Mix-and-Interpolate: A Training Strategy to Deal With Source-Biased Medical Data

Yuexiang Li, Jiawei Chen, Dong Wei, Yanchun Zhu, Jianrong Wu, Junfeng Xiong, Yadong Gang, Wenbo Sun, Haibo Xu, Tianyi Qian, Kai Ma, and Yefeng Zheng

***Abstract*—Till March 31st, 2021, the coronavirus disease 2019 (COVID-19) had reportedly infected more than 127 million people and caused over 2.5 million deaths worldwide. Timely diagnosis of COVID-19 is crucial for management of individual patients as well as containment of the highly contagious disease. Having realized the clinical value of non-contrast chest computed tomography (CT) for diagnosis of COVID-19, deep learning (DL) based automated methods have been proposed to aid the radiologists in reading the huge quantities of CT exams as a result of the pandemic. In this work, we address an overlooked problem for training deep convolutional neural networks for COVID-19 classification using real-world multi-source data, namely, the *data source bias* problem. The data source bias problem refers to the situation in which certain sources of data comprise only a single class of data, and training with such source-biased data may make the DL models learn to distinguish data sources instead of COVID-19. To overcome this problem, we propose MIx-aNd-Interpolate (MINI), a conceptually simple, easy-to-implement, efficient yet effective training strategy. The proposed MINI approach generates volumes of the absent class by combining the samples collected from different hospitals, which enlarges the sample space of the original source-biased dataset. Experimental results on a large collection of real patient data (1,221 COVID-19 and 1,520 negative CT images, and the latter consisting of 786 community acquired pneumonia and 734 non-pneumonia) from eight hospitals and health institutions show that: 1) MINI can improve COVID-19 classification performance upon the baseline (which does not deal with the source bias), and 2) MINI is superior to competing methods in terms of the extent of improvement.**

***Index Terms*—Source-biased data, interpolation, multi-source data, training strategy.**

I. Introduction

THE outbreak of the coronavirus disease 2019, or COVID-19 in short, was declared a pandemic on March 11th, 2020 by the World Health Organization. Till March 31st, 2021, reportedly over 127 million people had been diagnosed as confirmed COVID-19 cases, and more than 2.5 million people died of this highly contagious disease.[1] Early diagnosis of COVID-19 is crucial, both for timely treatment of individual patients and effective control of disease spread. On one hand, despite that the reverse transcription polymerase chain reaction (RT-PCR) assay of sputum or nasopharyngeal swab is considered as the gold standard for COVID-19 diagnosis, several studies have reported insufficient sensitivities (*i.e.,* high false negative rates) of RT-PCR for effective early diagnosis of presumptive patients [1]–[3]. On the other hand, considering that most COVID-19 patients exhibit respiratory symptoms (mainly pneumonia) [4], the non-contrast thoracic computed tomography (CT) can be an alternative solution for confirming COVID-19 cases from suspected cohorts.

From thoracic CT exams, typical chest CT findings of COVID-19 reported in the literature [2], [5] include diffuse or focal ground-glass opacities, particularly bilateral and peripheral ground glass, as well as consolidative pulmonary opacities. In addition, Bernheim *et al.* [5] noted the progression of chest CT manifestations from early to late stages, characterized by greater lung involvement, crazing paving and reverse halo signs, etc. Moreover, both [3] and [2] reported superior sensitivities of non-contrast chest CT to RT-PCR. Under such circumstances, computer aided diagnosis (CAD) systems based on deep learning (DL) methodologies [6]–[8] can support radiologists in the triage, quantification and trend analysis of COVID-19 diagnosis, at least from two aspects: i) reducing the heavy workload of exam reading and increasing the throughput capacity with efficient, automated pipelines (a thin-slice CT sequence of 300 slices can take a radiologist 5–15 minutes to diagnose); and ii) reducing the chance of fatigue- and omission-induced, and lack-of-experience related (*i.e.,* junior radiologists) misdiagnosis with objective and reliable CAD systems.

Facing the worldwide sudden outbreak of COVID-19, the research community of medical image analysis has

Manuscript received March 31, 2021; revised August 27, 2021; accepted October 5, 2021. Date of publication October 12, 2021; date of current version January 5, 2022. This work was supported by the Key-Area Research and Development Program of Guangdong Province, China under grant 2018B010111001, National Key R&D Program of China under grant 2018YFC2000702, and the Scientific and Technical Innovation 2030- "New Generation Artificial Intelligence" Project under grant 2020AAA0104100. *(Yuexiang Li, Jiawei Chen, and Dong Wei contributed equally to this work.) (Corresponding author: Yefeng Zheng.)*

Yuexiang Li, Jiawei Chen, Dong Wei, Kai Ma, and Yefeng Zheng are with the Tencent Jarvis Lab, Shenzhen 518000, China (e-mail: vicyxli@tencent.com; jiaweichen@tencent.com; donwei@tencent.com; kylekma@tencent.com; yefengzheng@tencent.com).

Yanchun Zhu, Jianrong Wu, Junfeng Xiong, and Tianyi Qian are with the Tencent Health, Shenzhen 518000, China (e-mail: noonezhu@tencent.com; edwinjrwu@tencent.com; francoxiong@tencent.com; tianyiqian@tencent.com).

Yadong Gang, Wenbo Sun, and Haibo Xu are with the Department of Radiology, Zhongnan Hospital of Wuhan University, Wuhan 430071, China (e-mail: gangyadong@hotmail.com; sunwb3@mail2.sysu.edu.cn; xuhaibo1120@hotmail.com).

Digital Object Identifier 10.1109/JBHI.2021.3119325

[1][Online]. Available: https://covid19.who.int/

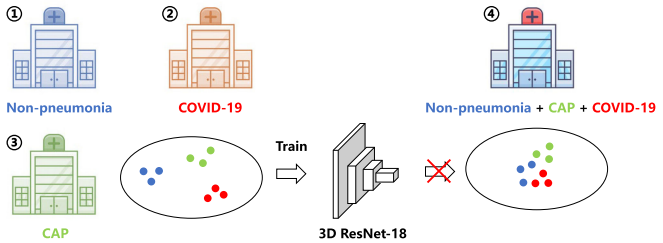


Fig. 1. The data bias problem: certain sources of data may comprise only a single class of data (*e.g.,* Source 2 has only positive cases (COVID-19), whereas Sources 1 and 3 have only negative cases—non-pneumonia and community acquired pneumonia (CAP), respectively), introducing a source-related bias in network training. The data bias problem occurred in reality while developing the CAD for COVID-19. Hospitals exclusively designated for accommodation of COVID-19 cases (such as the emergency specialty field hospitals built in Wuhan, China) provide mostly COVID-19 scans, whereas a physical examination institution provides mostly non-pneumonia cases, and the cases from conventional hospitals are usually CAP. Models trained with such multi-source data without dealing with the source bias cannot generalize to the data collected from an unseen centre (Source 4).

responded quickly and proposed new methods/systems tailored for imaging-based diagnosis of COVID-19. For example, Han *et al.* [9] proposed a novel attention-based deep 3D multiple instance learning to classify CT scans into three categories, *i.e.,* COVID-19, common pneumonia, and non-pneumonia, and achieved a high overall accuracy (97.9%) on a relatively small collection of data (460 CT scans). Similarly, both [10] and [11] proposed to discriminate between COVID-19 and community acquired pneumonia (CAP) from chest CT images and experimentally demonstrated moderate (88%) to high accuracies (95.5%), yet with much larger datasets (4,982 and 2,522 CT scans, respectively). Despite many studies have emerged to tackle various COVID-19 related tasks (more will be reviewed shortly), a practically relevant and important task regarding the unintended data bias is less explored. When collecting moderate to large scale training data from multiple sources, we notice that many sources provide data naturally dominated by a single class (Fig. 1). In reality, hospitals exclusively designated for accommodation of COVID-19 cases (such as the emergency specialty field hospitals built in Wuhan, China) provide mostly COVID-19 scans, whereas a physical examination institution provides mostly non-pneumonia cases, and the cases from conventional hospitals are usually CAP. As different sources are equipped with different CT scanners, scanning protocols and post-processing procedures, the generated CT scans are most likely to have implicit image characteristics representing the origins. Training with such biased data may lead to suboptimal and discriminatory outcomes, *i.e.,* associating model predictions with data sources instead of pneumonia types. As far as the authors are aware of, despite its practical relevance, this problem has not been considered yet for imaging-based CAD of COVID-19.

