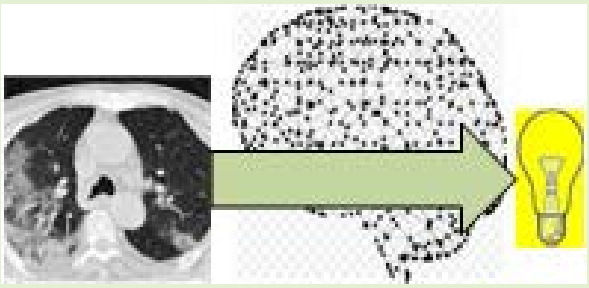
AVNC: Attention-Based VGG-Style Network for COVID-19 Diagnosis by CBAM

Shui-Hua Wang, *Senior Member, IEEE*, Steven Lawrence Fernandes, *Senior Member, IEEE*, Ziquan Zhu, and Yu-Dong Zhang, *Senior Member, IEEE*

***Abstract*—(Aim) To detect COVID-19 patients more accurately and more precisely, we proposed a novel artificial intelligence model. (Methods) We used previously proposed chest CT dataset containing four categories: COVID-19, community-acquired pneumonia, secondary pulmonary tuberculosis, and healthy subjects. First, we proposed a novel VGG-style base network (VSBN) as backbone network. Second, convolutional block attention module (CBAM) was introduced as attention module into our VSBN. Third, an improved multiple-way data augmentation method was used to resist overfitting of our AI model. In all, our model was dubbed as a 12-layer attention-based VGG-style network for COVID-19 (AVNC) (Results) This proposed AVNC achieved the sensitivity/precision/F1 per class all above 95%. Particularly, AVNC yielded a micro-averaged F1 score of 96.87%, which is higher than 11 state-of-the-art approaches. (Conclusion) This proposed AVNC is effective in recognizing COVID-19 diseases.**



***Index Terms*— Attention, covid-19, VGG, convolutional neural network, diagnosis, convolutional block attention module.**

I. Introduction

COVID-19 is an infectious disease. Till 6/Feb/2021, this COVID-19 pandemic caused more than 105.84 million confirmed cases and more than 2.31 million death tolls (US 465.8k deaths, Brazil 231.0k deaths, Mexico 165.7k deaths, India 154.9k deaths, UK 112.0k deaths, Italy 91.0k deaths, France 78.7k, Russia 76.6k, Germany 61.9k, Spain 61.3k, Iran 58.4k, Columbia 55.6k, etc.)

Manuscript received 8 February 2021; accepted 24 February 2021. Date of publication 26 February 2021; date of current version 14 September 2022. This work was supported in part by the Royal Society International Exchanges Cost Share Award, U.K., under Grant RP202G0230; in part by the Medical Research Council Confidence in Concept Award, U.K., under Grant MC_PC_17171; in part by the Hope Foundation for Cancer Research, U.K., under Grant RM60G0680; in part by the Fundamental Research Funds for the Central Universities under Grant CDLS-2020-03; and in part by the Key Laboratory of Child Development and Learning Science (Southeast University), Ministry of Education. The associate editor coordinating the review of this article and approving it for publication was Dr. Yin Zhang. *(Shui-Hua Wang and Steven Lawrence Fernandes contributed equally to this work.) (Corresponding authors: Ziquan Zhu; Yu-Dong Zhang.)*

Shui-Hua Wang is with the School of Mathematics and Actuarial Science, University of Leicester, Leicester LE1 7RH, U.K. (e-mail: shuihuawang@ieee.org).

Steven Lawrence Fernandes is with the Department of Computer Science, Design & Journalism, Creighton University, Omaha, NE 68178 USA (e-mail: stevenfernandes@creighton.edu).

Ziquan Zhu is with Science in Civil Engineering, University of Florida, Gainesville, FL 32608 USA (e-mail: zhu.ziquan@ufl.edu).

Yu-Dong Zhang is with the School of Informatics, University of Leicester, Leicester LE1 7RH, U.K. (e-mail: yudongzhang@ieee.org).

