PIXEL-LEVEL AND AFFINITY-LEVEL KNOWLEDGE DISTILLATION FOR UNSUPERVISED SEGMENTATION OF COVID-19 LESIONS

Rui Xu[⋆†] *Yufeng Wang*[§†] *Xinchen Ye*[⋆†] *Pengcheng Wu*[§†] *Yen-Wei Chen*[⋄†] *Fangyi Xu*[‖] *Wenchao Zhu*[‖]
Chao Chen[‖] *Yong Zhou*[‖] *Hongjie Hu*[‖] *Xiaofeng Qu*[‡] *Shoji Kido*[¶] *Noriyuki Tomiyama*[¶]

[⋆]International School of Information Science & Engineering, Dalian University of Technology, China
[†]DUT-RU Co-Research Center of Advanced ICT for Active Life, China
[§]College of Software, Dalian University of Technology, China
[⋄]College of Information Science and Engineering, Ritsumeikan University, Japan
[‖]Department of Radiology, Sir Run Run Shaw Hospital, Zhejiang University School of Medicine, China
[‡]Department of Radiology, the Second Hospital of Dalian Medical University, China
[¶]Graduate School of Medicine, Osaka University, Japan

ABSTRACT

Automatic segmentation of COVID-19 lesions is essential for computer-aided diagnosis. However, this task remains challenging because widely-used supervised based methods require large-scale annotated data that is difficult to obtain. Although an unsupervised method based on anomaly detection has shown promising results in [1], its performance is relatively poor. We address this problem by proposing a pixel-level and affinity-level knowledge distillation method. It obtains a pre-trained teacher network with rich semantic knowledge of CT images by constructing and training an auto-encoder at first, and then trains a student network with the same architecture as the teacher by distilling the teacher's knowledge only from normal CT images, and finally localizes COVID-19 lesions using the feature discrepancy between the teacher and the student networks. Besides, except for the traditional pixel-level distillation, we design the affinity-level distillation that takes into account the pairwise relationship of features to fully distill effective knowledge. We evaluate this method by using three different COVID-19 datasets and the experimental results show that the segmentation performance is largely improved when it is compared with the other existing unsupervised anomaly detection methods.

Index Terms— Knowledge Distillation, Unsupervised Segmentation, Anomaly Detection, Affinity-Level

1. INTRODUCTION

In recent years, Corona Virus Disease 2019 (COVID-19) has spread around the world which is seriously threatening people's lives and health. In the fight against this epidemic, computer-aided diagnosis (CAD) systems on chest CT images can play a key role in quick diagnosis and assessment of COVID-19 [2]. Usually, these CAD systems include an important processing step that automatically segments COVID-19 lesions on CT images, but most of them apply a supervised method for the lesion segmentation and require large-scale well-labeled datasets [3, 4, 5]. Unfortunately, such datasets are very difficult to obtain. They usually require multiple experienced radiologists to spend a lot of time for lesion labeling, which is time-consuming and labor-intensive. We notice that there are a lot of CT images exhibiting almost normal tissues inside lung regions in daily clinical screening. It could be helpful if we can make use of these normal CT images to develop an unsupervised segmentation method. We accomplish this by proposing an anomaly detection (AD) based method, which utilizes only normal CT images in training and can segment COVID-19 lesions (abnormal opacities) in testing.

Most researches on AD have focused on natural image related tasks. In recent years, there have been some studies devoted to apply it on medical images [6, 7, 8, 9, 10, 11]. Typically, they are based on a reconstruction error framework, which trains generative models, such as auto-encoders (AEs) or generative adversarial networks (GANs), by only using normal medical images, and then applies the trained models to reconstruct images with pathological changes or disorders. Since these models have only seen normal images, reconstruction error could be high at the locations with anomaly. However, these methods do not perform well to segment anomaly on medical images, especially when they are evaluated by traditional segmentation measures, such as the Dice similarity coefficient (DSC). This is due to poor reconstruction, over-generalization and low resolution on re-

This work was supported by National Natural Science Foundation of China (NSFC) under Grant 61772106, Grant 61702078 and Grant 61720106005, and by the Fundamental Research Funds for the Central Universities.