In this work, we propose a training strategy, namely MIx-aNd-Interpolate (MINI), to deal with the data bias problem occurred in real-world COVID-19 diagnosis applications by expanding the sample space. Particularly, the volumes of the absent class for each data source are generated by combining samples from another source (domain) via mixing and trilinear interpolation. It is worthwhile to mention that our MINI is easy-to-implement and computation-friendly; therefore, researchers can respond rapidly to the source bias problem using our training strategy, which is especially important facing the rapid spread of COVID-19.

In summary, our contributions are three folds:

1) Unlike existing literature, we address a practically relevant yet previously overlooked problem for training COVID-19 classification networks using real-world multi-source data, that is, certain data sources may comprise only a single class of data, introducing source bias in training.
2) We propose MIx-aNd-Interpolate (MINI), an intuitive and easy-to-implement training strategy, to deal with the specific problem.
3) We conduct thorough experiments to validate the effectiveness of MINI, study its behaviour and optimal configuration, and demonstrate its superiority to applicable competing methods.

II. Related Work

A. Imaging-Based Diagnosis of COVID-19

In the past few months, we have seen that a great number of imaging-based COVID-19 diagnosis systems have been proposed to use advanced deep learning (DL) techniques with various imaging modalities. Roy *et al.* [12] explored the application of DL methodologies for the classification and localization of COVID-19 markers using lung ultrasonography (LUS) images; Oh *et al.* [13] proposed a patch-based DL network architecture and employed majority voting for the final decision, to detect COVID-19 signs in chest X-rays (CXR). However, most methods were proposed for CT-based diagnosis, as CT images (especially those with thin slice below 2 mm) can capture more and finer details of the lung infection compared to LUS and CXR, thus having a higher clinical diagnostic value.

For CT images, rapid development of DL models in response to the sudden outbreak of COVID-19 is among the mostly visited topics. Gozes *et al.* [14] proposed to modify and adapt existing DL models that performed a similar task and combined them with initial clinical understanding of COVID-19 for rapid development of a new CAD system, at the early stage of the disease outbreak when both the data and annotation were scarce. The need for rapid model development was also considered in [15], which additionally addressed the data imbalance problem (*i.e.,* much fewer COVID-19 cases compared to the available CAP and non-pneumonia cases) using a self-supervised dual-track ranking strategy. Besides rapid development of the models, fast screening in practical use has also drawn attention, *e.g.,* Wang *et al.* [16] described a DL model that was able to process a whole CT volume—including both COVID-19 classification and lesion localization—in less than two seconds. Another frequently visited topic is to cope with the scarcity or difficulty in manual data annotation, for which both semi-supervised [17] and weakly supervised [12], [16] methodologies were proposed.

Unlike all the related work described above, this work deals with a problem that has been overlooked so far, *i.e.,* the data source bias problem, and proposes a simple yet effective approach (namely MIx-aNd-Interpolate, MINI) as a solution.

B. Unsupervised Domain Adaptation

To alleviate the problem caused by different imaging conditions of multiple centres, existing solutions applied the unsupervised domain adaptation technique to close the gap between the source and target domains. One common choice for domain adaptation is to directly align two different domains in the feature subspace [18], [19]. For example, Sun *et al.* [18] proposed the CORrelation Alignment (CORAL) method for unsupervised domain adaptation, which minimizes domain shift by aligning the second-order statistics of source and target distributions. In more recent studies, researchers tried to adopt generative adversarial networks (GANs) to reduce domain discrepancy [20]–[23] by optimizing an adversarial objective between the feature learning network and a domain discrimination network. Hoffman *et al.* [20] developed an method, called cycle-consistent adversarial domain adaptation (CyCADA), to guide transfer between domains according to a discriminatively trained network. The approach alleviated the divergence problem by enforcing consistency of the relevant semantics before and after adaptation. Meanwhile, the GAN-based domain adaptation is also widely used in the area of person re-identification (re-ID) [24], [25] and medical image processing, *e.g.,* for color normalization of histopathological slices [26], intensity standardization of magnetic resonance images [27] and cross-modality adaptation [28]. However, most existing domain adaptation approaches focus on transferring knowledge between two domains (source and target), which are difficult to adapt to tasks with multi-source data (*i.e.,* data from more than two centres) [29].

C. Domain Generalization

As an extension of domain adaptation, domain generalization (DG) aims to learn from multiple source domains to incorporate domain invariance into the model, in hopes that such invariance also holds in target domains. Most previous approaches attempted to learn a domain-invariant feature representation, typically through minimising the discrepancy between all source domains and assuming the resulting domain-invariant feature will generalize well for unseen target distributions [30]–[34]. More recently, meta-learning [35] starts to be used for solving DG problems. Bala *et al.* [36] learned an adaptive regularizer through meta-learning for cross-domain recognition. Li *et al.* [37] alternated domain-specific feature extractors and classifiers across domains via episodic training, but without using inner gradient descent update. Dou *et al.* [38] learned discriminative features that would allow for semantic coherence across meta-train and meta-test domains. However, meta-learning approaches are complicated to implement, and slow to train. Meanwhile, data augmentation strategies, such as gradient-based domain perturbation [39] or adversarially perturbed samples [40] demonstrated effectiveness for model generalization. These methods augment the source domain to a wider span of the training data space for enlarging the possibility of covering the span of the data in the target domain. A recent method with the state-of-the art performance is JiGen [41], which leverages self-supervised signals by solving jigsaw puzzles.

Despite achieving promising improvements to the model generalization, both unsupervised domain adaptation and generalization approaches require the source and target domains to consist of the same categories, which cannot always be fulfilled in practical applications, *e.g.,* when the problem of missing category occurs for multi-source data as formulated in this study.

Algorithm 1 Volume Generation by MINI

Input:
- $v_A, v_B \in R^{X\times Y\times Z}$: input CT samples from different domains (resampled to a uniform shape $X \times Y \times Z$)
- λ: the hyperparameter controlling the number of slices taken for interpolation

Output:
- $\tilde{v} \in R^{X\times Y\times Z}$: the generated CT volume

Functions:
- $Init(S)$: generate an all-zero volume of shape S
- $Rand()$: randomly yield a floating point number in range $[0, 1]$
- $mod(.)$: calculate the remainder
- $Intp(v, S, type)$: resize v to shape S using the designated interpolation method (*'trilinear'* or *'nearest'*)

Volume Generation Procedure:
Step 1 Initialization
  $v^{temp} \leftarrow Init(X \times Y \times 2Z)$
  $z \leftarrow 1$
  $k \leftarrow 1$
  $\alpha \leftarrow Rand()$
$\beta \leftarrow 1 - \alpha$
Step 2 Mixing
  for z in range($2Z$):
    if $mod(z, 2) == 1$:
      $v^{temp}(:, :, z) \leftarrow \alpha\, v_A(:, :, \lfloor \frac{z}{2} \rfloor)$
    else:
      $v^{temp}(:, :, z) \leftarrow \beta\, v_B(:, :, \lfloor \frac{z}{2} \rfloor)$
    $z \leftarrow z + 1$
Step 3 Interpolation
  for k in range($\lambda - 1$):
    $v^{temp} \leftarrow Intp(v^{temp}, X \times Y \times \frac{2Z}{2^k}$, *'trilinear'*)
    $k \leftarrow k + 1$
  $\tilde{v} \leftarrow Intp(v^{temp}, X \times Y \times Z$, *'nearest'*)