Digital Object Identifier 10.1109/JSEN.2021.3062442

Symptoms of COVID-19 vary; nevertheless, they entail fever and coughs. Some patients exhibit shortness of breath, loss of smell/taste, headache, dizziness, skin rashes, muscle pain, fatigue, etc. The symptoms may change over time [1].

Two types of diagnosis methods are available at present. One type is viral testing [2]. Currently, there are numerous walk-through testing site [3] at UK. There are also free NHS COVID-19 test kits available. The other type is imaging methods, which entails chest ultrasound, chest X-ray, and chest computerized tomography (CCT). Compared to viral testing, the imaging methods are much quicker. After the images are reconstructed by fixed computer programs, the radiologists need to check through the suspicious regions, and give a decision whether the scanned patient is positive or negative.

Among all the three imaging methods, CCT provides the highest sensitivity. The chest ultrasound is operator-dependent, which needs high level operation skills. Besides, it is hard to keep the ultrasound probe at the same position during scanning. For chest X-ray (radiograph), its limitations include poor soft tissue contrast and 2D image generation. In all, CCT can generate 3D image around the lung area, and give higher sensitivity diagnosis for COVID-19. The main biomarkers in CCT distinguishing COVID-19 are ground-glass opacities (GGOs) without pleural effusions [4].

Nevertheless, manual delineation by radiologists is tedious, labor-intensive, and easy to be influenced by inter and intra-expert factors. Artificial intelligence (AI) and deep

learning (DL) approaches have gained promising results in analyzing CCT images.

For example, Lu [5] presented a new extreme learning machine combined with bat algorithm (ELM-BA) approach. Li and Liu [6] presented the real-coded biogeography-based optimization (RCBBO) approach, which can be used in our COVID-19 recognition task. Szegedy, *et al.* [7] presented the GoogleNet (GN), which can be used in this study. Guo and Du [8] used ResNet18 (RN18) for ultrasound standard plane classification. Both GN and RN18 could be used in our COVID-19 task.

Satapathy [9] presented a five-layer deep network model using stochastic pooling (5LSP) to replace traditional max pooling. Satapathy and Zhu [10] expanded it to 7-layer stochastic pooling (7LSP). Ko, *et al.* [11] presented a fast-track COVID-19 classification network (FCONet). Ni, *et al.* [12] presented a deep learning approach (DLA) to characterize COVID-19 in CCT images. Cohen, *et al.* [13] presented a COVID Severity Score (CSS) network model. Wang, *et al.* [14] proposed a 3D deep DNN to detect COVID-19 (DeCovNet). Togacar, *et al.* [15] used SqueezeNet, MobileNetV2, and chose social mimic optimization (SMO) approach to select features for further fusion. The overall classification rate obtained with their proposed approach was 99.27%. Wang [16] combined graph convolutional network with traditional convolutional neural network to detect COVID-19. Wang [17] proposed $(L, 2)$ transfer feature learning method, and developed a novel discriminant correlation analysis fusion approach.

Nevertheless, previous studies tried to improve the performance of deep neural networks from three importance factors: depth, width, and cardinality. In this study, we try to use attention mechanism to adjust the structure of deep AI models. Our proposed AI model is named attention-based VGG-style network for COVID-19 (AVNC). The contributions of our work are five-folds:

(i) A VGG-style base network (VSBN) is proposed as backbone network
(ii) Convolutional block attention module (CBAM) is embedded to help include attention to our AI model.
(iii) An improved multiple-way data augmentation is proposed to resist overfitting
(iv) Gram-CAM is introduced to show the most important areas that AI are observing.
(v) The results of our proposed model (AVNC) is experimentally proven to outperform state-of-the-art approaches.

II. Dataset

This retrospective study was exempt by Institutional Review Board of local hospitals. Four categories were used in this study: (i) COVID-19 positive patients; (ii) community-acquired pneumonia (CAP); (iii) second pulmonary tuberculosis (SPT); and (iv) healthy control (HC).

For each subject, $m = \{1, 2, 3, 4\}$ slices were chosen via slice level selection (SLS) approach. For the three diseased groups, the slices displaying the largest number of lesions and size were chosen. For HC subjects, any slice within the 3D image was randomly chosen. The average selected slices m is defined as $m(c) = N_I(c)/N_P(c), c = 1, \cdots, 4$, where N_I stands for the number of images via SLS method, and N_P the number of patients, c stands for the category index.

