RCoNet: Deformable Mutual Information Maximization and High-Order Uncertainty-Aware Learning for Robust COVID-19 Detection

Shunjie Dong, Qianqian Yang, *Member, IEEE*, Yu Fu, Mei Tian, and Cheng Zhuo, *Senior Member, IEEE*

Abstract—The novel 2019 Coronavirus (COVID-19) infection has spread worldwide and is currently a major healthcare challenge around the world. Chest computed tomography (CT) and X-ray images have been well recognized to be two effective techniques for clinical COVID-19 disease diagnoses. Due to faster imaging time and considerably lower cost than CT, detecting COVID-19 in chest X-ray (CXR) images is preferred for efficient diagnosis, assessment, and treatment. However, considering the similarity between COVID-19 and pneumonia, CXR samples with deep features distributed near category boundaries are easily misclassified by the hyperplanes learned from limited training data. Moreover, most existing approaches for COVID-19 detection focus on the accuracy of prediction and overlook uncertainty estimation, which is particularly important when dealing with noisy datasets. To alleviate these concerns, we propose a novel deep network named $RCoNet_s^k$ for robust COVID-19 detection which employs Deformable Mutual Information Maximization (DeIM), Mixed High-order Moment Feature (MHMF), and Multiexpert Uncertainty-aware Learning (MUL). With DeIM, the mutual information (MI) between input data and the corresponding latent representations can be well estimated and maximized to capture compact and disentangled representational characteristics. Meanwhile, MHMF can fully explore the benefits of using high-order statistics and extract discriminative features of complex distributions in medical imaging. Finally, MUL creates multiple parallel dropout networks for each CXR image to evaluate uncertainty and thus prevent performance degradation caused by the noise in the data. The experimental results show that $RCoNet_s^k$ achieves the state-of-the-art performance on an open-source COVIDx dataset of 15 134 original CXR images across several metrics. Crucially, our method is shown to be more effective than existing methods with the presence of noise in the data.

Index Terms—Chest X-rays (CXRs), COVID-19, deformable mutual information maximization (DeIM), mixed high-order moment feature (MHMF), multiexpert uncertainty-aware learning (MUL), noisy data, $RCoNet_s^k$, uncertainty.

I. INTRODUCTION

CORONAVIRUS disease 2019 (COVID-19) causes an ongoing pandemic that significantly impacts everyone's life since it was first reported, with hundreds of thousands of deaths and millions of infections emerging in over 200 countries [1], [2]. As indicated by the World Health Organization (WHO), due to its highly contagious nature and lack of corresponding vaccines, the most effective method to control the spread of COVID-19 infection is to keep social distance and contact tracing. Hence, early and fast diagnosis of COVID-19 has become significantly essential to control further spreading, and such that the patients could be hospitalized and receive proper treatment in time.

Since the emergence of COVID-19, reverse transcription polymerase chain reaction (RT-PCR), as a viral nucleic acid detection method by gene sequencing, is the accepted standard for COVID-19 detection [3]. However, because of the low accuracy of RT-PCR and limited medical test kits in many hyperendemic regions or countries, it is challenging to detect every individual affected by COVID-19 rapidly [4], [5]. Therefore, alternative testing methods, which are faster and more reliable than RT-PCR, are urgently needed to combat the disease.

Since most COVID-19 positive patients were diagnosed with pneumonia, radiological examinations could help detect and assess the disease. Recently, chest computed tomography (CT) has been shown to be efficient and reliable to achieve a real-time clinical diagnosis of COVID-19, outperforming RT-PCR in terms of accuracy. Moreover, some deep learning-based methods have been proposed for COVID-19 detection using chest CT images [6]–[9]. For example, an adaptive feature selection approach was proposed in [10] for COVID-19 detection based on a trained deep forest model. In [11], an uncertainty vertex-weighted hypergraph learning method was designed to identify COVID-19 from community-acquired pneumonia (CAP) using CT images. However, the routine use of CT, which is conducted via expensive equipment, takes considerably more time than X-ray imaging and brings a massive burden on radiology departments. Compared to CT, X-rays could significantly speed up disease screening, and hence become a preferred method for disease diagnosis.
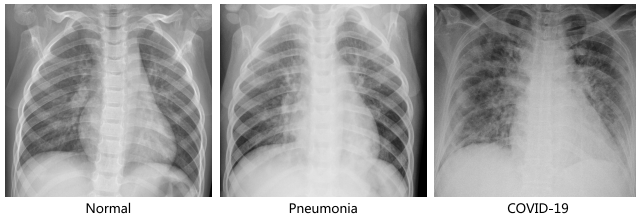
Fig. 1. Visual illustration of CXR images, including normal, pneumonia, and COVID-19.

Accordingly, deep learning-based methods for detecting COVID-19 with chest X-ray (CXR) have been developed and shown to be able to achieve accurate and speedy detection [12], [13]. For instance, a tailored convolution neural network platform trained on open-source dataset called COVIDNet in [14] was proposed for the detection of COVID-19 cases from CXR. Oh *et al.* [15] proposed a novel probabilistic gradient-weighted class activation map to enable infection segmentation and detection of COVID-19 on CXR images. Fig. 1 shows three samples from the COVIDx dataset [14] which contains three different classes: normal, pneumonia, and COVID-19. However, due to a similar pathological information between pneumonia and COVID-19 in the early stage, the CXR samples may have latent features distributed near the category boundaries, which can be easily misclassified by the hyperplane learned from the limited training data. Moreover, to the best of our knowledge, most of the existing methods for COVID-19 detection were designed to extract the lower-dimension latent representations which may not be able to fully capture statistical characteristic of complex distributions (i.e., non-Gaussian distribution presented in CXR images). Furthermore, quantifying uncertainty in COVID-19 detection is still a major yet challenging task for existing deep networks, especially with the presence of noise in the training samples (i.e., label noise and image noise).

To address the above problems, we propose a novel deep network architecture, referred to as $RCoNet_s^k$, for robust COVID-19 detection which, in particular, contains the following three modules, i.e., Deformable mutual Information Maximization (DeIM), Mixed High-order Moment Feature (MHMF) and Multiexpert Uncertainty-aware Learning (MUL):

1) The DeIM module estimates and maximizes the mutual information (MI) between input data and learned high-level representations, which pushes the model to learn the discriminative and compact features. We employ deformable convolution layers in this module which are able to explore disentangled spatial features and mitigate the negative effect of similar samples across different categories.
2) The MHMF module fully explores the benefits of using a mix of high-order moment statistics to better characterize the feature distributions in medical imaging and reduce the negative effects of noise.
3) The MUL creates multiple parallel dropout networks, each can be treated as an expert, to derive multiple experts-based diagnosis similar to clinical practices,

which improves the prediction accuracy. MUL also quantifies the prediction accuracy by obtaining the variance in prediction across different experts.
4) The experimental results show that our proposal achieves the state-of-the-art performance in terms of most metrics both on open source COVIDx dataset of 15134 original CXR images and that of noisy setting.

The remaining of this article is organized as follows: In Section II, we review related works on MI estimation and uncertainty learning as well. In Section III, after an overview of our proposed approach, we discuss the main components of $RCoNet_s^k$. In Section IV, we compare our proposed architecture with the existing deep learning-based methods evaluated on a publically available dataset of CXR images and also the same dataset but under noisy conditions. And we also conduct extensive experiments to demonstrate the benefits of DeIM, MHMF, and MUL on the system's performance. Finally, we conclude this article in Section V.

II. BACKGROUND AND RELATED WORKS

In this section, we introduce related works on MI estimation and uncertainty learning that lay the foundation of this article.

A. MI Estimation

MI, as a fundamental concept in information theory, is widely applied to unsupervised feature learning for quantifying the correlation between random variables. MI has been exploited in a wide range of domains and tasks, including biomedical sciences [16], blind source separation (BSS, e.g., independent component analysis [17]), feature selection [18], [19], and causal inference [20]. For example, the object tracking task considered in [21] was treated as a problem of optimizing the MI between features extracted from a video with most color information removed and those from the original full-color video. Closely related work presented in [22] considered learning representations to predict cross-modal correspondence by maximizing MI between features from the multiview encoders and the content of the held-out view. Moreover, Mutual Information Neural Estimation (MINE) proposed by [23] was designed to learn a general-purpose estimator of the MI between continuous variables based on dual representations of the Kullback-Leibler (KL)-divergence, which are scalable, flexible, and, most crucially, trainable via back-propagation. Inspired by MINE, our proposal estimates and maximizes the CXR image inputs and the corresponding latent representations to improve diagnosis performance.