III. Mix-and-Interpolate

Medical images from multicentres are often captured under different imaging conditions, making models trained in one domain (centre) frequently fail to generalize on another. Therefore, the generalization capacity of machine learning models to unseen domains draws increasing attention from the research community. To address the problem, recent studies have proposed various domain generalization methods [37], [38]. However, compared to the typical domain generalization, the challenge we

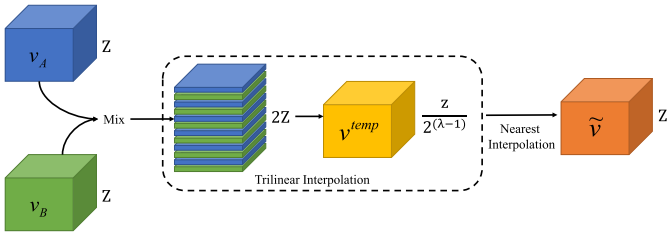


Fig. 2. Illustration of the proposed MINI. The proposed approach consists of three steps—mix, trilinear interpolation and nearest interpolation. A hyperparameter λ is used to control the amount of information to fuse along the z-axis.

face while developing the CAD system for COVID-19 diagnosis is more skewed. As mentioned earlier, many hospitals may provide only COVID-19 or non-pneumonia samples, which the current domain generalization methods fail to handle. Witnessed the fast outbreak of COVID-19, we propose a simple yet effective approach, namely MIx-aNd-Interpolate (MINI), to deal with such a training set containing biased multi-source data, which accordingly improves the generalization capacity of deep learning models.

A. Algorithm

The proposed MINI addresses the problem of biased multi-source data via extra data generation. Specifically, let $v \in R^{X\times Y\times Z}$ (where X, Y and Z are the shape of v) denote a 3D CT sample in the training set. Its corresponding class label is denoted as l. MINI aims to generate a new training sample $(\tilde{v}, \tilde{l})$ by combining two authentic training samples (v_A, l_A) and (v_B, l_B) from different domains (sources), *i.e.,* CT volumes of different patients captured in different hospitals. The generation process of $\tilde{v}$ is summarized in Algorithm 1 and illustrated in Fig. 2. In short, MINI generates samples by mixing the volume-wise information of v_A and v_B via interpolation. To generate samples having closer intensity distributions to the original samples, we stretch the grayscale of $\tilde{v}$ to the average interval of v_A and v_B after the volume generation.

On one hand, the proposed MINI is simple, which costs a negligible computational overhead like existing data augmentation approaches, *e.g.,* volume rotation and flipping. On the other hand, MINI-generated samples $(\tilde{v})$ can effectively narrow down the gap between domains and improve the model generalization. Such an efficient method can be readily applied for improving the performance of any network architecture.

B. Labeling Strategy

After the generation of volume $\tilde{v}$, we propose two strategies (hard and soft labeling) to yield the corresponding label, which are introduced below.

1) Hard Labeling: Taking binary classification as an example, this strategy assigns 0 (negative) or 1 (positive) to the generated volume, according to the original labels of v_A and v_B (*i.e.,* l_A and l_B). Denoting the assigned label as $\tilde{l}_h$, the rule of hard labeling can be formulated as:

$$\tilde{l}_h = \begin{cases} 0, & l_A = 0 \text{ and } l_B = 0; \\ 1, & l_A = 1 \text{ or } l_B = 1. \end{cases} \tag{1}$$

Specifically, the generated volume $(\tilde{v})$ is treated as positive (1) if there is at least one positive sample in v_A and v_B. Otherwise, a negative label (0) is assigned to $\tilde{v}$.

2) Soft Labeling: Consistent with the existing data augmentation approaches, the soft-labeling strategy is also adopted in our MINI. Particularly, the random factors α and β in Algorithm 1 are utilized to yield the soft label $(\tilde{l}_s)$ as a pair of labels $(\alpha\, l_A, \beta\, l_B)$.[2] The soft labeling can reflect the similarity of the generated sample to negative and positive samples, respectively, when v_A and v_B are from different categories; otherwise, this strategy equals to the hard labeling strategy.

C. Analysis of Hyperparameter λ

As presented in Algorithm 1, our MINI performs information fusion along the z-axis via trilinear interpolation.[3] Since the trilinear interpolation can only fuse the information from the adjacent slices, we propose an iterative interpolation strategy,[4] as presented in Algorithm 1 (Lines 18–22), to involve more slices from v_A and v_B for the per slice generation of $\tilde{v}$. The hyperparameter λ controls the number of slices from each training volume for the per slice generation; specifically, $2^{(\lambda-1)}$ slices from v_A and v_B are utilized to generate a single slice of $\tilde{v}$. The influence of this hyperparameter on performance will be empirically studied in the next section.

Additionally, as shown in Fig. 2, the proposed MINI adopts the neareast neighbor interpolation for shape unification, instead of the trilinear interpolation. The underlying reason is that the step of shape unification aims to reshape the compressed volume to the original size for neural network to process, which may not be sensitive to the choice of interpolation approach. To validate this claim, we conduct experiments and also present the results in the next section.

D. Comparison With Existing Approaches

We are aware that there are several existing approaches, *e.g.,* Mixup [42], Cutout [43], and CutMix [44], highly related to our MINI. They adopt different ways to produce synthetic samples. Specifically, Mixup generates the synthetic sample by fusing two authentic ones via pixel-wise summation (*e.g.,* $\tilde{v} = \frac{v_A + v_B}{2}$). Compared to Mixup, Cutout is a simpler approach, which yields new samples by cropping part of the image content. CutMix can be seen as a hybrid method combining Mixup and Cutout. It crops part of v_A and fill the content of v_B to the cropped area. In this section, a comprehensive comparison (Table I) is presented

[2]The optimization process of the mixed label $(\alpha\, l_A, \beta\, l_B)$ is the same as Mixup [42].

[3]Trilinear interpolation is a method of multivariate interpolation on a 3-dimensional regular grid. It approximates the value of intermediate voxel by taking the surrounding eight corner voxels into consideration. For detailed information, please refer to http://paulbourke.net/miscellaneous/interpolation.

[4]A series of v_{temp}s is generated (one smaller than the preceding one), and only the last one (in the loop) is retained (all preceding ones being overwritten).

TABLE I
COMPARISON BETWEEN EXISTING APPROACHES AND OUR MINI

	Cutout	CutMix	Mixup	MINI
Full volume region	×	√	√	√
Mixed label	×	√	√	√
Complete lesion	×	×	√	√
Cross slice fusion	×	×	×	√

to illustrate the advantage of the proposed MINI, compared to the existing approaches.

As shown in Table, only the Cutout approach encounters the problem of lacking usage of '"full volume region"' and '"mixed label"'. In particular, lacking usage of '"full volume region"' is caused by the discard operation in Cutout, which decreases the information contained in the newly generated samples, *i.e.,* part of the volume (all black) provides no useful information at all. Moreover, the Cutout approach generates new samples without fusing information from two volumes. Hence, it assigns the original label to the generated samples, instead of mixing the labels from the fusing volumes ('"mixed label"').

Some existing approaches (Cutout and CutMix) face the problem of '"complete lesion"', since they block/replace part of the volume content to generate new samples. When these approaches happen to crop and discard the lesion areas, the label of the generated volume may change (*i.e.,* COVID-19 samples without lesion area become negative samples). However, these approaches still assign the original labels to the generated volumes. Training with such '"incorrectly-labeled"' data, the performance of deep learning network may thereby degrade.