TABLE I
Subjects and Images of Four Categories

Class Index	Class Title	N_P	N_I	$m = N_I/N_P$
$c = 1$	COVID-19	125	284	2.27
$c = 2$	CAP	123	281	2.28
$c = 3$	SPT	134	293	2.18
$c = 4$	HC	139	306	2.20
Total		521	1164	2.23

In total, we enrolled $\sum_{c=1}^{4} N_p(c) = 521$ subjects, and produced $\sum_{c=1}^{4} N_I(c) = 1164$ slice images using the SLS method. Table I lists the demographics of the four-category subject cohort with the values of triplets $[m, N_P, N_I]$, where m of the total set equals to $m = \sum_{c=1}^{4} N_I(c) / \sum_{c=1}^{4} N_p(c) = 2.23$.

Three experienced radiologists (Two juniors: F_{J1} and F_{J2}, and one senior: F_S) were convened to curate all the images $\{x(k)\}$, where $x(k)$ means k-th CCT images, L stands for the labelling of each individual radiologist. The final labelling L_F of the CCT scan $x(k)$ is obtained by

$$L_F[x(k)] = \begin{cases} L[F_{J1}, x(k)], & L[F_{J1}, x(k)] == L[F_{J2}, x(k)] \\ \text{MAV}\{F_a[x(k)]\}, & \text{otherwise} \end{cases} \tag{1}$$

where MAV denotes majority voting, F_a the labelling of all three radiologists, viz.,

$$L_a[x(k)] = [L(F_{J1}, x(k)), L(F_{J2}, x(k)), L(F_S, x(k))] \tag{2}$$

The above equation denotes that at the situation of disagreement between the analyses of two junior radiologists (F_{J1}, F_{j2}), it is necessary to consult a senior radiologist (F_S) to reach a MAV-type consensus. The CCT and labelling data are available on request due to privacy/ethical restrictions.

III. Methodology

A. Preprocessing

Suppose the original dataset containing four categories is named $O_{orig} = \{o_{orig}(1), \cdots, o_{orig}(i), \cdots\}$, where $o_{orig}(i)$ is the i-th original image.

Figure 1 presents the flowchart of our preprocessing procedure: (i) RGB to grayscale; (ii) histogram stretch; (iii) margin crop; and (iv) resizing. Suppose the dataset generated by each step is O_i, O_{ii}, O_{iii}, and O_{prep}, we can yield

$$\begin{aligned} O_i &= F_{Gray}(O_{Orig}|\text{RGB} \rightarrow \text{Grayscale}) \\ &= \{o_i(1), \ldots, o_i(i), \ldots\} \end{aligned} \tag{3}$$

$$O_{ii} = F_{HS}(O_i) = \{o_{ii}(1), \cdots, o_{ii}(i), \ldots\} \tag{4}$$

$$\begin{aligned} O_{iii} &= F_{Crop}\left(O_{ii}, \left[c^t, c^b, c^l, c^r\right]\right) \\ &= \{o_{iii}(1), \cdots, o_{iii}(i), \cdots\} \end{aligned} \tag{5}$$

$$\begin{aligned} O_{prep} &= F_{DS}(O_{iii}, [256256]) \\ &= \{o_{prep}(1), \ldots, o_{prep}(i), \ldots\} \end{aligned} \tag{6}$$

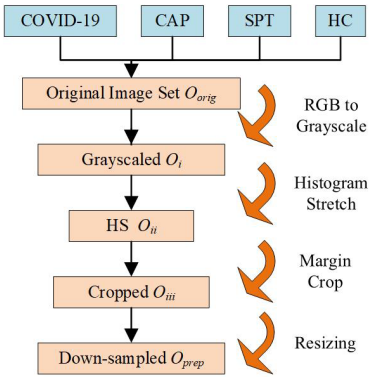


Fig. 1. Preprocessing on raw dataset (HS: histogram stretch).

