analysis*, Volume 54, May 2019, Pages 30-44.

[51] T. Schlegl, P. Seebock, S. M. Waldstein, U. Schmidt-Erfurth, and G. Langs, "Unsupervised anomaly detection with generative adversarial networks to guide marker discovery," in *IPMI*, Springer, 2017, pp. 146-157.

[52] D. Zimmerer, S. A. Kohl, J. Petersen, F. Isensee, and K. H. Maier-Hein, "Context-encoding variational autoencoder for unsupervised anomaly detection," in *MIDL*, 2019

[53] X. Chen and E. Konukoglu, "Unsupervised detection of lesions in brain MRI using constrained adversarial auto-encoders," in *MIDL*, 2018

[54] S. You, K. C. Tezcan, X. Chen, and E. Konukoglu, "Unsupervised lesion detection via image restoration with a normative prior," in *MIDL* PMLR, 2019, pp. 540-556.

[55] N. Pawlowski *et al.*, "Unsupervised lesion detection in brain ct using bayesian convolutional autoencoders, in *MIDL*, 2018

[56] Baur C., Wiestler B., Albarqouni S., Navab N, "Deep Autoencoding Models for Unsupervised Anomaly Segmentation in Brain MR Images," In:*International MICCAI Brainlesion Workshop* vol 11383, 2019, Springer, pp. 161-169

[57] G. Pang, C. Shen and A. Hengel, "Deep anomaly detection with deviation networks," in *SIGKDD*, 2019, pp. 353-362.

[58] J. Zhang *et al.* "Viral Pneumonia Screening on Chest X-ray Images Using Confidence-Aware Anomaly Detection," *arxiv:2003.12338* 2020.

[59] E. Nalisnick, A. Matsukawa, Y. W. Teh, D. Gorur, and B. Lakshminarayanan, "Do Deep Generative Models Know What They Dont Know?" in *ICLR*, 2019

[60] P. Bergmann, M. Fauser, D. Sattlegger and C. Steger, "Uninformed Students: Student-Teacher Anomaly Detection With Discriminative Latent Embeddings," in *CVPR* 2020, pp. 4183-4192.

[61] D. Gong *et al.*, "Memorizing Normality to Detect Anomaly: Memory-Augmented Deep Autoencoder for Unsupervised Anomaly Detection," in *ICCV*, 2019, pp. 1705-1714.

[62] P. Bergmann, S. Lowe, M. Fauser, D. Sattlegger, and C. Steger, "Improving Unsupervised Defect Segmentation by Applying Structural Similarity to Autoencoders," *VISAPP*, 2019

[63] P. F. Christ *et al.* "Automatic Liver and Lesion Segmentation in CT Using Cascaded Fully Convolutional Neural Networks and 3D Conditional Random Fields," in *MICCAI*, 2016, Springer, pp. 415-423.

[64] R. Yan and N. Tokuda, "Analysis and Recognition of Medical Images: 1. Elastic Deformation Transformation." In: T*Communicating with Virtual Worlds. CGS CG International Series.* 1993, Springer, Tokyo.

[65] "LUNA16." https://luna16.grand-challenge.org/Home/

[66] Armato III *et al.* "Data From LIDC-IDRI," *The Cancer Imaging Archive.*, 2015, DOI:10.7937/K9/TCIA.2015.LO9QL9SX.

[67] Armato SG *et al.* "The Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI): A completed reference database of lung nodules on CT scans," *Medical Physics*, vol. 38, pp. 915-931, 2011, DOI:10.1118/1.3528204.

[68] F. Isensee, P. F. Jger, S. A. A. Kohl, J. Petersen and K. H. Maier-Hein, "Automated Design of Deep Learning Methods for Biomedical Image Segmentation," *arXiv:1904.08128*, 2020.

[69] K. W. Clark *et al.* "The cancer imaging archive (tcia): Maintaining and operating a public information repository," *J. Digital Imaging*, no. 6, pp. 10451057, 2013.

[70] K. Kiser *et al.* "Data from thethoracic volume and pleural effusion segmentations in diseased lungs for benchmarking chest ct processing pipelines," *The Cancer Imaging Archive*, 2020.

[71] H. J. Aerts *et al.* "Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach," *Nature communications*, vol. 5, no. 1, pp. 19, 2014.

[72] "StructSeg" https://structseg2019.grand-challenge.org

[73] "Medical Segmentation Decathlon" http://medicaldecathlon.com/