Last but not least, the existing approaches share a common shortage when applied to 3D volumes—the lack of cross-slice information fusion along the z-axis. Particularly, Mixup, Cutout, and CutMix perform slice-wise fusion using voxel-wise summation, region cropping, and region filling, respectively. As a result, the information cannot be transmitted and fused properly between adjacent slices. Current studies [45] demonstrate that the integration of spatial information plays an important role for 3D medical image processing. In this regard, our MINI adopts the trilinear interpolation to properly fuse per slice information during volume generation.

IV. EXPERIMENTS

In this section, we evaluate our MINI for its capability to improve the model generalization on multi-domain data. We first conduct extensive experiments to evaluate the impacts of the hyperparameter λ, labeling strategies and interpolation approaches. Next, the proposed MINI is compared with the state-of-the-art approaches to validate its effectiveness. Finally, we also show that our MINI can effectively deal with an increasing number of domains/centres, *i.e.,* a training set with more diverse samples.

A. Experimental Settings

1) Dataset: We collected 2,173 CT volumes from seven hospitals exclusively for model training, where each of them provided one of the three categories: COVID-19 (1,046), CAP (652) and non-pneumonia (475). The detailed information of the collected data is shown in Table II. In this study, we focus on the COVID-19 diagnosis task, which aims to separate COVID-19 (positive) patients from the negatives (*i.e.,* CAP and non-pneumonia). The DL models trained with the multi-source data are then evaluated on a test set, consisting of 175 COVID-19, 134 CAP and 259 non-pneumonia CT samples, collected from the eighth hospital. This experimental setting enables us to validate the generalization of models trained with multi-source biased data on an unseen domain.

2) Evaluation Metric: The sensitivity (SEN), specificity (SPE), and F1 score are adopted as the metrics to evaluate the performance of COVID-19 diagnosis, which can be written as:

$$SEN = \frac{N_{tp}}{N_{tp} + N_{fn}}, \quad SPE = \frac{N_{tn}}{N_{tn} + N_{fp}}, \tag{2}$$

$$F_1 = \frac{2\,N_{tp}}{2\,N_{tp} + N_{fp} + N_{fn}}, \tag{3}$$

where N_{tn}, N_{tp}, N_{fp}, and N_{fn} represent the numbers of true negatives, true positives, false positives, and false negatives, respectively. The average classification accuracy (ACC) and area under curve (AUC) are also adopted as metrics for the performance evaluation.[5]

3) Implementation: All experiments are implemented using PyTorch and according to the same training protocol. The 3D ResNet-18 [46] is adopted as the backbone, which is trained with a mini-batch size of 16. The initial learning rate is set to 0.001, which decreases with a gamma of 0.1 every 30 epochs of training. We train the network for 200 epochs. The Adam solver [47] is used for optimization.

It is worthwhile to mention that all the augmentation algorithms generate the equal numbers of samples during network training for a fair comparison. Concretely, the online data augmentation approach is adopted—the model decides whether to augment the input volume based on a random process. For the Mixup, Cutout, CutMix and our MINI, we first draw a random value (p) according to a uniform distribution in [0, 1]. If p is larger than 0.5, the model randomly selects another sample from the batch and performs corresponding augmentation approach with the selected and input volumes to generate a new sample for network training. For the conventional training strategy, an augmentation approach is randomly selected from the pool (*i.e.,* `torchvision.transforms`, which includes rotation, flipping and elastic deformation, etc.) for new sample generation if $p > 0.5$. The input sample is maintained and fed to the neural network if $p \leq 0.5$. Hence, the total number of samples for network training is exactly the same (*i.e.,* the size of training set) for different training strategies.

During inference, the well-trained network outputs the possibility for the positive class. If the possibility is larger than 0.5, the patient is identified as COVID-19. Note that the proposed MINI and existing approaches (*e.g.,* Mixup, Cutout and CutMix) are used only for training. After that, the trained classification

[5]The F_1 score, providing a comprehensive evaluation of classification performance, is taken as the primary metric, while the others are also listed for reference.

TABLE II
DETAILED INFORMATION OF MULTI-SOURCE DATASET

	Centre	No. of CT Volumes	No. slices	Slice thickness (mm)	Matrix size	Pixel spacing (mm)
Training	Hospital #1 ⋆	305	38–629	0.625–7.5	512×512	0.604–0.977
	Hospital #2 ⋆	741	79–547	1.0–1.25	512×512	0.580–0.912
	Hospital #3 ◇	144	47–553	1.0–5.0	512×512	0.330–0.977
	Hospital #4 ◇	217	125–407	1.0–1.5	512×512	0.387–0.977
	Hospital #5 ◇	291	156–591	0.6–1.5	512×512	0.324–0.977
	Hospital #6 ●	201	29–131	5.0	512×512	0.430–0.977
	Hospital #7 ●	274	56–950	0.6–2.0	512×512	0.396–0.977
Test	Hospital #8 ⋆ ◇ ●	568	65–710	0.6–3.0	512×512	0.342–0.998

⋆, ◇, ● Indicate the COVID-19, CAP and non-pneumonia, respectively.

TABLE III
PERFORMANCE (%) ANALYSIS OF MINI WITH DIFFERENT VALUES OF HYPERPARAMETER λ (SEN–SENSITIVITY, SPE–SPECIFICITY, ACC–ACCURACY, AUC–AREA UNDER CURVE)

λ	SEN	SPE	ACC	AUC	F_1
2	84.3	87.3	86.4	92.7	79.3
3	**86.6**	**87.3**	**87.1**	**92.8**	**80.6**
4	85.4	87.3	86.7	92.7	79.9
5	84.3	86.2	85.7	92.7	78.4

TABLE IV
COVID-19 CLASSIFICATION PERFORMANCE (%) OF MINI USING DIFFERENT LABELING STRATEGIES WITH $\lambda = 3$ (SEN–SENSITIVITY, SPE–SPECIFICITY, ACC–ACCURACY, AUC–AREA UNDER CURVE)

	SEN	SPE	ACC	AUC	F_1
w/o MINI	77.0	87.0	83.9	89.6	74.7
Hard labeling	**88.1**	85.6	86.3	92.6	79.9
Soft labeling	86.6	**87.3**	**87.1**	**92.8**	**80.6**

TABLE V
COVID-19 CLASSIFICATION PERFORMANCE (%) OF MINI USING DIFFERENT UNTERPOLATION APPROACHES FOR INFORMATION FUSION (SEN–SENSITIVITY, SPE–SPECIFICITY, ACC–ACCURACY, AUC–AREA UNDER CURVE, P. T.–PROCESSING TIME FOR INTERPOLATION, NEIG.–NEIGHBOR). HERE, MINI USES THE SOFT LABELING STRATEGY WITH $\lambda = 3$

	SEN	SPE	ACC	AUC	F_1	P. T.
Nearest neig.	82.3	84.2	83.6	89.1	75.6	0.17 ms
Areal [48]	84.0	86.3	85.6	91.6	78.2	0.20 ms
Trilinear	**86.6**	**87.3**	**87.1**	**92.8**	**80.6**	0.22 ms

network can be used for COVID-19 classification without any data augmentation.

B. Empirical Experiments on Design Options

We conduct an ablation study to evaluate the influence of the hyperparameter λ and different labeling strategies, and present the evaluation results in this section.

1) Hyperparameter λ: The hyperparameter λ controls the amount of information fused along the z-axis—a larger λ leads to a fusion of more slices for the per slice generation. To this end, the proposed MINI with different values of λ is assessed and the corresponding COVID-19 classification performance is presented in Table III. Here, the soft-labeling strategy is used to yield the labels for generated samples.

As we can see, the best COVID-19 classification performance is achieved by setting λ to 3, *i.e.,* fusing the information of four adjacent slices. The overall classification performance is observed to decrease as λ further increases. The underlying reason for the performance degradation may be the information loss while compacting the volume ($X \times Y \times 2Z$) to a smaller shape $\left(X \times Y \times \frac{2Z}{2^{(\lambda-1)}}\right)$.