TABLE II
IMAGE SIZE AND STORAGE PER IMAGE AT EACH PREPROCESSING STEP

Preprocess	Variable	Size (per image)
Original	o_{orig}	$1024^2 \times 3 = 3{,}145{,}728$
Step 1	o_i	$1024^2 \times 1 = 1{,}048{,}576$
Step 2	o_{ii}	$1024^2 \times 1 = 1{,}048{,}576$
Step 3	o_{iii}	$624^2 \times 1 = 389{,}376$
Output	o_{prep}	$256^2 \times 1 = 65{,}536$

where F_{Gray} means the grayscale operation. F_{HS} stands for the histogram stretch operation, which maps the intensity value range to the standard range $[0, 255]$. F_{Crop} stands for the margin crop function. c^t means the pixels to be cropped of top side, which equals 150 in this study. Other three factors c^b, c^l, c^r stand for the cropped pixels to the bottom side, left side, and right side, respectively. They are all chosen as 200. F_{DS} stands for the downsampling (DS) function. [256256] means the size of downsampled image.

Table II compares the size per image after every preprocessing step. We can see here after preprocessing procedure, the size ratio of o_{prep} over o_{orig} is only $(256/2014)^2/3 = 2.08\%$. Thus, the preprocessing has four major advantages: First, it can remove the unrelated regions (like the checkup bed on the bottom side, and texts on right side side). Second, the color redundant information will be removed. Third, Contrast will be enhanced so the lesions become clearer. Fourth, the sizes of the image are compressed to save storage.

B. VGG-Style Base Network

VGG is a typical CNN architecture, and is considered to be one of excellent model architecture till date [18]. After investigating through the recent VGG networks, we propose a similar VGG-style base network (VSBN) model for our task. Note that VGG-style are popularly used in many recent networks, such as References [19], [20].

VGG-16 has an input size of $224 \times 224 \times 3$ (See S1 in Figure 2a), After the 1st convolution block (CB), which consists of two repetitions of 2 convolutional layers with 64 kernels with sizes of 3×3, abbreviate as $2 \times (64\ 3 \times 3)$, and one max pooling layers, the output is $112 \times 112 \times 64$ (See S2 in Figure 2a). The 2nd CB $2 \times (128\ 3 \times 3)$, 3rd CB $3 \times (256\ 3 \times 3)$, 4th CB $3 \times (512\ 3 \times 3)$, and 5th CB $3 \times (512\ 3 \times 3)$ generate the activation maps with sizes of $56 \times 56 \times 128$, $28 \times 28 \times 256$, $14 \times 14 \times 512$, and $7 \times 7 \times 512$, respectively (See S3-S6 in Figure 2a).

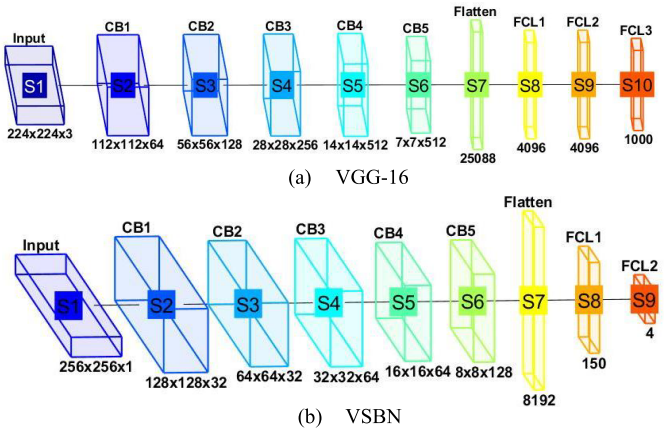


Fig. 2. AM comparison.

TABLE III
SIMILARITIES BETWEEN VSBN AND VGG-16

Index	Similarity Aspect
1	Using small convolution kernels (3×3)
2	Using small-kernel max pooling with size of (2×2)
3	Several repetitions of conv layers followed by max pooling
4	Fully-connected layers at the end
5	Size of feature maps shrinks as it goes from input to output
6	Channel number increase as it goes from input layer to the last conv layer, and then decreases as to output layer.