B. Uncertainty in Deep Learning

Aiming at combating the significant negative effects of uncertainty in deep neural networks, uncertainty learning has been getting lots of research attention, which facilitates the reliability assessment and solves risk-based decision-making problems [24]–[26]. In recent years, various frameworks have been proposed to characterize the uncertainty in the model parameters of deep neural networks, referred to as *model*
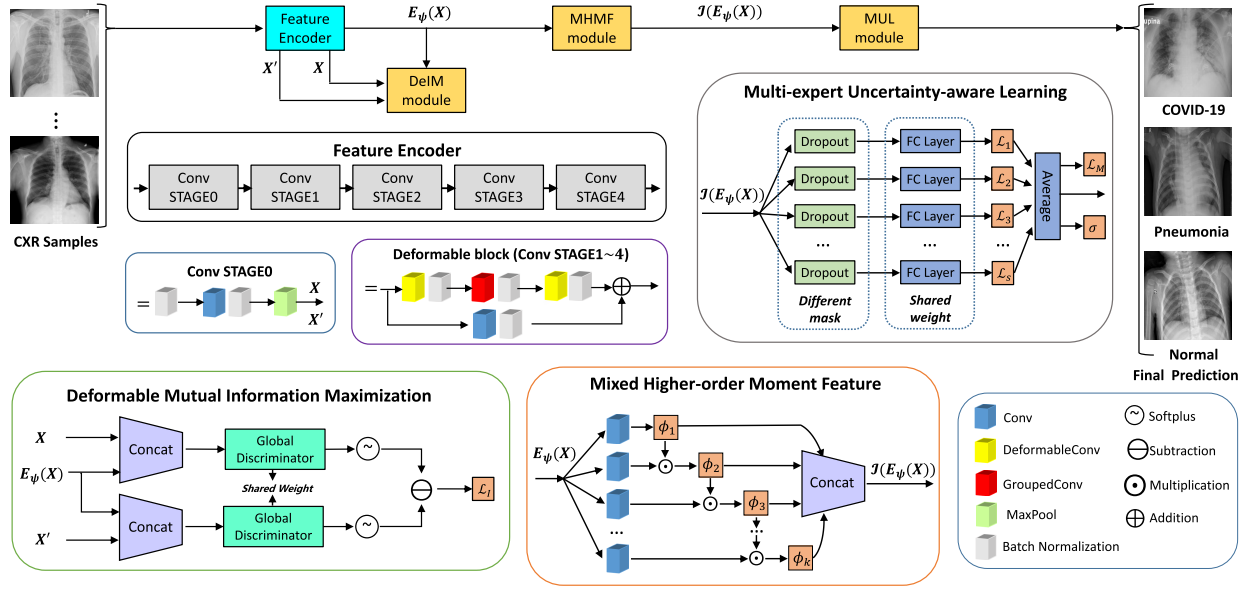
Fig. 2.   Architecture of RCoNet$_s^k$ for COVID-19 detection.

uncertainty, due to the limited size of training data [27], [28], which can be reduced by collecting more training data [25], [29], [30]. Meanwhile, another kind of uncertainty in deep learning, referred to as *data uncertainty*, measures the noise inherent in given training data, and hence cannot be eliminated by having more training data [31]. To combat these two kinds of uncertainty, lots of works on various computer vision tasks, i.e., face recognition [24], semantic segmentation [32], object detection [33], and person reidentification [34], have introduced deep uncertainty learning to improve the robustness of deep learning model and interpretability of discriminant. For face recognition task in [25], an uncertainty-aware probabilistic face embedding (PFE) was proposed to represent face images as distributions by utilizing data uncertainty. Exploiting the advantage of Bayesian deep neural networks, one recent study [35] leveraged the model uncertainty for analysis and learning of face representations. To our knowledge, our proposal is the first work that utilizes the high-order moment statistics and multiple expert networks to estimate uncertainty for COVID-19 detection using CXR images.

III. METHOD

In this section, we introduce the novel RCoNet$_s^k$ for robust COVID-19 detection, which incorporates DeIM, MHMF, and MUL, as illustrated in Fig. 2. k is the number of levels of moment features that are combined in MHMF, and s is the number of the expert network in MUL, which will be further clarified in the sequel. The CXR images are first processed by DeIM which consists of a stack of deformable convolution layers, extracting discriminative features. The compact features are then fed into MHMF module to generate mixed high-order moment latent features, reducing negative effects caused by similar images and noise. The proposed MUL utilizes the learned high-order features to generate final diagnoses.

A. Deformable Mutual Information Estimation and Maximization

Due to the similarity between COVID-19 and pneumonia in the latent space, we propose DeIM to extract discriminative and informative features, reducing the negative influence caused by the lack of distinctiveness in the deep features. In particular, we train the model by maximizing the MI between the input and corresponding latent representation.

We use a stack of five convolutional stages, as shown in Fig. 2, to encode inputs into latent representations, which is denoted by a differentiable parametric function E_ψ

$$E_\psi : \mathcal{X} \to \mathcal{Z} \tag{1}$$

where ψ denotes the set of all the trainable parameters in these layers, and $\mathcal{X}$ and $\mathcal{Z}$ denote the input and output spaces, respectively.

The detailed architecture of each convolutional stage is presented in Fig. 2, which consists of several convolutional layers each followed by a batch normalization layer. Note that we employ deformable convolutional layers which can better extract spatial information of the irregular infected area compared to conventional convolutional layers. More specifically, regular convolution operates on predefined rectangular grid from an input image or a set of input feature maps, while the deformable convolution operates on deformable grids that each grid point is moved by a learnable offset. For example, the receptive grid $\mathcal{P}$ of a regular convolution with kernel size 3×3 is fixed and can be given by

$$\mathcal{P} = \{(-1, -1), (-1, 0), \dots, (0, 1), (1, 1)\} \tag{2}$$

while, for deformable convolution, the receptive grid is moved by the learned offsets $\Delta p_n \in \mathbb{R}^2$ and the output is given as follows:

$$b(p_0) = \sum_{P_n \in \mathcal{P}} w(p_n) \cdot a(p_0 + p_n + \Delta p_n) \tag{3}$$

where $b(p_0)$ denotes the value at location p_0 on the output feature map b, p_n enumerates the locations in $\mathcal{P}$, $w(p_n)$ represents the weight at location p_n of the kernel, and $a(\cdot)$ is value at given location on the input feature map. We can see that with the introduction of offsets Δp_n, the receptive grid is no longer fixed to be a rectangle, and instead is deformable.

We optimize E_ψ by maximizing the MI between the input and the output, i.e., $I(X; Z)$, where $Z \triangleq E_\psi(X)$. The precise MI requires knowledge probability density functions (PDFs) of X and Z, which is intractable to obtain in practice. To overcome this issue, MINE proposed in [23] estimates MI by using a lower-bound on the Donsker-Varadhan representation [36] of the KL-divergence

$$
\begin{aligned}
I(X; Z) &:= D_{KL}(\mathbb{J} || \mathbb{M}) \geq \widehat{I}_\theta^{(DV)}(X; Z) \\
&:= \mathbb{E}_\mathbb{J}[T_\theta(x, z)] - \log \mathbb{E}_\mathbb{M}\big[e^{T_\theta(x,z)}\big]
\end{aligned}
\tag{4}
$$

where $\mathbb{J}$ represents the joint probability of X and Z, i.e., $\mathbb{J} \triangleq P(X, Z)$, and $\mathbb{M}$ denotes the product of marginal probabilities of X and Z, $\mathbb{M} \triangleq P(X)P(Z)$. $T_\theta : \mathcal{X} \times \mathcal{Z} \to \mathbb{R}$ denotes a *global discriminator* modeled by a neural network with parameters θ, which is trained to maximize $\widehat{I}_\theta^{(DV)}(X; Z)$ to approximate the actual MI. Hence, we can simultaneously estimate and maximize $I(X; E_\psi(X))$ by maximizing $\widehat{I}_\theta^{(DV)}(X; Z)$

$$
(\widehat{\theta}, \widehat{\psi}) = \arg\max_{\theta, \psi} \ \widehat{I}_\theta^{(DV)}(X; E_\psi(X)).
\tag{5}
$$

Since the encoder E_ψ and the MI estimator T_θ are optimized simultaneously with the same objective function, we can share some layers between them, and replace the T_θ with $T_{\theta,\psi}$ to account for this fact.