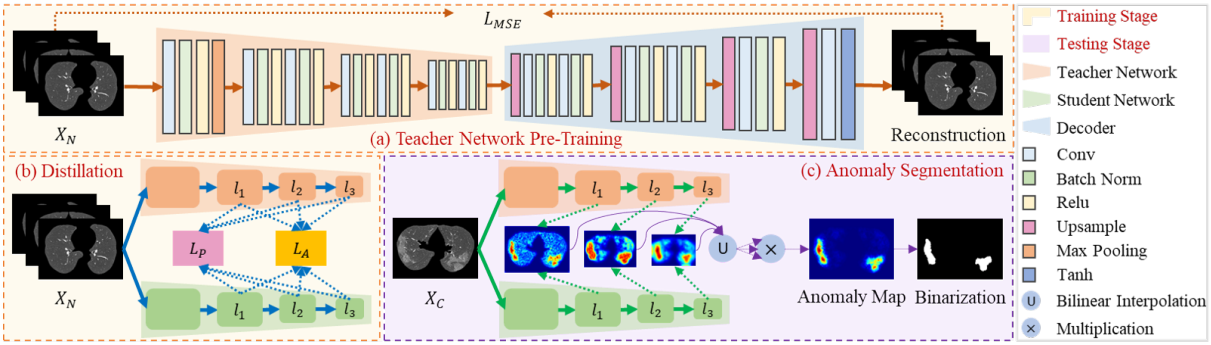

Fig. 1. An overview of the proposed method.

constructed images.

Recently, a knowledge distillation based framework has been invented to segment anomaly on natural image and shows very promising results [12, 13, 14]. It is based on a teacher-student learning strategy, and distills knowledge from a powerful pre-trained teacher network to a student network by only using normal natural images. When an abnormal image is input into the two networks, anomaly can be located by using the discrepancy of their outputs. We apply this framework on our task (method-a in Table 1) and find that it can outperform our pervious study (AE-CW-GAN in Table 2) [1], where an AE method combined with GAN is proposed to segment COVID-19 lesions. However, we notice that the vanilla application of this framework cannot exert its full capacity. The teacher network is typically pre-trained on ImageNet [15], which is quite different from CT image datasets. Besides, the knowledge distillation is only carried on pixel-level, ignoring relationship between similar pixels. Therefore, we address these problems by proposing a pixel-level and affinity-level knowledge distillation method, which exploits AE on CT images to obtain a more reasonable teacher network with rich lung semantic knowledge and distills useful knowledge to student network via pixel-level and affinity-level.

The main contributions are summarized as follows. 1) According to our knowledge, this is the first work applying the distillation based AD framework on the unsupervised segmentation of COVID-19 lesions. 2) To fully explore the capacity of the distillation based AD framework for our task, we exploit AE to obtain a more reasonable teacher network pre-trained on normal CT images and design a novel pixel-level and affinity-level distillation strategy. 3) We extensively evaluate our method on two public COVID-19 datasets [16][17] and one private dataset, and experimental results show our method outperforms the other state-of-the-art AD methods.

2. PROPOSED METHOD

In this paper, we propose an unsupervised method to segment COVID-19 lesions only by using normal CT images in training. Since it is too difficult to develop an unsupervised method directly on the whole scope of CT images, we restrict that our method operates only inside lung regions that can be extracted by any lung segmentation methods [18]. Thus, the input of our method is a series of 2D CT images cropped inside lungs.

The overview of the proposed method is shown in Fig. 1. Our method mainly consists of three steps, i.e., a) teacher network pre-training, b) pixel-level and affinity-level distillation and c) anomaly segmentation. At first, we construct an AE that is trained by the reconstruction of normal CT images. The encoder part of the AE is a powerful feature extractor that contains rich lung semantics information. Thus, we keep the encoder and take it as a pre-trained teacher network for the next step. Then, a teacher-student distillation procedure is carried out to train a student network with the same architecture as the teacher's. In order to allow the student network to learn more efficient knowledge, we design a novel strategy that can distill knowledge not only at the corresponding pixel locations (pixel-level) but also according to the feature similarity (affinity-level). Finally, in the testing stage, CT images with COVID-19 lesions are fed into both of the teacher and student networks, and the differences on multi-level features of both networks are integrated to identify where the lesions exist.