Afterwards, the AM is flattened into a vector of 25,088 neurons, and passed into three fully-connected layers with neurons of 4096, 4096, and 1000, respectively (See S7-S10 in Figure 2a).

Our proposed VSBN model is similar to VGG-16 network. Its activation maps (AMs) of each convolutional block and fully-connected layers (FCLs) are shown in Figure 2(b). The similarities are in terms of six following aspects as shown in Table III.

C. Attention-Based VGG-Style Network for COVID-19

To improve the performance of the deep neural network, many researches are done with respects to either depth, or width, or cardinality. Recently, Woo, *et al.* [21] proposed a novel convolutional block attention module (CBAM), which mainly features in the attention mechanism. Attention not only tells the neural network model where to focus, it also improves the representation of interests.

Attention plays an essential role of human visual systems (HVSs) [22]. Figure 3 shows an example of HVS, where image formation is captured by lens of cornea of the eyeballs. Then, iris utilizes the photoreceptor sensitivity to control the exposure. The information flow is then sent to rod cells and cone cells in the retina. Finally, the neural firing is passed to brain for further processing.

Humans do not try to process the whole scenarios at one time; nevertheless, humans take full use of partial glimpses and focus on salient features selectively so as to seize a better visual structure [23]. The core idea of attention mechanism is to refine the three-dimensional feature maps by learning

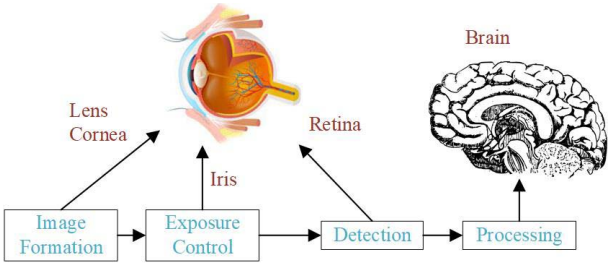


Fig. 3. Diagram of an HVS system.

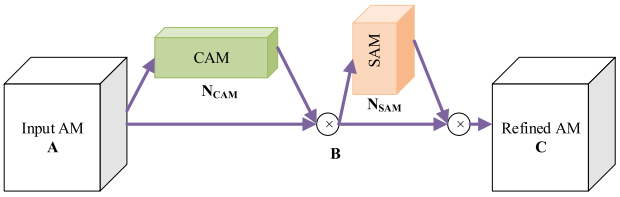


Fig. 4. Relation of CBAM and CAM and SAM.

channel attention and spatial attention, respectively. Thus, the AI model using attention mechanism (i) will focus on those really important and salient features, (ii) performs more effective, and (iii) becomes more robust to noisy inputs.

CBAM, a typical attention mechanism-based module, entails two sequential submodules: channel attention module (CAM) and spatial attention module (SAM). Figure 4 shows the overall relation of CBAM and CAM and SAM.

Suppose we have a temporary activation map. Its input is $\mathbf{A} \in \mathbb{R}^{C\times H\times W}$. The CBAM will apply 1D CAM $\mathbf{N_{CAM}} \in \mathbb{R}^{C\times 1\times 1}$ and a 2D SAM $\mathbf{N_{SAM}} \in \mathbb{R}^{1\times H\times W}$ in sequence to the input $\mathbf{A}$, as shown in Figure 4. Thus, we have the channel-refined activation map and the final activation map as:

$$\begin{cases} \mathbf{B} = \mathbf{N_{CAM}}\left(\mathbf{A}\right) \otimes \mathbf{A} \\ \mathbf{C} = \mathbf{N_{SAM}}\left(\mathbf{B}\right) \otimes \mathbf{B} \end{cases} \tag{7}$$

where $\otimes$ stands for the element-wise multiplication. If the two operands are not with the same dimension, then the values are broadcasted (copied) in such ways the spatial attentional values are broadcasted along the channel dimension, and the channel attention values are broadcasted along the spatial dimension.