Since we are primarily interested in maximizing the MI rather than estimating the precise value, we can alternatively use a Jensen-Shannon MI estimator (JSD) [37], which offers more interpretable tradeoff

$$
\begin{aligned}
\widehat{I}_{\theta,\psi}^{(DeJSD)}(X; E_\psi(X)) := &\ \mathbb{E}_\mathbb{P}\big[-\log\big(1 + e^{-T_{\theta,\psi}(x, E_\psi(x))}\big)\big] \\
&- \mathbb{E}_{\mathbb{P} \times \widetilde{\mathbb{P}}}\Big[\log\Big(1 + e^{T_{\theta,\psi}(x', E_\psi(x))}\Big)\Big]
\end{aligned}
\tag{6}
$$

where x is an input sample of an empirical probability distribution $\mathbb{P}$, x' denotes a fake sample from distribution $\widetilde{\mathbb{P}}$, where $\widetilde{\mathbb{P}} = \mathbb{P}$. This estimator is illustrated by the DeIM block shown in Fig. 2, which has the latent representation $E_\psi(x)$, the input sample x and the fake sample x' as input, and the difference between the outputs of the two softplus operations as the estimation of MI.

Another alternative MI estimator is called Noise-Contrastive Estimator (NCE) [38], which is defined as

$$
\begin{aligned}
&\widehat{I}_{\theta,\psi}^{(DeNCE)}(X; E_\psi'(X)) \\
&:= \mathbb{E}_\mathbb{P}\Bigg[T_{\theta,\psi}(x, E_\psi'(x)) - \mathbb{E}_{\widetilde{\mathbb{P}}}\Bigg[\log \sum_{x'} e^{T_{\theta,\psi}(x', E_\psi'(x))}\Bigg]\Bigg].
\end{aligned}
\tag{7}
$$

The experiments have found that using the NCE estimator outperforms the JSD estimator in some cases, but appears to be quite similar most of the time.

The existing works [39] that implement these estimators use some latent representation of x, which is then merged with some randomly generated features to obtain "fake" samples
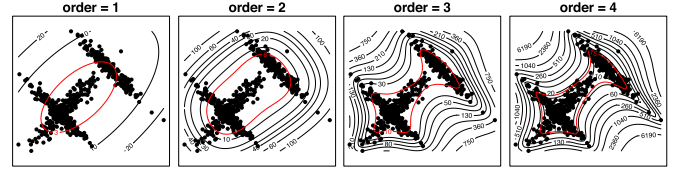

Fig. 3. Data points from three Gaussian distributions and the corresponding moment feature of order 1 to 4.

that satisfy $\mathbb{P} = \widetilde{\mathbb{P}}$. In contrast, we use the samples from other categories as the "fake" samples, i.e., x', instead. For example, if the input is a pneumonia sample, then the fake sample is either a normal or COVID sample. We note that this can push the learned encoder to derive more distinguishable features for samples from different categories.

B. Mixed High-Order Moment Feature

The presence of the image noise and label noise in CXR datasets may cause image latent representations generated by deep neural networks to be scattered in the entire feature space. To deal with this issue, [24], [25], [34] represent each image as a Gaussian distribution, that is defined by a mean (a standard feature vector) and a variance. However, the deep features of CXR samples we considered in this article typically follow a complex, non-Gaussian distribution [40], [41], which cannot be fully captured by its first-order (mean) or second-order statistics (variance).

We seek a better combination of different orders of statistics to more precisely characterize the latent representation of the CXR images. We illustrate the moment features of different orders [42] in Fig. 3, where we plot 350 data points in $\mathbb{R}^2$ sampled from a distribution that combines three different Gaussian distributions. We can observe that the high-order moment features are more expressive of statistical characteristic compared to low-order one. More specifically, they capture the shape of the cloud of samples more accurately. Therefore, we include the MHMF module in the proposed model, as shown in Fig. 2, which outputs a combination of high-order moment features with the latent representation $E_\psi(X)$ as input. This will potentially solve the scattering problem and capture the subtle differences between CXR images of similar categories, i.e., pneumonia and COVID-19 in our case.

We show how to obtain the complicated high-order moment feature in the following. Define rth order moment feature as $\phi_r(\boldsymbol{a})$, where $\boldsymbol{a} \in \mathbb{R}^{H \times W \times C}$ denotes a latent feature map of dimension $H \times W \times C$. Lots of recent works adopt the Kronecker product to compute high-order moment feature [41]. However, calculating Kronecker product of high dimensional feature maps is significantly computational intensive, and hence infeasible for real-world applications. Inspired by [43]–[45], we approximate $\phi_r(\boldsymbol{a})$ by exploiting r random projectors which rely on certain factorization schemes, such as Random Maclaurin [46]. We use 1×1 convolution kernels as the random projectors to estimate the expectations of high-order moment features. That is

$$
\phi_r(\boldsymbol{a}) \approx \mathcal{K}_1(\boldsymbol{a}) \odot \mathcal{K}_2(\boldsymbol{a}) \odot \cdots \odot \mathcal{K}_r(\boldsymbol{a}) \in \mathbb{R}^{H \times W \times C}
\tag{8}
$$

where $\odot$ represents the Hadamard (element-wise) product, and $\mathcal{K}_1, \mathcal{K}_2, \ldots, \mathcal{K}_r$ are 1×1 convolution kernels with random weights.

Note that Random Maclaurin produces an estimator that is independent of the input distribution, which causes the estimated high-order moments to contain noninformative high-order moment components. We eliminate these components by learning the weights of the projectors, i.e., the 1×1 convolution kernels, from the data. Also, note that the Hadamard product of a number of random projectors may end up with the estimated high-order moment features to be similar to low-order ones. To solve this problem, we use a recursive way to estimate the high-order moments instead

$$\phi_r(\boldsymbol{a}) = \phi_{r-1}(\boldsymbol{a}) \odot \mathcal{K}_r(\boldsymbol{a}). \tag{9}$$

Since different order moments capture different informative statistics, we design the MHMF module to keep the estimated moments of different levels of order, as shown in Fig. 2, the output of which is given as

$$\mathcal{J}(\boldsymbol{a}) = [\phi_1(\boldsymbol{a}); \phi_2(\boldsymbol{a}); \ldots; \phi_r(\boldsymbol{a})] \in \mathbb{R}^{H \times W \times rC}. \tag{10}$$

Hence, $\mathcal{J}(\boldsymbol{a})$ is rich enough to capture the complicated statistics, and produce discriminative features for the input of different categories.

C. Multiexpert Uncertainty-Aware Learning

The MHMF module, as described in Section III-B, generates MHMFs of each sample in the latent space, which we aim to further exploit to derive compact and disentangled information for COVID-19 detection. Meanwhile, quantifying uncertainty in disease detection is undoubtedly significant to understand the confidence level of computer-based diagnoses. Motivated by the clinical practices, we present a novel neural network in this section, referred to as MUL, which takes in the MHMFs and outputs the prediction and the quantification of the diagnostic uncertainty caused by the noise in the data.