2) Labeling Strategy: The results of MINI using the different labeling strategies are listed in Tabel IV. The 3D ResNet-18 trained without MINI is also involved for comparison.

It can be observed that using either hard or soft labeling, our MINI substantially improves the overall COVID-19 classification performance, compared to the baseline (w/o MINI). Since the soft labeling can better reflect the information contained in the generated volumes, *i.e.,* the similarity of $\tilde{v}$ to the negative and positive samples, it yields a larger improvement to the baseline (*i.e.,* $+5.9\%$ to F_1), compared to the hard labeling.[6] Nevertheless, the MINI with hard labeling yields a higher sensitivity than the one using soft labeling. The reason may be that the hard labeling strategy generates more positive samples (*i.e.,* $\tilde{l} = 1$ if l_A or l_B equals to 1) for the network to learn, which biases the model to make more positive predictions.

3) Interpolation Approach: As previously mentioned, our MINI adopts the trilinear interpolation to downsample the fused volume and the nearest neighbor to uniform the volume size. To analyze the influence caused by the selections, we evaluate the performance of the proposed MINI with different combinations of interpolation approaches. The widely-used areal interpolation [48] is also adopted for evaluation.

a) *Downsampling:*. We downsample the mixed volume to fuse the information of two authentic training samples. In this experiment, we maintain the nearest neighbor interpolation for upsampling and evaluate the performance of MINI with different interpolation approaches for information fusion. The experimental results are presented in Table V. The processing time for each interpolation approach is also evaluated. It can be observed that our MINI achieves the best COVID-19 classification performance using the trilinear interpolation, due to its better capacity of information fusion, with the similar processing time, compared to the others. Note that the processing time for each interpolation approach is tested with a single Tesla V100.

[6] In the following experiments, we report the COVID-19 classification performance of MINI with $\lambda = 3$ and soft-labeling.

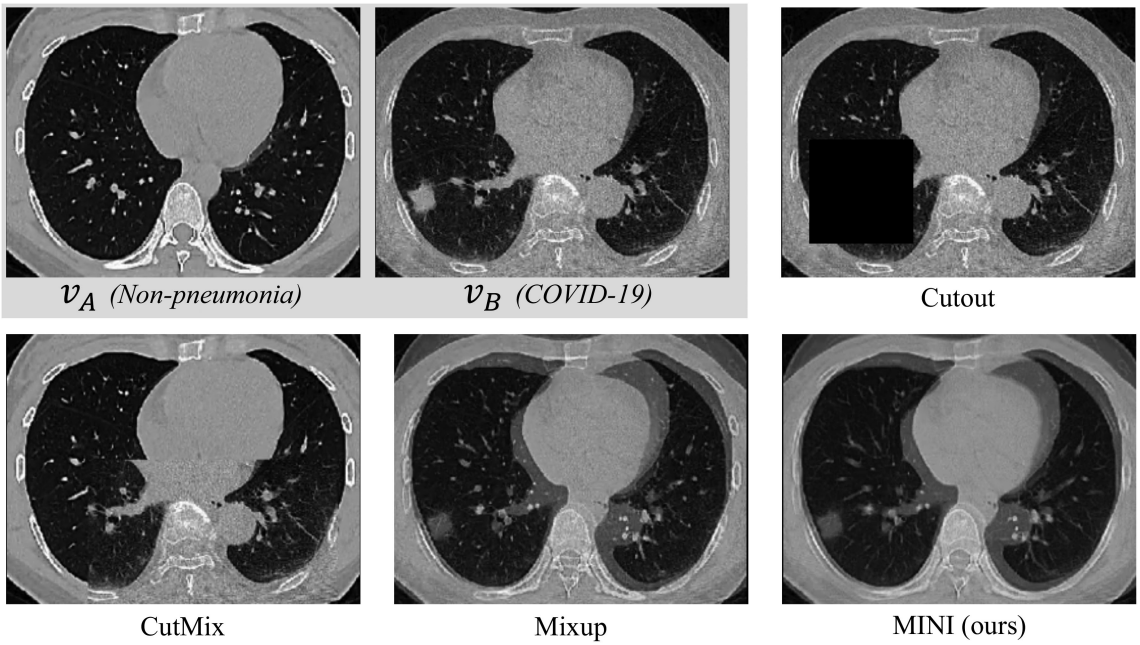


Fig. 3. Visualization of samples generated by different augmentation approaches. The samples generated by Cutout and CutMix face the risk of mistakenly discarding the lesion areas. Mixup and our MINI can alleviate the problem—maintaining the lesion areas via voxel-summation and trilinear interpolation, respectively.

TABLE VI
COVID-19 CLASSIFICATION PERFORMANCE (%) OF MINI USING DIFFERENT INTERPOLATION APPROACHES FOR SHAPE UNIFICATION (SEN–SENSITIVITY, SPE–SPECIFICITY, ACC–ACCURACY, AUC–AREA UNDER CURVE). HERE, MINI USES THE SOFT LABELING STRATEGY WITH $\lambda = 3$

	SEN	SPE	ACC	AUC	F_1
Trilinear	**86.8**	86.3	**88.0**	92.2	80.4
Areal [48]	86.3	87.0	86.8	**93.1**	80.1
Nearest neighbor	86.6	**87.3**	87.1	92.8	**80.6**

b) *Shape Unification:*. We also assess the performance of the proposed MINI with different interpolation approaches for shape unification. The experimental results are presented in Table VI. No significant difference between the accuracies achieved by different interpolation approaches is observed, which demonstrates that our MINI is insensitive to the choice of upsampling. Note that the trilinear interpolation is adopted for downsampling in this experiment.

C. Comparison with State-of-the-Art

In this section, the proposed MINI is compared with several state-of-the-art approaches, including Cutout, CutMix, and Mixup. For a fair comparison, we extend the typical 2D Cutout and CutMix to 3D versions, *i.e.,* cropping and replacing sub-volume for Cutout and CutMix, respectively.

1) Visualization of Generated Volumes: The volumes generated by different approaches are illustrated in Fig. 3. A non-pneumonia sample (v_A) and a COVID-19 sample (v_B) are adopted for volume generation. As shown in Fig. 3, volumes yielded by Cutout and CutMix may encounter the problem

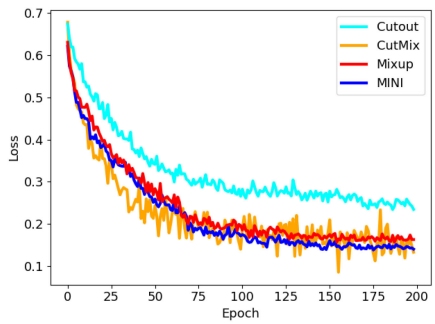


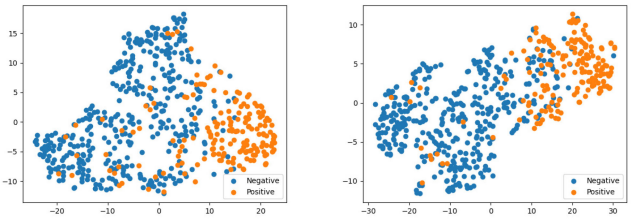
Fig. 4. Training curves for models using existing data augmentation approaches and our MINI. The model with MINI is observed to better converge (*i.e.,* more stable with the lower loss values).

of incomplete lesions (*i.e.,* the lesion area may be cropped or replaced), which makes the generated volume ($\tilde{v}$) contain inconsistent information with its label ($\tilde{l}$). For example, the Cutout still assigns the positive label (1) to $\tilde{l}$ even if the lesion area is cropped. In contrast, Mixup and our MINI can alleviate the problem by excellently maintaining the lesions. By comparing the volumes generated by Mixup and MINI, we can observe that the proposed MINI prefers to preserve the primary information and discard unnecessary details. In our experiments, the neural networks are observed to be easier to train and converge with such volumes. As shown in Fig. 4, the model using MINI has a better convergence (*i.e.,* lower training loss) than the one with Mixup.