First, we define the CAM. Both max pooling f_{mp} and average pooling f_{mp} are utilized, generating two features $\mathbf{D_{ap}}$ and $\mathbf{D_{mp}}$.

$$\begin{cases} \mathbf{D_{ap}} = f_{ap}\left(\mathbf{A}\right) \\ \mathbf{D_{mp}} = f_{mp}\left(\mathbf{A}\right) \end{cases} \tag{8}$$

Both are then forwarded to a shared multi-layer perceptron (MLP) to generate the output features, which are then merged using element-wise summation $\oplus$. The merged sum is finally sent to the sigmoid function σ. Mathematically

$$\mathbf{N_{CAM}}\left(\mathbf{A}\right) = \sigma\left\{MLP\left[\mathbf{D_{ap}}\right] \oplus MLP\left[\mathbf{D_{mp}}\right]\right\} \tag{9}$$

To reduce the parameter resources, the hidden size of MLP is set to $\mathbb{R}^{C/r\times 1\times 1}$, where r is defined as the reduction ratio. Suppose $\mathbf{W_0} \in \mathbb{R}^{C/r\times C}$ and $\mathbf{W_1} \in \mathbb{R}^{C\times C/r}$ stand for the MLP weights, respectively, we can rephrase equation (9) as

$$\mathbf{N_{CAM}}\left(\mathbf{A}\right) = \sigma\left\{\mathbf{W_1}\left[\mathbf{W_0}\left(\mathbf{D_{ap}}\right)\right] \oplus \mathbf{W_1}\left[\mathbf{W_0}\left(\mathbf{D_{mp}}\right)\right]\right\} \tag{10}$$

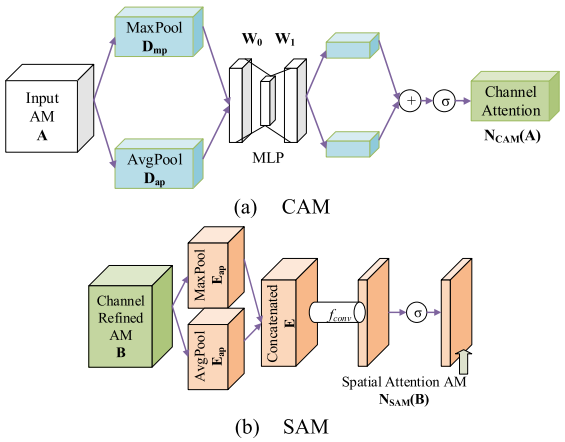


Fig. 5. Flowchart of two blocks in SAM.

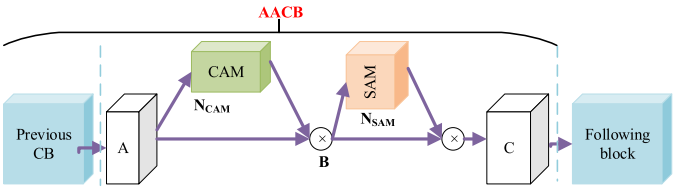


Fig. 6. Illustration of AACB: Integration of CBAM with VGG-style base network.

Note $\mathbf{W_0}$ and $\mathbf{W_1}$ are shared by both $\mathbf{D_{ap}}$ and $\mathbf{D_{mp}}$. Figure 5(a) shows the flowchart of CAM. Note that squeeze-and-excitation (SE) [24] method is similar to CAM.

Second, we describe the detailed procedures of SAM. The spatial attention module $\mathbf{N_{SAM}}$ is a complementary step to the previous channel attention module $\mathbf{N_{CAM}}$. The average pooling f_{ap} and max pooling f_{mp} are applied to the channel-refined activation map $\mathbf{B}$, and we get

$$\begin{cases} \mathbf{E_{ap}} = f_{ap}\left(\mathbf{B}\right) \\ \mathbf{E_{mp}} = f_{mp}\left(\mathbf{B}\right) \end{cases} \tag{11}$$

Both $\mathbf{E_{ap}}$ and $\mathbf{E_{mp}}$ are two dimensional activation maps: $\mathbf{E_{ap}} \in \mathbb{R}^{1\times H\times W} \wedge \mathbf{E_{mp}} \in \mathbb{R}^{1\times H\times W}$. They are concatenated together along the channel dimension as $\mathbf{E} = \mathbf{concat}\left(\mathbf{E_{ap}}, \mathbf{E_{mp}}\right)$.The concatenated activation map is then passed into a standard 7×7 convolution f_{conv} and followed by a sigmoid function σ. In all, we have: $\mathbf{N_{SAM}}\left(\mathbf{B}\right) = \sigma\left\{f_{conv}\left[\mathbf{E}\right]\right\}$. The flowchart of SAM is drawn in Figure 5(b). The output $\mathbf{N_{SAM}}\left(\mathbf{B}\right)$ is then element-wisely multiplied with $\mathbf{B}$, as shown in Equation (7).