The structure of the MUL module is shown in Fig. 2, which consists of multiple dropout layers that process the output from MHMF in parallel, each of which together with the following several fully connected layers can be regarded as an expert for COVID-19 detection. We note that each dropout layer uses different masks which results in different subsets of latent information to be kept, while the following fully connected layers share the same weights across different experts. The masks for the dropout layers are generated randomly at each iteration during training but fixed during the inference time. We denote the input-output function of each expert by $C_e^j(\cdot)$, $j = 1, \ldots, N$, where N is the total number of experts. Hence, we have the classification loss $\mathcal{L}_e^j$ of jth expert given as follows:

$$\mathcal{L}_e^j = \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}_w\big(C_e^j(\mathcal{J}(E_\psi(x_i))), y_i\big) \tag{11}$$

where n represents the total number of labeled CXR samples, and y_i denotes the one-hot representation of the class label, $i = 1, \ldots, n$, and we recall that $\mathcal{J}(\cdot)$ denotes the MHMF

operation given in (10) and $E_\psi(\cdot)$ is the preprocessing step on the CXR samples. Note that, the total number of COVID-19 cases is much smaller than non-COVID cases, i.e., normal and pneumonia cases. This imbalance in the dataset leads to a high ratio of false-negative classification. To mitigate this negative effect, we employ a weighted cross-entropy $\mathcal{L}_w(\cdot)$ given as follows:

$$\mathcal{L}_w(\widehat{y}_i, y_i) = -\frac{1}{C} \sum_{c=1}^{C} \lambda_c \cdot y_{i,c} \log \widehat{y}_{i,c} \tag{12}$$

where C is the total number of classes, $y_{i,c}$ is the cth element of y_i, and $\widehat{y}_{i,c}$ denotes the corresponding prediction. λ_c represents the weight that controls how much the error on class c contributes to the loss, $c = 1, \ldots, C$. Finally, the loss $\mathcal{L}_M$ of the whole MUL module is derived by averaging the loss values of all the experts

$$\mathcal{L}_M = \frac{1}{N} \sum_{j=1}^{N} \mathcal{L}_e^j. \tag{13}$$

We use the variance of classification loss $\mathcal{L}_e^j$ with regards to the average loss $\mathcal{L}_M$ to quantify the uncertainty, denoted by σ, which is given as

$$\sigma = \frac{1}{N} \sum_{j=1}^{N} \big(\mathcal{L}_M - \mathcal{L}_e^j\big)^2. \tag{14}$$

The proposed MUL module improves the diagnostic accuracy as the final prediction combines the results from multiple experts, and also mitigates the negative effects caused by the noise in the data by introducing the dropout layers. Moreover, the experiments have revealed that the more experts in the MUL module the faster the system converges during training.

D. Training

The whole architecture of RCoNet$_s^k$ is presented in Fig. 2, where the CXR images are first processed by a stack of deformable convolution layers, and then are transformed to high-order moment latent features by the MHMF module. Finally, the MUL module utilizes the learned high-order features to generate final diagnoses. The loss used to optimize RCoNet$_s^k$ is given as follows:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_M - \alpha \mathcal{L}_I \tag{15}$$

where $\mathcal{L}_M$ is the prediction loss given by (13), and $\mathcal{L}_I$ denotes the MI between the input X and the latent representation $E_\psi(X)$ estimated by either (6) or (7). α is a positive hyper-parameter that governs how much $\mathcal{L}_M$ and $\mathcal{L}_I$ contribute to the total loss. During training, the trainable parameters of the whole systems are updated iteratively to minimize $\mathcal{L}_{\text{total}}$, which is to jointly minimize the prediction loss $\mathcal{L}_M$ thus to improve the accuracy and maximize the MI $\mathcal{L}_I$.

IV. EXPERIMENTS AND RESULTS

A. Dataset

We use a public CXR dataset, referred to as *COVIDx*, to evaluate the proposed model, which is

TABLE I
DETAILS OF PATIENT DATA USED FOR TRAINING AND TESTING

Data	Number of Patients Per Class			Total Patients
	Normal	Pneumonia	COVID-19	
Train	7966	5451	207	13624
Test	885	594	31	1510

TABLE II
DETAILS OF 10% NOISY PATIENT DATA USED FOR TRAINING

Training Date	Clean	Noise	Total
Normal	7170	796 (Peumonia+COVID-19)	7966
Pneumonia	4906	545 (COVID-19+Normal)	5451
COVID-19	187	20 (Peumonia+Normal)	207

published by the authors of COVID-Net [14]. This dataset contains a total of 13975 CXR images from 13870 patients of 3 classes: (a) normal (no infections); (b) pneumonia (non-COVID-19 pneumonia); (c) COVID-19. It contains samples from five open source available data repositories https://github.com/lindawangg/COVID-Net/blob/master/docs/COVIDx.md. Three random CXR samples of these three classes are shown in Fig. 1. To reduce the negative effect caused by extremely unbalanced training samples, i.e., a very limited number of COVID-19 positive cases compared to the other two categories, we further include other open-source CXR datasets from https://www.kaggle.com/c/rsna-pneumonia-detection-challenge/data. Following [14], [47], the dataset is finally divided into 13624 training and 1510 test samples. The numbers of samples from different categories used for training and testing are summarized in Table I. Moreover, we also adopted various data augmentation techniques to generate more COVID-19 training samples, such as flipping, translation, rotation using random five different angles, to tackle the data imbalance issue such that the proposed model can learn an effective mechanism of detecting COVID-19.

B. Evaluation Metrics

In our experiments, we use the following six metrics to evaluate the COVID-19 detection performance of different approaches:

1) Accuracy (ACC): ACC calculates the proportion of images that are correctly identified. ACC = (TP + TN/TP + TN + FP + FN).
2) Sensitivity (SEN): SEN is the ratio of the positive cases that have been correctly detected to all the positive cases. SEN = (TP/TP + FN).
3) Specificity (SPE): SPE is the ratio of the negative cases that have been correctly classified to all the negative cases. SPE = (TN/TN + FP).
4) Balance (BAC): BAC is the mean value of SEN and SPE. BAC = (SEN + SPE/2).
5) Positive Predictive Value (PPV): PPV is the ratio of correctly detected positive cases to all cases that are detected to be positive. PPV = (TP/TP + FP).
6) F1-score (F1): F1 uses a combination of accuracy and sensitivity to calculate a balanced average result. F1 = (2 × ACC × SEN/ACC + SEN).

TN, TP, FN and FP represent the total number of true negatives, true positives, false negatives, and false positives, respectively.

C. Compared Methods

We compare the proposed $RCoNet_s^k$ with the following five existing deep learning methods for COVID-19 detection:

1) PbCNN [15]: A patch-based convolutional neural network (CNN) with a relatively small number of trainable parameters.
2) COVID-Net [14]: A tailored deep CNN that uses a projection-expansion-projection design pattern.
3) DenseNet-121 [48]: A densely connected convolutional network that connects each layer to every other layer in a feed-forward fashion.
4) CoroNet [49]: A deep CNN model based on Xception architecture pretrained on ImageNet dataset.
5) ReCoNet [47]: A residual image-based COVID-19 detection network that exploits a CNN-based multi-level preprocessing filter block and a multitask learning loss.

D. Implementation

We implement our $RCoNet_s^k$ using the PyTorch library and apply ResNeXt [50] as the backbone network. We train the model with the Adam optimizer with an initial learning rate of 2×10^{-4} and a weight decay factor of 1×10^{-4}. All the experiments are run on an NVIDIA GeForce GTX 1080Ti GPU. We set the batch size to be 8, and resize all images to 224×224 pixels. The hyperparameter α in the loss function given in (15) is set to be within the range of $[0, 0.4]$. The drop rate of each dropout layer in the MUL module is randomly chosen from $\{0.1, 0.3, 0.5\}$. The loss weight λ_c for each category, which is used to calculate the weighted sum of the loss as given in (12), is set to be 1, 1, and 20 for the normal, pneumonia, COVID-19 samples, respectively, corresponding to the number of training samples in each. We adopt fivefold cross-validation and evaluate our proposed model with a different number of order moments for the MHMF module k, and a different number of experts s.