2.1. Teacher Network Pre-Training

Since CT images are quite different from natural images, models pre-trained on ImageNet not only contain less-representative features for lungs but also can lead bias for knowledge distillation. Thus, we train an AE to obtain a teacher network, instead of directly using a pre-trained ImageNet models as in [14]. Given that a series of normal CT images cropped inside lungs is denoted as X_N, we input an image $x_n \in X_N$ into the AE to obtain a reconstructed image $\hat{x_n}$. The AE is comprised by an encoder network f and a decoder network g. They are constructed by stacking a series of convolution, batch normalization, relu, pooling and upsampling layers, and the detailed architectures are illustrated by Fig 1 (a). We train the AE by minimizing a mean squared error loss $\mathcal{L}_{MSE}$ defined by Eq. 1. The encoder f of the trained AE is treated as a teacher network for the following knowledge distillation.

$$\mathcal{L}_{MSE} = \sum_{x_n \in X_N} \|x_n - \hat{x_n}\|_2 = \sum_{x_n \in X_N} \|x_n - g(f(x_n))\|_2 \tag{1}$$

2.2. Pixel-Level and Affinity-Level Distillation

Previous works on anomaly detection [12, 13, 14] usually train a student network by distilling teacher's knowledge at each pixel-level location on different levels of feature maps. However, such a pixel-level distillation ignores the pairwise relationship of features, which is valuable for dense prediction tasks, such as the segmentation of anomaly. If one location on an image is identified as anomaly, the surrounding places with similar features are also more likely to be anomaly. Thus, we design an affinity-level distillation together with the pixel-level distillation to ensure that the student can learn more efficient knowledge from the teacher.

Pixel-Level Distillation Given a normal CT image x_n in X_N, $f_t^l(x_n)$ and $f_s^l(x_n)$ represent feature maps of different levels for the teacher and student networks respectively, where $l = \{l_1, l_2, l_3\}$. If w_l and h_l are denoted as the width and height of a feature map, we define $\mathcal{L}_P^l(x_n)$ as the pixel-level distillation loss on the l-th feature map for the image x_n. Its definition is given by Eq. 2.

$$\mathcal{L}_P^l(x_n) = \frac{1}{w_l h_l} \sum_{i=1}^{h_l} \sum_{j=1}^{w_l} \left\| \frac{f_t^l(x_n)_{ij}}{\left\| f_t^l(x_n)_{ij} \right\|_2} - \frac{f_s^l(x_n)_{ij}}{\left\| f_s^l(x_n)_{ij} \right\|_2} \right\|_2^2 \tag{2}$$

The total pixel-level distillation loss is defined by summarizing the losses for all images on different levels in the training dataset X_N, and its definition is given by Eq. 3.

$$\mathcal{L}_P = \sum_{x_n \in X_N} \sum_{l=l_1,l_2,l_3} \mathcal{L}_P^l(x_n) \tag{3}$$

Affinity-Level Distillation Inspired by [19, 20] that non-local correlation is considered to strengthen connections between pixels, we also transfer non-local knowledge on affinity level by calculating pairwise similarity between pixels. Supposing feature dimension f^l is $w_l \times h_l \times c$, the reshaping function $\mathbb{R}$ recasts f^l as $\mathbb{R}(f^l)$ with the dimension of $w_l h_l \times c$. The affinity matrix A is given by:

$$A(f^l) = \sigma\left(\mathbb{R}(f^l)\right) \otimes \sigma\left(\mathbb{R}^T(f^l)\right) \tag{4}$$

where $\sigma(\cdot)$ is the sigmoid operation, $\otimes$ is the matrix multiplication and T is the transpose operator. The affinity-level distillation loss is defined as Eq. 5,

$$\mathcal{L}_A = \sum_{x_n \in X_N} \sum_{l=l_1,l_2,l_3} \left\| A\left(f_t^l(x_n)\right) - A\left(f_s^l(x_n)\right) \right\|_1 \tag{5}$$

The total loss for knowledge distillation $\mathcal{L}_{total}$ is expressed in Eq. 6.

$$\mathcal{L}_{total} = \mathcal{L}_P + \alpha \mathcal{L}_A \tag{6}$$

where α is an adjustment parameter. Note that, $\mathcal{L}_{total}$ should be imposed on the training of the student network, while the teacher network is fixed.