The attention mechanism CBAM is embedded into our VGG-style base network. The integration is shown in Figure 6. For the activation map A of each convolution block, the two consecutive attention modules (channel and spatial) are added, and the refined features C are sent to the next CBs. Here AACB means the attention attached convolution block, which is composed of one CB and following attention modules.

Table IV itemizes the structure of proposed attention-based VGG-style network for COVID-19 (AVNC) model, where the structure is almost the same as previous VSBN model, except that each CB are replaced by corresponding AACB.

The size of input is $256\times 256\times 1$. The first AACB contains 2 repetitions of 32 kernels with size of 3×3, followed by max pooling with size of 2×2, and attention modules (both channel

TABLE IV
STRUCTURE OF PROPOSED 12-LAYER AVNC MODEL

Index	Name	Kernel Parameter	Size of Output
1	Input		256x256x1
2	AACB-1	[3x3, 32]x2	128x128x32
3	AACB-2	[3x3, 32]x2	64x64x32
4	AACB-3	[3x3, 64]x2	32x32x64
5	AACB-4	[3x3, 64]x2	16x16x64
6	AACB-5	[3x3, 128]x2	8x8x128
7	Flatten		1x1x8192
8	FCL-1	150x8192, 150x1	1x1x150
9	FCL-2	4x150, 4x1	1x1x4
10	Softmax		1x1x4

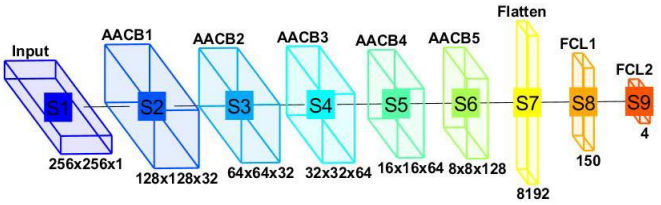


Fig. 7. AMs of AVNC.

and spatial). The output after 1st AACB is $128 \times 128 \times 32$. Consequently, the output of second to fifth AACB is $64 \times 64 \times 32$, $32 \times 32 \times 64$, $16 \times 16 \times 64$, and $8 \times 8 \times 128$, respectively (See S1-S6 in Figure 7).

At the flatten stage, the feature map is vectorized as $1 \times 1 \times 8192$, where from the previous AM we get $8 \times 8 \times 128 = 8192$. This 8192-element vector is then submitted to 2-layer FCLs with 150 neurons and 4 neurons, respectively (See S7-S9 in Figure 7). The softmax function is defined as $\sigma : \mathbb{R}^K \mapsto \mathbb{R}^K$. Suppose the input $z = (z_1, \ldots, z_K) \in \mathbb{R}^K$, we have $\sigma(z)_i = \exp(z_i) / \sum_{j=1}^{K} \exp(z_j), \forall i = 1, \ldots, K$.

As default, batch normalization and dropout are embedded in our proposed AVNC model. Gradient-weighted class activation mapping (Grad-CAM) [25] is employed to give explanations that how our AVNC model makes the decision. Grad-CAM demystifies the AI model by utilizing the gradient of the categorization score regarding the convolutional features. S6 in Figure 7 will be used to generate the heatmap where red colors stand for the most interesting area our AVNC model pay attentions to.

D. Improved Multiple-Way Data Augmentation

This small four-category dataset (1164 images) causes our AVNC model easily overfitted. To help prevent the overfitting take place, an improved multiple-way data augmentation (IMDA) inspired from Ref. [16] was proposed. Ref. [16] proposed a 14-way MDA, which includes 7 data augmentation techniques to the original image $o(i)$ and its horizontal-mirrored (HM) image $o_{HM}(i)$.