To evaluate the performance of the proposed model with the presence of label noise, we derive a noisy dataset from the given dataset in the following way: we randomly select a given percentage of training samples in each category and assign wrong labels to these samples. In particular, to ensure that the fake COVID-19 samples are less than the real ones, we assign the COVID-19 labels to select normal and pneumonia samples in a way the number of normal and pneumonia samples assigned with the COVID-19 label equals the number of COVID-19 samples assigned with either normal or pneumonia label. We show a realization of the derived noisy dataset when the percentage of fake samples is set to be 10% in Table II.

E. Results and Discussions

1) Performance on Clean Data: The numerical results on the clean dataset without any artificial noise added are shown in Table III. The results are presented in the form of

TABLE III

PERFORMANCE COMPARISON OF DIFFERENT APPROACHES FOR COVID-19 DETECTION ON THE COVIDx DATASET

| Method	ACC (%)	SEN (%)	SPE (%)	BAC (%)	PPV (%)	F1 (%)	Param (M)	FLOPs (G)
PbCNN [15]	88.90±1.63	85.90±1.69	96.40±2.10	91.15±1.31	88.65±1.52	87.37±2.14	11.60	-
COVID-Net [14]	95.10±1.34	91.37±1.37	95.76±2.04	93.57±0.89	94.73±0.97	93.20±0.85	117.4	15.10
DenseNet-121 [48]	97.40±1.67	96.08±0.88	97.23±1.01	96.66±1.21	96.05±1.00	96.74±1.04	7.61	**5.59**
CoroNet [49]	95.00±1.58	96.90±1.57	97.50±1.93	97.20±1.07	95.00±1.03	95.60±0.95	33.00	-
ReCoNet [47]	97.48±1.05	97.39±1.67	97.53±1.28	97.46±0.87	97.17±0.76	97.43±0.59	**2.52**	7.68
$RCoNet_4^1$	96.12±0.33	95.71±0.41	96.38±0.29	96.05±0.20	95.86±0.62	95.91±0.56	6.73	7.61
$RCoNet_4^2$	96.78±0.57	96.48±0.69	96.91±0.74	96.70±0.34	96.94±0.53	96.63±0.58	6.74	7.70
$RCoNet_4^3$	97.46±0.43	97.25±0.79	97.62±0.40	97.44±0.82	97.59±0.91	97.35±0.38	6.75	7.79
$RCoNet_4^4$	**97.89±0.53**	97.33±0.45	**98.24±0.39**	**97.79±0.62**	**97.93±0.74**	97.61±0.48	6.77	7.91
$RCoNet_4^5$	97.50±0.62	**97.76±0.87**	97.18±0.63	97.47±0.73	97.10±0.91	**97.63±0.71**	6.77	8.00
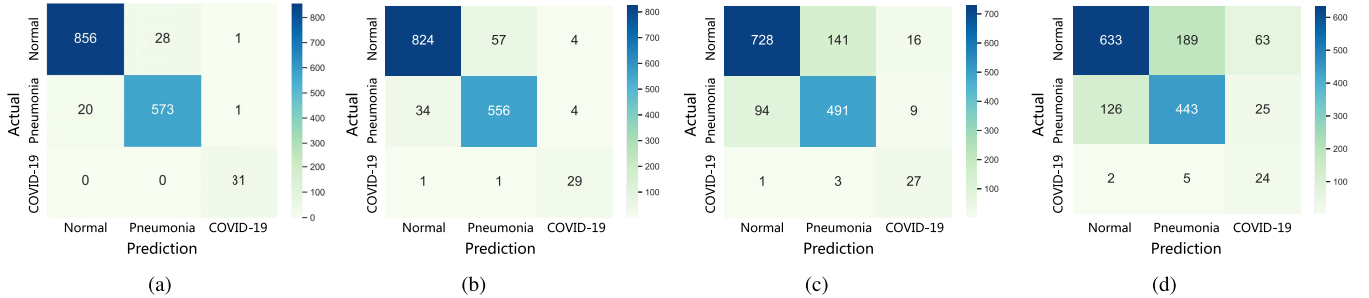
Fig. 4. Confusion matrices of the proposed $RCoNet_s^k$ trained on noisy dataset with different percentages of noisy samples. (a) Clean. (b) 10% Noise. (c) 20% Noise. (d) 30% Noise.

TABLE IV

PERFORMANCE COMPARISON OF DIFFERENT APPROACHES ON COVIDx DATASET WITH NOISY SAMPLES

Noise	Method	ACC (%)	SEN (%)	SPE (%)
10%	PbCNN [15]	83.22	81.98	89.01
	COVID-Net [14]	91.03	87.94	90.62
	DenseNet-121 [48]	91.97	87.94	92.17
	CoroNet [49]	89.45	88.74	90.06
	ReCoNet [47]	91.63	90.82	91.16
	$RCoNet_4^3$	92.78	92.21	**93.51**
	$RCoNet_4^4$	**92.98**	**93.39**	93.12
	$RCoNet_4^5$	92.01	91.41	92.76
20%	PbCNN [15]	78.42	75.90	80.29
	COVID-Net [14]	82.51	82.77	81.95
	DenseNet-121 [48]	82.16	81.01	82.21
	CoroNet [49]	82.33	81.10	81.89
	ReCoNet [47]	83.26	82.72	83.17
	$RCoNet_4^3$	84.18	84.56	85.79
	$RCoNet_4^4$	84.30	**84.01**	**85.99**
	$RCoNet_4^5$	**84.34**	83.96	85.21
30%	PbCNN [15]	67.76	66.47	70.61
	COVID-Net [14]	71.98	70.13	71.55
	DenseNet-121 [48]	72.74	72.36	72.96
	CoroNet [49]	71.87	72.02	71.54
	ReCoNet [47]	73.26	72.53	73.11
	$RCoNet_4^3$	74.56	74.20	75.54
	$RCoNet_4^4$	74.69	**74.51**	**76.94**
	$RCoNet_4^5$	**74.88**	74.37	75.21

$a \pm b$, where a and b denote the average and variance values of each metric on five independent experiments, respectively. We can see that $RCoNet_4^5$, i.e., the proposed model with $k = 4$ levels of mixed moment features and $s = 4$ experts, achieves notable performance improvement over the comparison methods in terms of most metrics considered, including ACC, SPE, BAC, PPV, and F1 score. We note the performance of $RCoNet_s^k$

can be further improved with a different set of k and s. For instance, $RCoNet_4^5$ achieves a better SEN and F1 score than $RCoNet_4^4$. The higher ACC and F1 score validate that $RCoNet_s^k$ is able to obtain latent features, i.e., the mixed moment features of different levels of order, that maintain interclass separability and intraclass compactness better than other models. Note that $RCoNet_4^5$ leads to a higher SEN than all other methods, which is particularly important to COVID-19 detection since successfully detecting COVID-19 positive cases is the key to control the spread of this super contagious disease. Moreover, it can be observed that $RCoNet_s^k$ has smaller variance compared to the others, which demonstrates the robustness and stability of our model.