2.3. Anomaly Segmentation

The student network is learned to produce similar features as the teacher's when normal CT images are given in knowledge distillation. Thus, when CT images with COVID-19 lesions are sent into both networks, their features should also be similar at the locations where normal tissues exist. With regard to locations with COVID-19 lesions, the student network should produce features that are different from the teacher's, due to that the student has never been taught to be similar with the teacher on abnormal lung tissues. Therefore, the feature discrepancy between both networks can be exploited to segment COVID-19 lesions.

Given that a CT image in a COVID-19 dataset is denoted as $x_c \in X_C$, we can calculate the differences between feature maps on l-th level by $d^l(x_c) = \mathcal{L}_P^l(x_c)$. Then, we fuse them to generate a final anomaly map by using Eq. 7.

$$d(x_c) = \prod_{l=l_1,l_2,l_3} \mathrm{U}(d^l(x_c)) \tag{7}$$

where $\mathrm{U}(\cdot)$ is the upsampling operation and $\prod(\cdot)$ is element-wise multiplication.

Finally, we get a binary segmentation result of COVID-19 lesions $y(x_c)$ by setting a threshold T at each pixel location (i, j) on the final anomaly map, as the definition in Eq. 8.

$$y(x_c)_{ij} = \begin{cases} 1 & d(x_c)_{ij} > T \\ 0 & d(x_c)_{ij} \leq T \end{cases} \tag{8}$$

3. EXPERIMENTS

3.1. Datasets & Pre-processing

We collect a normal dataset of 69 CT volumes with normal lungs for training, and utilize three COVID-19 datasets for testing. The testing datasets includes a private COVID-19 dataset of 65 CT volumes, and two public COVID-19 datasets named as Coronacases [16] and Radiopaedia [17] respectively, both of which contain 10 CT volumes. All CT volumes are conducted by a pre-processing step comprised by lung segmentation [18], resampling and cropping to obtain a series of 2D CT images only containing regions inside lungs, before they are used in the training and testing. The obtained 2D CT images have the identical spatial size of 448×320 and the identical spacing of $0.6738 \times 0.6738 mm^3$ in the axial plane. Finally, the normal dataset contains 18538 CT images, while the testing datasets contains 5825, 1350 and 492 images respectively.

Table 1. Ablation study for the proposed method.

Methods	Pre-Training		Distillation		Private COVID-19 Dataset			Coronacase Dataset [16]			Radiopaedia Dataset [17]		
	ImageNet	CT	Pixel-L.	Affinity-L.	AUROC	AUPRC	DSC	AUROC	AUPRC	DSC	AUROC	AUPRC	DSC
method-a	✓		✓		0.9034	0.2231	0.2479	0.9557	0.5200	0.4926	0.9305	0.3512	0.3522
method-b		✓	✓		0.9424	0.3393	0.4086	0.9679	0.5448	0.5798	0.9510	0.4554	0.4810
method-c (proposed)		✓	✓	✓	**0.9495**	**0.3715**	**0.4302**	**0.9711**	**0.5651**	**0.5886**	**0.9582**	**0.5026**	**0.5310**

Table 2. The quantitative results of our method compared to other state-of-the-art AD methods.