2) Quantitative Comparison: It is worthwhile to mention that we repeat this experiment five times to reduce the influence caused by random nature of network training—such a setting

TABLE VII
COMPARISON OF COVID-19 CLASSIFICATION PERFORMANCE (%) WITH EXISTING AUGMENTATION APPROACHES (SEN–SENSITIVITY, SPE–SPECIFICITY, ACC–ACCURACY, AUC–AREA UNDER CURVE, HL–HARD LABELING, SL–SOFT LABELING)

	SEN	SPE	ACC	AUC	F_1
Baseline	77.0 ± 5.5	87.0 ± 2.9	83.9 ± 1.1	89.6 ± 1.1	74.7 ± 1.8
Cutout [43]	85.6 ± 2.1	85.0 ± 0.7	85.2 ± 0.5	91.1 ± 1.3	78.1 ± 0.9
CutMix [44]	85.9 ± 1.7	86.2 ± 1.3	86.1 ± 0.5	91.3 ± 0.3	79.2 ± 0.5
Mixup [42]	86.3 ± 0.6	86.3 ± 0.2	86.3 ± 0.3	92.2 ± 0.5	79.6 ± 0.5
Mixup *w.* **HL**	87.6 ± 0.9	83.8 ± 2.7	84.9 ± 1.7	92.1 ± 1.3	78.2 ± 1.8
MINI *w.* **HL**	$\mathbf{88.1 \pm 1.3}$	85.6 ± 1.5	86.3 ± 0.8	92.6 ± 0.6	79.9 ± 0.9
MINI *w.* **SL**	86.6 ± 2.2	$\mathbf{87.3 \pm 2.8}$	$\mathbf{87.1 \pm 1.5}$	$\mathbf{92.8 \pm 0.8}$	$\mathbf{80.6 \pm 1.5}$



(a) (b)

Fig. 5. Comparison of t-SNE between models without (a) and with MINI (b). Our MINI is observed to compact the cluster of negative in the feature space.

also enables us to conduct a statistical significance analysis of performance improvement. Hence, the results reported in the section are the average results of five repeated experiments with corresponding standard deviations. The performance of COVID-19 identification yielded by networks trained with different strategies are presented in Table VII. In general, the networks trained with generated volumes achieve consistent improvements for COVID-19 identification, compared to the baseline which is trained with only the original volumes. The inconsistent information contained in the volumes generated by Cutout and CutMix degrades their improvements. The network trained with our MINI with soft labeling achieves the best performance in most metrics (*i.e.,* 87.3%, 87.1%, 92.8% and 80.6% for SPE, ACC, AUC and F_1, respectively). Specifically, the F_1 of our MINI is $+1.0\%$ and $+5.9\%$ higher than the following-up (Mixup) and baseline, respectively.

Statistical Significance. A t-test validation is conducted using the results of repeated experiments to validate the statistical significance between our MINI and the following-up Mixup. A p-value of 0.039 is obtained, which indicates that the accuracy improvement (F_1) produced by our MINI is statistically significant at the 5% significance level.

To compare the feature representations learned by frameworks with and without MINI, we visualize the embedded features of the test set using t-SNE [49] in Fig. 5. It can be observed that the negative cluster becomes more compact using our MINI (Fig. 5(b)), compared to the baseline (Fig. 5(a)), which is easier for the classifier to draw the decision boundary.

3) Comparison Against Radiologists: We also conduct a human-machine confrontation to evaluate the potential of our deep learning framework for clinical applications. Since the labels for network training are provided by senior radiologists via comprehensive consideration of the RT-PCR results and clinical information, which is the upper limit of our model accuracy, a junior radiologist is invited to compete with our framework. A new test set, consisting of 150 negatives and 150 positives, is collected for human-machine confrontation. The experimental results are presented in Table VIII. It can be observed that the framework trained with our MINI outperforms the radiologist with all metrics, which demonstrates its potential for clinical application.

TABLE VIII
COMPARISON WITH RADIOLOGISTS ON AN INDEPENDENT TEST SET WITH 300 PATIENTS

	SEN	SPE	ACC	F_1
Radiologist	83.7	83.2	83.3	83.3
MINI (ours)	**86.3**	**84.0**	**83.9**	**85.0**

TABLE IX
DETAILED INFORMATION OF TRAINING SETS CONTAINING SAMPLES FROM DIFFERENT NUMBERS OF CENTRES

No. of centres	**3**	**4**	**5**	**6**	**7**
Hospital #1 ⋆	-	250	250	250	250
Hospital #2 ⋆	500	250	250	250	250
Hospital #3 ◇	-	-	-	-	50
Hospital #4 ◇	-	-	125	125	100
Hospital #5 ◇	250	250	125	125	100
Hospital #6 •	-	-	-	125	125
Hospital #7 •	250	250	250	125	125

⋆, ◇, • Indicate the COVID-19, CAP And Non-Pneumonia Respectively.

D. Performance With Different Numbers of Centres

To investigate the influence of multi-source biased data on the generalization capacity of deep learning models, we conduct an experiment evaluating the performance of the 3D ResNet-18 trained on data collected from different numbers of centres, while fixing the size of the training set to 1000 (500 COVID-19250 CAP and 250 non-pneumonia volumes). The training samples become more diverse as the number of centres increases. The detailed information about the training set composition with different numbers of centers is shown in Table IX.

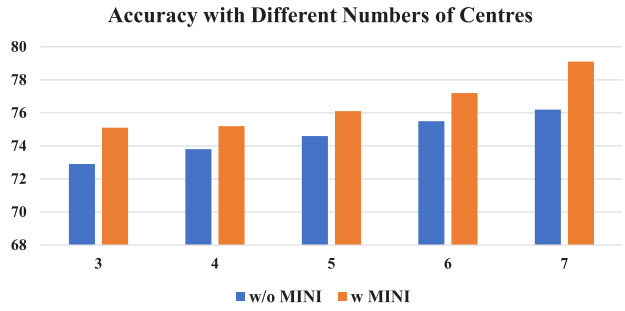


Fig. 6. The F_1 of COVID-19 identification yielded by networks trained with data collected from different numbers of centres. Here, soft labeling is adopted to yield the labels for generated volumes.

The evaluation results are presented in Fig. 6, where F_1 is adopted as the evaluation metric for the performance of COVID-19 identification. It can be observed that as the diversity of the training set increases, the source-biased problem is alleviated, *i.e.,* F_1 of the baseline (*i.e.,* without MINI) increases from around 73% to around 76% when data from four more centres is included.

Meanwhile, our MINI can effectively deal with the data source bias problem: an improvement of about $+2\%$ is achieved upon the baseline with the training set consisting of only three centres; in addition, MINI continues to substantially improve the F_1 score when more centres are included (*e.g.,* the F_1 score of the network trained with data from seven centres reaches 79% using MINI). Note that we use common augmentation approaches (*e.g.,* vertical flipping) to generate the equal numbers of additional samples to that with MINI for a fair comparison. The experimental results demonstrate the effectiveness of our MINI approach to the data source bias problem encountered in the practical implementation of CAD systems for COVID-19 diagnosis.

V. Discussion

Although human observers may not be able to distinguish the subtle differences between CT volumes captured at different centres, such differences may be explicit to DL networks. To gain insights into the data source bias problem, we try to train a deep learning network to classify the sources of data. The same training and test sets are used to train and evaluate the deep learning network on data source classification (*i.e.,* a single-label 7-class task). The experimental results show that the well-trained model can successfully identify the data sources, achieving an accuracy (ACC) of 96.40%. Therefore, the well-trained model can easily classify the positive/negative samples from a source biased dataset by their data originalities. Such a model often fails to generalize to an unseen domain (*i.e.,* a new centre) with data from all the disease categories. Taking a broader view, it can be easily seen that the data source bias problem may also happen for DL-based classification of diseases beyond COVID-19 which involves multi-centre data, and our proposed MINI can be widely used in such scenario for handy performance improvement, thus making a broader impact.