2) Complexity Discussion: We also evaluate the complexity of the proposed model in terms of numbers of parameters and computational cost, i.e., Float-point operations (FLOPs), which is presented in Table III. It can be observed that the proposed model has much fewer parameters than several existing methods, except ReCoNet. However, we note that the FLOPs of $RCoNet_s^k$ is quite close to that of ReCoNet, which means it takes a similar amount of time to diagnose COVID-19 from CXR images by these two models. We can also observe that the increase of k, i.e., the number of mixed moment features, only causes a small or even neglectable amount of increase in the number of parameters and FLOPs as well, which suggests that we can improve the performance of the proposed model by optimizing k, without the concern on the significant increase of the complexity. As for s, the number of experts in MUL, we select 4 which is confirmed to have better performance with a bit of computational cost increase after a great number of experiments.
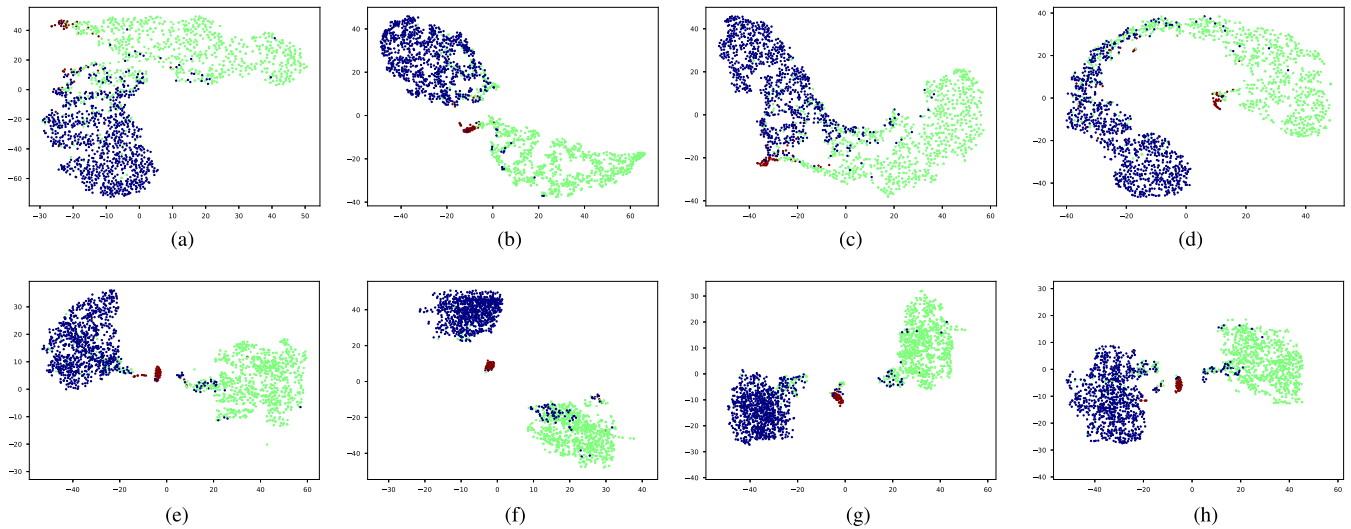
Fig. 5. t-SNE visualization of the latent features generated by different methods. Blue, green and red dots represent normal, pneumonia and COVID-19 samples, respectively. (a) Baseline. (b) RCoNet-D. (c) RCoNet-M. (d) RCoNet-DM. (e) $RCoNet_4^2$. (f) $RCoNet_4^3$. (g) $RCoNet_4^4$. (h) $RCoNet_4^5$.

3) Performance on Noisy Data: We further compare the proposed model to the existing ones when there is noise present in the training dataset. We generate three noisy training datasets in an aforementioned way from the clean dataset with 10%, 20%, and 30% samples with wrong labels, respectively. The results, which we take the averages from five independent experiments, are presented in Table IV. It can be easily seen that the more fake samples we add, the more it degrades the performance of all the methods. Note that the proposed $RCoNet_4^4$ still gets the state-of-the-art results in all considered cases with different percentages of noisy samples in the training dataset. Moreover, the performance gain over the existing methods slightly increases with the ratio of noisy samples, verifying that our model is more robust to the noise. Note that the extreme case of 30% noisy samples leads to great performance degradation of all the models. In practice, the percentage of label noise is usually around 10% to 20%. We present the confusion matrices in Fig. 4 to summarize the prediction accuracy of different categories. We can observe that, although with very limited number of COVID-19, our model still maintains high accuracy of detecting COVID-19 cases, even with the presence of noisy samples.

4) Uncertainty Estimation: One remarkable advantage of our model is the ability to quantify the uncertainty in the final prediction, which is significantly crucial for COVID-19 detection. This is done by obtaining the variance in the output of different experts in MUL as described in Section III-C. The larger the variance is, the more different experts disagree with each other, and, hence, the more uncertain the model is about the final prediction. We present two CXR samples in Fig. 6, including the predictions and the corresponding uncertainty level by $RCoNet_s^k$. We can see that the correctly classified CXR image has a low uncertainty level about its prediction, i.e., 0.0094, and the misclassified CXR sample with a high uncertainty level, i.e., 0.4792, suggests that an alternative way of diagnosis should be sought to correct this prediction. This
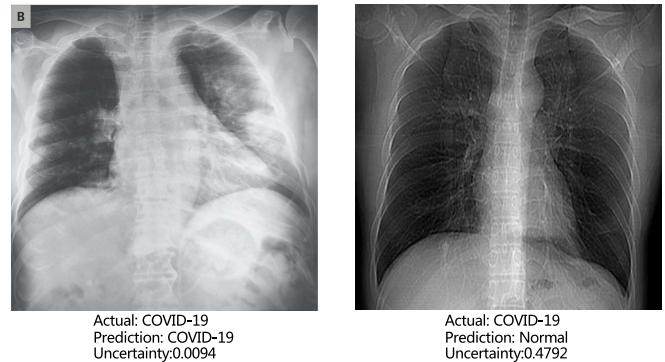

Actual: COVID-19
Prediction: COVID-19
Uncertainty:0.0094

Actual: COVID-19
Prediction: Normal
Uncertainty:0.4792

Fig. 6. Example CXR samples with their predictions and the corresponding uncertainty levels by $RCoNet_4^4$.

greatly improves the reliability of the prediction by $RCoNet_s^k$, and reduces the chance of misdiagnosis. We also show in Fig. 7 the average uncertainty levels of $RCoNet_s^k$ trained on clean and noisy datasets with different ratios of noisy samples. It can be observed that the uncertainty level increases almost linearly with the percentage of noisy samples in the dataset, which highlights the negative impact of noise on model training.

F. Analysis

We further numerically analyze the benefits of the three key modules of $RCoNet_s^k$, i.e., the DeIM, MHMF and MUL modules in this section.

1) Effectiveness of DeIM: As shown in Fig. 5, we utilize t-stochastic neighbor embedding (SNE) method [51] to visualize the latent features, which are generated by the bottleneck layers of the baseline model, i.e., ResNeXt, $RCoNet_s^k$ and three variants of $RCoNet_s^k$: (a) RCoNet-D: a model contains only DeIM; (b) RCoNet-M: a model contains only MUL; (c) RCoNet-DM: a model contains DeIM and MUL but not MHMF. Comparing the latent feature distribution by the baseline model shown in Fig. 5(a), and that by RCoNet-D

TABLE V
IMPACT OF THE MHMF AND MUL ON THE MODEL PERFORMANCE

$RCoNet_s^k$	s=1	s=2	s=3	s=4	s=5	s=6	s=7
k=1	95.4	95.7	95.9	96.1	96.1	96.0	95.8
k=2	96.3	96.4	96.6	96.8	96.8	96.7	96.4
k=3	97.2	97.2	97.3	97.5	97.4	97.3	97.3
k=4	97.4	97.6	97.8	**97.9**	**97.9**	97.7	97.5
k=5	97.2	97.3	97.3	97.4	97.5	97.5	97.3
k=6	96.8	97.0	97.0	97.1	97.0	96.9	96.9
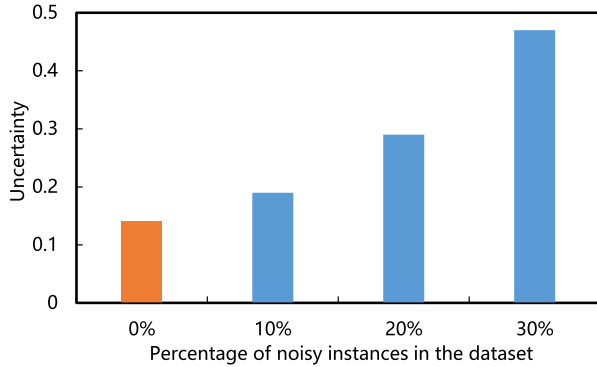
Fig. 7.   Comparison on uncertainty level of the predictions by $RCoNet_4^4$.

Fig. 8.   Prediction accuracy by $RCoNet_4^4$ with regards to different values of α.

Fig. 9.   Comparison on the learning trajectories of different models.

presented in Fig. 5(b), we can tell that the introduction of DeIM leads to better class separation in the latent space.