Methods	Private COVID-19 Dataset			Coronacase Dataset [16]			Radiopaedia Dataset [17]		
	AUROC	AUPRC	DSC	AUROC	AUPRC	DSC	AUROC	AUPRC	DSC
AE [7]	0.9232	0.1377	0.2196	0.9469	0.2588	0.3634	0.9291	0.2425	0.3710
CAE [7]	0.8972	0.1129	0.1841	0.9198	0.2235	0.2966	0.9116	0.1883	0.2857
VAE [21]	0.9264	0.1407	0.2261	0.9490	0.2648	0.3646	0.9313	0.2529	0.3820
VCAE [21]	0.8842	0.0816	0.1445	0.9048	0.1532	0.2321	0.9031	0.1609	0.2733
AE-CW-GAN [1]	0.8938	0.1003	0.1709	0.9159	0.1922	0.2746	0.9076	0.1728	0.2813
MKD [13]	0.8875	0.0830	0.1475	0.8924	0.1142	0.1929	0.9132	0.2241	0.3176
STPM [14]	0.9034	0.2231	0.2479	0.9557	0.5200	0.4926	0.9305	0.3512	0.3522
Proposed	**0.9495**	**0.3715**	**0.4302**	**0.9711**	**0.5651**	**0.5886**	**0.9582**	**0.5026**	**0.5310**

3.2. Implement Details & Evaluation Metrics

All our experiments are implemented by a Pytorch framework and carried out on a TITAN RTX GPU. The AE is initialized randomly and optimized by using an Adam optimizer with the learning rate of 0.0002 in 50 epochs. Knowledge distillation is conducted in 2 epochs by using a stochastic gradient descent (SGD) optimizer with a momentum of 0.9, a weight decay of 1e-4, and a learning rate of 0.01. We set the parameter α in Eq. 6 to be 0.001.

We follow the previous work on anomaly segmentation in medical images [21] to evaluate our method by using three metrics, including two threshold-independent metrics, which are the area under the ROC curve (AUROC) and the area under the precision-recall curve (AUPRC), and a threshold-related metric called as DSC, the threshold is deduced from the best f1-score calculated by precision and recall.

3.3. Ablation Study

We perform the ablation study to verify the key designs of the proposed method, and list the detailed results in Table 1. The method-a, only using a ImageNet pre-trained teacher network and the pixel-level distillation, is a vanilla application of the distillation based AD framework on our task. Although it performs relative well, the capacity based on distillation framework cannot be fully exploited. Thus, we develop two enhanced versions, named as the method-b and method-c. The method-b replaces the ImageNet teacher network by using an encoder of the AE pre-trained on normal CT images, and the method-c is further enhanced by using knowledge distillation on both pixel-level and affinity-level. It can be seen that the performance can be gradually enhanced by using the CT pre-trained teacher network and the new strategy of distillation.

3.4. Comparison to the State-of-the-Art Methods

We compare the proposed method with seven state-of-the-art AD methods in Table 2. Five of them is based on the reconstruction error framework, including AE [7], convolution auto-encoder (CAE) [7], variational auto-encoder (VAE) [21], variational convolution auto-encoder (VCAE) [21], and AE method combined with GAN (AE-CW-GAN) [1]. The other two methods named as MKD [13] and STPM [14] are based on the knowledge distillation framework and originally proposed to segment anomaly on natural images. It can be seen that the proposed method can outperform all of these state-of-the-art methods and show the best performance on the unsupervised segmentation of COVID-19 lesions. Finally, we give some examples of visualization results by using the proposed method on the three testing datasets in Fig. 2.

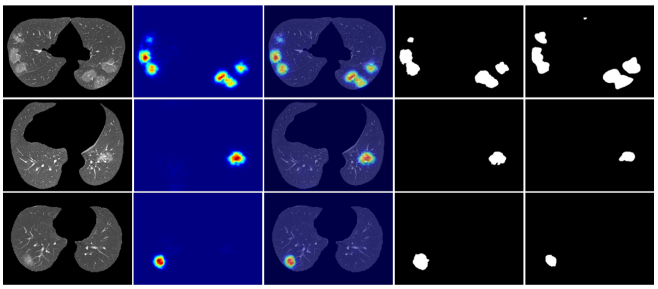

Fig. 2. Examples of visualization results for lesions segmentation by using the proposed method (method-c) on Private COVID-19 dataset, Coronacase and Radiopaedia. (a) CT images, (b) anomaly map, (c) anomaly map overlaid on the CT image, (d) binary map of anomaly map, (e) ground truth.