As revealed by our empirical experiment, the data source bias problem can be mitigated by increasing the number of data sources. As shown in Fig. 6, the F_1 score without MINI increases from around 73% to around 76% while keeping the numbers of training data (both total and class-specific) unchanged. We conjecture that this is because the addition of data from more centres narrows the domain gap between centres, to a limited extent. Nonetheless, our proposed MINI can further narrow down the domain gaps and thus improves the classification performance for all configurations of centres. The experimental results imply that simply adding data from more centres (which is often expensive or even impractical) is not enough for dealing with the source bias problem, and the proposed MINI can be an effective solution despite the number of centres involved.

We study two labeling strategies for the proposed MINI in this work, *i.e.,* hard versus soft labeling. As shown in Table IV, both strategies brought apparent general performance improvement upon the baseline, indicating effectiveness of MINI regardless of the labeling strategy. Meanwhile, hard labeling leads to a higher sensitivity than soft labeling, whereas the latter leads to better specificity and overall performance (indicated by higher AUC and F_1) than the former. The underlying reason may be that the hard labeling strategy probably assigns more positive labels to the generated volumes (the generated volume is labeled positive as long as at least one of v_A and v_B is positive), resulting in elevated sensitivity. A similar trend can be observed in Table VII, where the hard labeling strategy still achieves the highest sensitivity (and comparable values for other metrics) when compared to competing methods, and the soft labeling achieves the highest specificity, the best overall performance, and the second best sensitivity among the competing methods. We believe the decision on which labeling strategy to use is up to the practical need: if the extremely high sensitivity is required (with reasonable overall performance at the same time), then the hard labeling strategy should be preferred; otherwise, the soft labeling strategy can be picked for its more balanced performance.

Although existing approaches (*e.g.,* Mixup) can also alleviate the source bias problem, *i.e.,* yielding improvements upon the baseline, as shown in Table VII, the proposed MINI approach outperforms them by around $+1\%$. The underlying reason is that our MINI takes the cross slice information into account. The voxel-wise summation adopted by Mixup may decrease the contrast of lesion areas against the background tissue while performing a fusion of COVID-19 and normal volumes. In contrast, our MINI integrates the information of multiple slices via trilinear interpolation, which assists to construct a 3D view of lesion area and accordingly reduce the loss of lesion-related information when fusing COVID-19 and normal volumes. To validate the importance of lesion area for COVID-19 identification, we invite experienced radiologists to provide the ground truth of lesion areas and visualize the activation maps of neural network for COVID-19 samples. The visualization results are presented in Fig. 7. The annotations provided by radiologists are marked with red rectangles. The visualization results demonstrate that the neural network identifies COVID-19 patients mainly based on the lesion area.

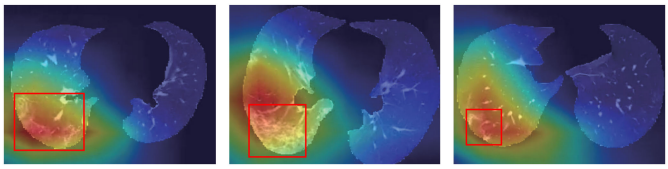


Fig. 7. Visualization of the activation maps of neural network for COVID-19 samples. The annotations provided by radiologists are drawn with red rectangles.

TABLE X
CLASSIFICATION ACCURACY (ACC %) OF A MULTI-CLASS CLASSIFICATION FRAMEWORK TRAINED WITH AND WITHOUT MINI

	Normal	CAP	COVID-19	Mean
w/o MINI	54.1	**73.9**	84.0	70.7
w MINI	**72.6**	61.2	**89.7**	**74.5**

As an immediate response to the COVID-19 pandemic, this work focuses on the differentiation of COVID-19 from both CAP and non-pneumonia cases. However, it would be more useful in clinical routine to develop a CAD system that can also distinguish non-pneumonia and CAP cases. In this regard, we train a multi-class classification framework and report its performance on the test set in Table X. The mean accuracy of the model using our MINI is 74.5%, which is $+3.8\%$ higher than vanilla one.

Last but not least, we identify limitations of this work that warrant future work. Consistent to the previous works [42], [44], the proposed MINI is implemented to fuse the information of two volumes. However, we notice that our MINI can be further expanded to deal with multiple (more than two) samples, which results in a higher diversity of generated volumes. To this end, we will try to improve the performance of the proposed MINI along this direction in our future study. Meanwhile, interpretability of deep learning models [50] is drawing increasing attention of the community of medical image analysis, which can provide evidences to back up as well as boost users' confidence in the networks' predictions. Although the class activation maps in Fig. 7 provide some clues illustrating the mechanism of our model for COVID-19 identification, we plan to deeply explore along this direction and further enhance the model interpretability in the future.

VI. CONCLUSION

In this work, we looked into the *data source bias* problem when using multi-centre data for training deep neural networks for COVID-19 classification. Subject to practical conditions, a certain centre may provide samples of only a single class (*e.g.,* COVID-19), while another centre may provide those of another class (*e.g.,* non-pneumonia). Training with such data without any countermeasure may potentially bias the trained networks towards identifying sources of data instead of the pneumonia. Despite being practically relevant, this problem had been overlooked until this work, which presented MIx-aNd-Interpolate (MINI)—a conceptually simple, computationally efficient, and effective training strategy dealing with the source bias problem. Specifically, MINI generated volumes of the absent class by combining samples collected from different centres, which enlarged the sample space of the original source-biased dataset, and worked as a lightweight online data augmentation strategy. The experimental results on a large-scale real-patient dataset composed of samples collected from eight medical centres demonstrated 1) the effectiveness of the proposed MINI in dealing with the data source bias and thus improving classification performance upon the baseline, and 2) its superiority to several recently proposed strategies that can be employed against the source bias problem. In the future, we expect MINI to be broadly applied to training classification networks for other diseases beyond COVID-19.

REFERENCES

[1] T. Ai *et al.*, "Correlation of chest CT and RT-PCR testing in cronavirus disease 2019 (COVID-19) in China: A report of 1014 cases," *Radiology*, vol. 296, no. 2, pp. E32–E40, 2020, Art. no. 200642.

[2] Y. Fang *et al.*, "Sensitivity of chest CT for COVID-19: Comparison to RT-PCR," *Radiology*, vol. 296, no. 2, pp. E115–E117, 2020, Art. no. 200432.

[3] X. Xie, Z. Zhong, W. Zhao, C. Zheng, F. Wang, and J. Liu, "Chest CT for typical coronavirus disease 2019 (COVID-19) pneumonia: Relationship to negative RT-PCR testing," *Radiology*, vol. 296, no. 2, pp. E41–E45, 2020, Art. no. 200343.

[4] C. Huang *et al.*, "Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China," *Lancet*, vol. 395, no. 10223, pp. 497–506, 2020.

[5] A. Bernheim *et al.*, "Chest CT findings in coronavirus disease-19 (COVID-19): Relationship to duration of infection," *Radiology*, vol. 295, no. 3, 2020, Art. no. 200463.

[6] C. Butt, J. Gill, D. Chun, and B. A. Babu, "Deep learning system to screen coronavirus disease 2019 pneumonia," *Appl. Intell.*, vol. 6, no. 10, pp. 1–7, 2020.

[7] L. Li *et al.*, "Artificial intelligence distinguishes COVID-19 from community acquired pneumonia on chest CT," *Radiology*, vol. 296, no. 2, pp. 65–71, 2020.

[8] K. Zhang *et al.*, "Clinically applicable AI system for accurate diagnosis, quantitative measurements, and prognosis of COVID-19 pneumonia using computed tomography," *Cell*, vol. 181, no. 6, pp. 1423–1433, 2020.