2) Effectiveness of MHMF: We can observe in Fig. 5(a)–(d) that the latent features of the COVID-19 samples, generated by the models without MHMF, always distribute around the category boundary, and are not quite separable from those of some pneumonia samples. Meanwhile, the latent feature distributions presented in Fig. 5(e)–(h) derived by the models with MHMF show significant separability between different categories, which implies that MHMF can extract discriminative features. We also include numerical results of $RCoNet_s^k$, trained and tested on COVIDx dataset, with regards to different values of k, i.e., the number of levels of the moment features to be mixed, and s, i.e., the number of experts, in Table V in terms of accuracy. We can observe that, for a given value of s, the accuracy increases first with the value of k but decreases after k is larger than 4. It demonstrates that including more levels of moment feature could improve the model performance. However, the overly high-order moments may lead to performance degradation, which may be because these features are not useful for COVID detection.

3) Effectiveness of MUL: From Table V, we observe that, for a given value of k, accuracy increases first with the value of s but saturates around $s = 5$. This implies that having more experts in MUL can increase the prediction accuracy but it is not necessary to have too many.

4) Parameter Sensitivity and Convergence: We evaluate how sensitive the model performance in terms of accuracy to the value of α. We show the average accuracy of five independent experiments by $RCoNet_4^4$ trained on the dataset with different ratios of noisy samples in Fig. 8. As we can see, the larger α, which means the prediction loss, i.e., $\mathcal{L}_M$, contributes less to the total loss, not necessarily leads to
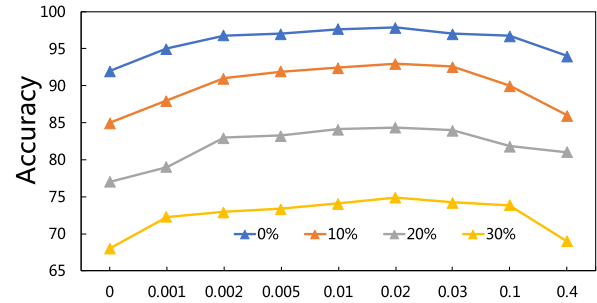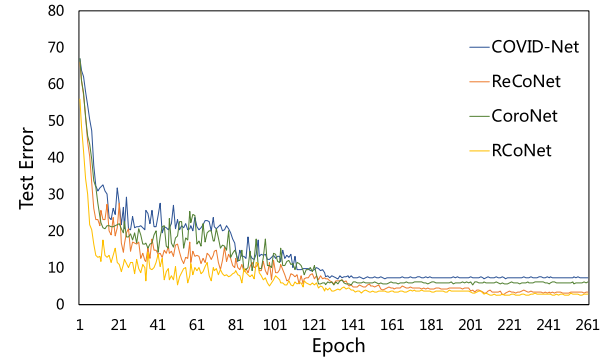
degradation in the accuracy. This means maximizing the MI between the input and the latent features could keep useful information within the latent features, thus improving the prediction accuracy. We have also shown the learning curves of different models in Fig. 9, which shows that $RCoNet_4^4$ converges slightly faster than the others, including COVID-Net, ReCoNet and CoroNet.

V. CONCLUSION

In this article, we proposed a novel deep network model, named $RCoNet_s^k$, for robust COVID-19 detection, which contains three key components, i.e., DeIM, MHMF and MUL. DeIM estimates and maximizes the MI between input data and the latent representations simultaneously to obtain the category separability in the latent space. MHMF overcomes the limited expressive capability of low-order statistics, and uses a combination of both low and high order moment features to extract more informative and discriminative features. MUL generates the final diagnosis and the uncertainty estimation by combining the output of multiple parallel dropout networks, each as an expert. We numerically validated that the proposed RCoNet trained on either the public COVIDx dataset or the noisy version of it outperforms the existing methods in terms of all the metrics considered. We noted that these three modules can be easily implemented into other frameworks for different tasks.

REFERENCES

[1] 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.e11–1433.e11, Jun. 2020.

[2] 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.

[3] "Artificial intelligence-enabled rapid diagnosis of patients with COVID-19," *Nature Med.*, vol. 26, pp. 1224–1228, 2020.

[4] W. Xie, C. Jacobs, J.-P. Charbonnier, and B. van Ginneken, "Relational modeling for robust and efficient pulmonary lobe segmentation in CT scans," *IEEE Trans. Med. Imag.*, vol. 39, no. 8, pp. 2664–2675, Aug. 2020.

[5] 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.

[6] H. X. Bai *et al.*, "AI augmentation of radiologist performance in distinguishing COVID-19 from pneumonia of other etiology on chest CT," *Radiology*, vol. 296, no. 3, 2020, Art. no. 201491.

[7] A. A. Ardakani, A. R. Kanafi, U. R. Acharya, N. Khadem, and A. Mohammadi, "Application of deep learning technique to manage COVID-19 in routine clinical practice using CT images: Results of 10 convolutional neural networks," *Comput. Biol. Med.*, vol. 121, Jun. 2020, Art. no. 103795.

[8] 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.

[9] 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.

[10] L. Sun *et al.*, "Adaptive feature selection guided deep forest for COVID-19 classification with chest CT," 2020, *arXiv:2005.03264*. [Online]. Available: http://arxiv.org/abs/2005.03264

[11] D. Donglin *et al.*, "Hypergraph learning for identification of COVID-19 with CT imaging," *Med. Image Anal.*, vol. 68, Feb. 2020, Art. no. 101910.

[12] Z. Y. Zu *et al.*, "Coronavirus disease 2019 (COVID-19): A perspective from China," *Radiology*, vol. 296, no. 6, 2020, Art. no. 200490.

[13] M. Siddhartha and A. Santra, "COVIDLite: A depth-wise separable deep neural network with white balance and CLAHE for detection of COVID-19," 2020, *arXiv:2006.13873*. [Online]. Available: http://arxiv.org/abs/2006.13873

[14] L. Wang and A. Wong, "COVID-Net: A tailored deep convolutional neural network design for detection of COVID-19 cases from chest X-ray images," 2020, *arXiv:2003.09871*. [Online]. Available: http://arxiv.org/abs/2003.09871

[15] 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.

[16] F. Maes, A. Collignon, D. Vandermeulen, G. Marchal, and P. Suetens, "Multimodality image registration by maximization of mutual information," *IEEE Trans. Med. Imag.*, vol. 16, no. 2, pp. 187–198, Apr. 1997.

[17] A. Hyvärinen and E. Oja, "Independent component analysis: Algorithms and applications," *Neural Netw.*, vol. 13, nos. 4–5, pp. 411–430, Jun. 2000.

[18] N. Kwak and C.-H. Choi, "Input feature selection by mutual information based on Parzen window," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 12, pp. 1667–1671, Dec. 2002.

[19] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 8, pp. 1226–1238, Aug. 2005.

[20] A. J. Butte and I. S. Kohane, "Mutual information relevance networks: Functional genomic clustering using pairwise entropy measurements," in *Biocomputing*. Singapore: World Scientific, 1999, pp. 418–429.

[21] C. Vondrick, A. Shrivastava, A. Fathi, S. Guadarrama, and K. Murphy, "Tracking emerges by colorizing videos," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 391–408.

[22] R. Arandjelovic and A. Zisserman, "Look, listen and learn," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 609–617.

[23] M. I. Belghazi *et al.*, "Mutual information neural estimation," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 531–540.

[24] J. Chang, Z. Lan, C. Cheng, and Y. Wei, "Data uncertainty learning in face recognition," 2020, *arXiv:2003.11339*. [Online]. Available: http://arxiv.org/abs/2003.11339

[25] Y. Shi and A. Jain, "Probabilistic face embeddings," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6902–6911.

[26] A. Kendall, V. Badrinarayanan, and R. Cipolla, "Bayesian SegNet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding," 2015, *arXiv:1511.02680*. [Online]. Available: http://arxiv.org/abs/1511.02680

[27] C. Blundell, J. Cornebise, K. Kavukcuoglu, and D. Wierstra, "Weight uncertainty in neural networks," 2015, *arXiv:1505.05424*. [Online]. Available: http://arxiv.org/abs/1505.05424

[28] Y. Gal, "Uncertainty in deep learning," Ph.D. dissertation, Univ. Cambridge, Cambridge, U.K., 2016.