4. CONCLUSION

We proposed a knowledge distillation based AD method for unsupervised segmentation of COVID-19 lesions. By exploiting a teacher network pre-trained in AE on normal CT images and carefully designing a pixel-level and affinity-level distillation strategy, our method can outperform seven state-of-the-art AD methods in the evaluation of three COVID-19 datasets.

5. REFERENCES

[1] Rui Xu, Xiao Cao, and et al., "Unsupervised detection of pulmonary opacities for computer-aided diagnosis of covid-19 on ct images," in *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE, 2021, pp. 9007–9014.

[2] Tao Ai, Zhenlu Yang, and et al., "Correlation of chest ct and rt-pcr testing in coronavirus disease 2019 (covid-19) in china: a report of 1014 cases," *Radiology*, p. 200642, 2020.

[3] Jun Chen, Lianlian Wu, and et al., "Deep learning-based model for detecting 2019 novel coronavirus pneumonia on high-resolution computed tomography," *Scientific reports*, vol. 10, no. 1, pp. 1–11, 2020.

[4] Yukun Cao, Zhanwei Xu, and et al., "Longitudinal assessment of covid-19 using a deep learning–based quantitative ct pipeline: illustration of two cases," *Radiology: Cardiothoracic Imaging*, vol. 2, no. 2, pp. e200082, 2020.

[5] Lu Huang, Rui Han, and et al., "Serial quantitative chest ct assessment of covid-19: a deep learning approach," *Radiology: Cardiothoracic Imaging*, vol. 2, no. 2, pp. e200075, 2020.

[6] Thomas Schlegl, Philipp Seeböck, and et al., "Unsupervised anomaly detection with generative adversarial networks to guide marker discovery," in *International conference on information processing in medical imaging*. Springer, 2017, pp. 146–157.

[7] Christoph Baur, Benedikt Wiestler, and et al., "Deep autoencoding models for unsupervised anomaly segmentation in brain mr images," in *International MICCAI Brainlesion Workshop*. Springer, 2018, pp. 161–169.

[8] Nick Pawlowski, Matthew CH Lee, and et al., "Unsupervised lesion detection in brain ct using bayesian convolutional autoencoders," *openreview.net*, 2018.

[9] David Zimmerer, Simon AA Kohl, and et al., "Context-encoding variational autoencoder for unsupervised anomaly detection," *arXiv preprint arXiv:1812.05941*, 2018.

[10] Thomas Schlegl, Philipp Seeböck, and et al., "f-anogan: Fast unsupervised anomaly detection with generative adversarial networks," *Medical image analysis*, vol. 54, pp. 30–44, 2019.

[11] Xiaoran Chen, Suhang You, and et al., "Unsupervised lesion detection via image restoration with a normative prior," *Medical image analysis*, vol. 64, pp. 101713, 2020.

[12] Paul Bergmann, Michael Fauser, and et al., "Uninformed students: Student-teacher anomaly detection with discriminative latent embeddings," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 4183–4192.

[13] Mohammadreza Salehi, Niousha Sadjadi, and et al., "Multiresolution knowledge distillation for anomaly detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 14902–14912.

[14] Guodong Wang, Shumin Han, and et al., "Student-teacher feature pyramid matching for unsupervised anomaly detection," *arXiv preprint arXiv:2103.04257*, 2021.

[15] Jia Deng, Wei Dong, and et al., "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[16] "Coronacases," `https://coronacases.org/`.

[17] "Radiopaedia," `https://radiopaedia.org/articles/covid-19-4`.

[18] Rui Xu, Yi Wang, and et al., "Bg-net: Boundary-guided network for lung segmentation on clinical ct images," in *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE, 2021, pp. 8782–8788.

[19] Yukang Gan, Xiangyu Xu, and et al., "Monocular depth estimation with affinity, vertical pooling, and label enhancement," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 224–239.

[20] Xiaolong Wang, Ross Girshick, and et al., "Non-local neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7794–7803.

[21] Christoph Baur, Stefan Denner, and et al., "Autoencoders for unsupervised anomaly segmentation in brain mr images: a comparative study," *Medical Image Analysis*, p. 101952, 2021.