[9] Z. Han *et al.*, "Accurate screening of COVID-19 using attention based deep 3D multiple instance learning," *IEEE Trans. Med. Imag.*, vol. 39, no. 8, pp. 2584–2594, Aug. 2020.

[10] X. Ouyang *et al.*, "Dual-sampling attention network for diagnosis of COVID-19 from community acquired pneumonia," *IEEE Trans. Med. Imag.*, vol. 39, no. 8, pp. 2595–2605, Aug. 2020.

[11] H. Kang *et al.*, "Diagnosis of coronavirus disease 2019 (COVID-19) with structured latent multi-view representation learning," *IEEE Trans. Med. Imag.*, vol. 39, no. 8, pp. 2606–2614, Aug. 2020.

[12] S. Roy *et al.*, "Deep learning for classification and localization of COVID-19 markers in point-of-care lung ultrasound," *IEEE Trans. Med. Imag.*, vol. 39, no. 8, pp. 2676–2687, Aug. 2020.

[13] Y. Oh, S. Park, and J. C. Ye, "Deep learning COVID-19 features on CXR using limited training data sets," *IEEE Trans. Med. Imag.*, vol. 39, no. 8, pp. 2688–2700, Aug. 2020.

[14] O. Gozes *et al.*, "Rapid AI development cycle for the coronavirus (COVID-19) pandemic: Initial results for automated detection & patient monitoring using deep learning CT image analysis," 2020, *arXiv:2003.05037*.

[15] Y. Li *et al.*, "Efficient and effective training of COVID-19 classification networks with self-supervised dual-track learning to rank," *IEEE J. Biomed. Health Informat.*, vol. 24, no. 10, pp. 2787–2797, Oct. 2020.

[16] X. Wang *et al.*, "A weakly-supervised framework for COVID-19 classification and lesion localization from chest CT," *IEEE Trans. Med. Imag.*, vol. 39, no. 8, pp. 2615–2625, Aug. 2020.

[17] D.-P. Fan *et al.*, "Inf-Net: Automatic COVID-19 lung infection segmentation from CT images," *IEEE Trans. Med. Imag.*, vol. 39, no. 8, pp. 2626–2637, Aug. 2020.

[18] B. Sun, J. Feng, and K. Saenko, "Return of frustratingly easy domain adaptation," in *Proc. AAAI Conf. Artif. Intell.*, 2016, pp. 2058–2065.

[19] Y. Ganin *et al.*, "Domain adversarial training of neural networks," *J. Mach. Learn. Res.*, vol. 17, no. 1, pp. 2096–2030, 2017.
[20] J. Hoffman, E. Tzeng, T. Park, and J.-Y. Zhu, "CyCADA: Cycle-consistent adversarial domain adaptation," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 1989–1998.
[21] S. Sankaranarayanan, Y. Balaji, C. D. Castillo, and R. Chellappa, "Generate to adapt: Aligning domains using generative adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 8503–8512.
[22] S. Huang, C. Lin, S. Chen, Y. Wu, P. Hsu, and S. Lai, "AugGAN: Cross domain adaptation with GAN-based data augmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 718–731.
[23] M. H. Chen, Z. Kira, G. AlRegib, J. Woo, R. Chen, and J. Zheng, "Temporal attentive alignment for large-scale video domain adaptation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 6320–6329.
[24] W. Deng, L. Zheng, Q. Ye, G. Kang, Y. Yang, and J. Jiao, "Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 994–1003.
[25] L. Wei, S. Zhang, W. Gao, and Q. Tian, "Person transfer GAN to bridge domain gap for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 79–88.
[26] F. G. Zanjani, S. Zinger, B. E. Bejnordi, J. A. W. M. van der Laak, and P. H. N. de With, "Stain normalization of histopathology images using generative adversarial networks," in *Proc. IEEE Int. Symp. Biomed. Imag.*, 2018, pp. 573–577.
[27] Y. Gao, Y. Liu, Y. Wang, Z. Shi, and J. Yu, "A universal intensity standardization method based on a many-to-one weak-paired cycle generative adversarial network for magnetic resonance images," *IEEE Trans. Med. Imag.*, vol. 38, no. 9, pp. 2059–2069, Sep. 2019.
[28] Z. Zhang, L. Yang, and Y. Zheng, "Translating and segmenting multimodal medical volumes with cycle- and shape-consistency generative adversarial network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 9242–9251.
[29] X. Peng, Q. Bai, X. Xia, Z. Huang, K. Saenko, and B. Wang, "Moment matching for multi-source domain adaptation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 1406–1415.
[30] K. Muandet, D. Balduzzi, and B. Schölkopf, "Domain generalization via invariant feature representation," in *Proc. Int. Conf. Mach. Learn.*, 2013, pp. 10–18.
[31] M. Ghifary, W. Bastiaan Kleijn, M. Zhang, and D. Balduzzi, "Domain generalization for object recognition with multi-task autoencoders," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 2551–2559.
[32] S. Motiian, M. Piccirilli, D. A. Adjeroh, and G. Doretto, "Unified deep supervised domain adaptation and generalization," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 5715–5725.
[33] Y. Li *et al.*, "Deep domain generalization via conditional invariant adversarial networks," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 624–639.
[34] H. Li, S. Jialin Pan, S. Wang, and A. C. Kot, "Domain generalization with adversarial feature learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 5400–5409.
[35] T. Hospedales, A. Antoniou, P. Micaelli, and A. Storkey, "Meta-learning in neural networks: A survey," 2020, *arXiv:2004.05439*.
[36] Y. Balaji, S. Sankaranarayanan, and R. Chellappa, "MetaReg: Towards domain generalization using meta-regularization," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 998–1008.
[37] D. Li, J. Zhang, Y. Yang, C. Liu, Y.-Z. Song, and T. M. Hospedales, "Episodic training for domain generalization," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 1446–1455.
[38] Q. Dou, D. C. de Castro, K. Kamnitsas, and B. Glocker, "Domain generalization via model-agnostic learning of semantic features," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 6450–6461.
[39] S. Shankar, V. Piratla, S. Chakrabarti, S. Chaudhuri, P. Jyothi, and S. Sarawagi, "Generalizing across domains via cross-gradient training," 2018, *arXiv:1804.10745*.
[40] R. Volpi, H. Namkoong, O. Sener, J. C. Duchi, V. Murino, and S. Savarese, "Generalizing to unseen domains via adversarial data augmentation," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 5334–5344.
[41] F. M. Carlucci, A. D'Innocente, S. Bucci, B. Caputo, and T. Tommasi, "Domain generalization by solving Jigsaw puzzles," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 2229–2238.
[42] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "Mixup: Beyond empirical risk minimization," 2017, *arXiv:1710.09412*.
[43] T. DeVries and G. W. Taylor, "Improved regularization of convolutional neural networks with cutout," 2017, *arXiv:1708.04552*.
[44] S. Yun, D. Han, S. J. Oh, S. Chun, J. Choe, and Y. Yoo, "CutMix: Regularization strategy to train strong classifiers with localizable features," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 6022–6031.
[45] Q. Dou *et al.*, "3D deeply supervised network for automated segmentation of volumetric medical images," *Med. Image Anal.*, vol. 41, pp. 40–54, 2017.
[46] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.
[47] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.
[48] A. Comber and W. Zeng, "Spatial interpolation using areal features: A review of methods and opportunities using new forms of data with coded illustrations," *Geography Compass*, vol. 13, no. 10, 2019, Art. no. e12465.
[49] L. Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, 2008.
[50] S. Chakraborty *et al.*, "Interpretability of deep learning models: A survey of results," in *Proc. IEEE Smartworld, Ubiquitous Intell. Comput., Adv. Trusted Computed, Scalable Comput. Commun., Cloud Big Data Comput., Internet People Smart City Innov.*, 2017, pp. 1–6.