[29] D. J. C. MacKay, "A practical Bayesian framework for backpropagation networks," *Neural Comput.*, vol. 4, no. 3, pp. 448–472, May 1992.

[30] R. M. Neal, *Bayesian Learning for Neural Networks*, vol. 118. Berlin, Germany: Springer, 2012.

[31] A. Kendall and Y. Gal, "What uncertainties do we need in Bayesian deep learning for computer vision?" in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5574–5584.

[32] S. Isobe and S. Arai, "Deep convolutional encoder-decoder network with model uncertainty for semantic segmentation," in *Proc. IEEE Int. Conf. Innov. Intell. Syst. Appl. (INISTA)*, Jul. 2017, pp. 365–370.

[33] J. Choi, D. Chun, H. Kim, and H.-J. Lee, "Gaussian YOLOv3: An accurate and fast object detector using localization uncertainty for autonomous driving," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 502–511.

[34] T. Yu, D. Li, Y. Yang, T. Hospedales, and T. Xiang, "Robust person re-identification by modelling feature uncertainty," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 552–561.

[35] U. Zafar *et al.*, "Face recognition with Bayesian convolutional networks for robust surveillance systems," *EURASIP J. Image Video Process.*, vol. 2019, no. 1, p. 10, Dec. 2019.

[36] M. D. Donsker and S. R. S. Varadhan, "Asymptotic evaluation of certain Markov process expectations for large time, I," *Commun. Pure Appl. Math.*, vol. 28, no. 1, pp. 1–47, Sep. 2010.

[37] S. Nowozin, B. Cseke, and R. Tomioka, "f-GAN: Training generative neural samplers using variational divergence minimization," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 271–279.

[38] M. U. Gutmann and A. Hyvärinen, "Noise-contrastive estimation of unnormalized statistical models, with applications to natural image statistics," *J. Mach. Learn. Res.*, vol. 13, pp. 307–361, Feb. 2012.

[39] P. Bachman, R. D. Hjelm, and W. Buchwalter, "Learning representations by maximizing mutual information across views," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 15509–15519.

[40] J. Xu, P. Ye, Q. Li, H. Du, Y. Liu, and D. Doermann, "Blind image quality assessment based on high order statistics aggregation," *IEEE Trans. Image Process.*, vol. 25, no. 9, pp. 4444–4457, Sep. 2016.

[41] C. Chen *et al.*, "HoMM: Higher-order moment matching for unsupervised domain adaptation," 2019, *arXiv:1912.11976*. [Online]. Available: http://arxiv.org/abs/1912.11976

[42] E. Pauwels and J. B. Lasserre, "Sorting out typicality with the inverse moment matrix sos polynomial," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 190–198.

[43] P. Jacob, D. Picard, A. Histace, and E. Klein, "Metric learning with HORDE: High-order regularizer for deep embeddings," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6539–6548.

[44] H. Jégou and O. Chum, "Negative evidences and co-occurences in image retrieval: The benefit of pca and whitening," in *Proc. Eur. Conf. Comput. Vis.* Berlin, Germany: Springer, 2012, pp. 774–787.

[45] M. Opitz, G. Waltner, H. Possegger, and H. Bischof, "BIER—Boosting independent embeddings robustly," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 5189–5198.

[46] P. Kar and H. Karnick, "Random feature maps for dot product kernels," in *Proc. Artif. Intell. Statist.*, 2012, pp. 583–591.

[47] "ReCoNet: Multi-level preprocessing of chest X-rays for COVID-19 detection using convolutional neural networks," *medRxiv*, to be published.

[48] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4700–4708.

[49] A. I. Khan, J. L. Shah, and M. M. Bhat, "CoroNet: A deep neural network for detection and diagnosis of COVID-19 from chest X-ray images," *Comput. Methods Programs Biomed.*, vol. 196, Nov. 2020, Art. no. 105581.

[50] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5987–5995.

[51] J. Donahue *et al.*, "DeCAF: A deep convolutional activation feature for generic visual recognition," in *Proc. Int. Conf. Mach. Learn.*, 2014, pp. 647–655.

Shunjie Dong received the B.S. degree in integrated circuit design and integration system from Xidian University, Xi'an, China, in 2018. He is currently pursuing the Ph.D. degree with the College of Information Science and Electronic Engineering, Zhejiang University, Hangzhou, China.

Since 2018, he has been advised by Prof. Cheng Zhuo. His research interests include medical image analysis and machine learning.

Qianqian Yang (Member, IEEE) received the B.S. degree in automation from Chongqing University, Chongqing, China, in 2011, the M.S. degree in control engineering from Zhejiang University, Hangzhou, China, in 2014, and the Ph.D. degree in electrical and electronic engineering from Imperial College London, London, U.K., in 2019.

She has held visiting positions at CentraleSupelec Gif-sur-Yvette, France, in 2016, and the New York University Tandon School of Engineering, New York, NY, USA, from 2017 to 2018. After her Ph.D., she served as a Post-Doctoral Research Associate for Imperial College London, and as a Machine Learning Researcher for Sensyne Health Plc, Oxford, U.K. She is currently a Tenure-Tracked Professor with the Department of Information Science and Electronic Engineering, Zhejiang University. Her main research interests include wireless communications, information theory, machine learning and medical imaging. She serves as a Reviewer for IEEE Transactions on Information Theory (TIT), IEEE Transactions on Communications (TCOM), IEEE Transactions on Wireless Communications (TWC), etc.

Yu Fu received the B.S. degree in computer science and technology from Sichuan Agricultural University, Ya'an, China, in 2017, and the M.S. degree in computer software and theory from Lanzhou University, Lanzhou, China, in 2020. He is currently pursuing the Ph.D. degree in electronic science and technology with the College of Information Science and Electronic Engineering, Zhejiang University, Hangzhou, China.

Since 2020, he has been advised by Prof. Cheng Zhuo. His research interests include medical image analysis and brain imaging.
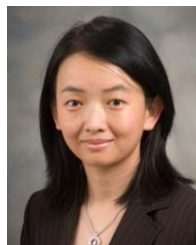
Mei Tian is the Director of Medical Imaging and Nuclear Medicine Program of Zhejiang University, Hangzhou, China, the Vice President of Zhejiang University Medical Center, and the Vice President of Hangzhou Riverside Hospital, Zhejiang University School of Medicine. Her current research interests include radiology, nuclear medicine, and molecular imaging.

Prof. Tian was a recipient of the Merit Award from the Radiological Society of North America (RSNA), the International Young Investigator Grant from RSNA, the International Development and Education Grant from ASCO, the Asian and Oceanian Distinguished Young Investigator Award from the Japanese Society of Nuclear Medicine (JSNM), and the Japan Society for the Promotion of Science (JSPS) Fellowship. She has served the Associate Editor or an Editorial Board Member for the official journals of the WMIS, the Society of Nuclear Medicine and Molecular Imaging (SNMMI), the European Association of Nuclear Medicine (EANM), the British Society for Nanomedicine (BSNM), and JSNM, and the Editorial Consultant of *The LANCET*.

Cheng Zhuo (Senior Member, IEEE) received the B.S. (Hons.) and M.S. degrees in electronic engineering from Zhejiang University, Hangzhou, China, in 2005 and 2007, respectively, and the Ph.D. degree in computer science and engineering from the University of Michigan, Ann Arbor, MI, USA, in 2010.

He is currently a Professor with the College of Information Science and Electronic Engineering, International Joint Innovation Center, Zhejiang University. His current research interests include medical imaging, deep learning, and hardware acceleration.

Dr. Zhuo was a recipient of three Best Paper nominations in DAC'16, CSTIC'18 and ICCAD'20, 2012 Association for Computing Machinery (ACM) SIGDA Technical Leadership Award, and 2017 Japan Society for the Promotion of Science (JSPS) Invitation Fellowship. He has served on the Technical Program and Organization Committees of many international conferences and as an Associate Editor for IEEE Technology Computer-Aided Design (TCAD), ACM *Transactions on Design Automation of Electronic System* (TODAES), and Elsevier *Integration*.