Our IMDA method includes a new DA method: salt-and-pepper noise (SPN) to both $o(i)$ and $o_{HM}(i)$. The SPN has already been successfully proven effective in medical image recognition.

First, eight DA transforms are utilized. They can be divided into three types: (i) geometric; (ii) photometric; and (iii) noise-injection. We use $d_m^{DA}, m = 1, \ldots, 8$ to denote each DA operation. Note each we assume DA operation

TABLE V
DATASET SPLITTING

Category	Non-test (10-fold CV)	Test (S runs)	Total
COVID-19	$\lvert O_1^n \rvert = 227$	$\lvert O_1^t \rvert = 57$	$\lvert O_1 \rvert = 284$
CAP	$\lvert O_2^n \rvert = 225$	$\lvert O_2^t \rvert = 56$	$\lvert O_2 \rvert = 281$
SPT	$\lvert O_3^n \rvert = 234$	$\lvert O_3^t \rvert = 59$	$\lvert O_3 \rvert = 293$
HC	$\lvert O_4^n \rvert = 245$	$\lvert O_4^t \rvert = 61$	$\lvert O_4 \rvert = 306$

d_m^{DA} yields $P(m)$ new generated images. The first temporary enhanced dataset $\mathbf{D_1}(i)$ is obtained by $\mathbf{D_1}(i) = f^C\left\{d_m^{DA}[o(i)] \,|m = 1, \cdots, 8\right\}$, where f^C stands for concatenation function. The number of images in $|\mathbf{D_1}(i)| = \sum_{m=1}^{8} P(m)$.

Second, the horizontal image is generated by HM function f^{HM} as $o_{HM}(i) = f^{HM}[o(i)]$.

Third, eight DA transforms are carried upon the mirrored image $o_{HM}(i)$, and we obtain the second temporary enhanced dataset $\mathbf{D_2}(i) = f^C\left\{d_m^{DA}[o_{HM}(i)] \,|m = 1, \cdots, 8\right\}$, where $|\mathbf{D_2}(i)| = |\mathbf{D_1}(i)|$.

Fourth, the raw image, the mirrored image, the first and second temporary enhanced dataset are all combined. The final dataset is

$$\mathbf{O}(i) = f^C \begin{bmatrix} o(i) & o_{HM}(i) \\ \mathbf{D_1}(i) & \mathbf{D_2}(i) \end{bmatrix} \tag{12}$$

In this study, we set $P(1) = \cdots = P(8) = P = 30$. We tested greater value of W, but it does not bring significant improvement. Hence, one image will generate $|\mathbf{O}(i)| = |o(i)| + |o_{HM}(i)| + |\mathbf{D_1}(i)| + |\mathbf{D_2}(i)| = 1 + 1 + 8P + 8P = 16P + 2$ images (including raw image $o(i)$).

Table V itemizes the non-test, and test set for each category. The whole dataset O contains four non-overlapping categories $O = \{O_k | k = 1, \cdots, 4\}$. For each category, the corresponding dataset O_k will be split into non-test set and test set $O_k \rightarrow \left\{O_k^n, O_k^t\right\}, k = 1, 2, 3, 4$, where the superscript n and t stand for non-test and test, respectively.

Our algorithm is composed of two phases. At Phase I, 10-fold cross validation was used for validation on the non-test set (See O_k^n in Table V) to select the best hyperparameters and best network structure. Due to the page limit, results on Phase I were not reported in Section IV. Afterwards at Phase II, we train our model using non-test set (See O_k^t in Table V) S times with different initial seeds, and obtain the test results over the test set $O_k^t, k = 1, \cdots, 4$. After combining the S runs, we obtain a summation of test confusion matrix (TCM) $\mathbf{E^t}$. The ideal TCM only have entries on the diagonal line as

$$\mathbf{E^t} = S \times \begin{bmatrix} \left|O_1^t\right| & 0 & 0 & 0 \\ 0 & \left|O_2^t\right| & 0 & 0 \\ 0 & 0 & \left|O_3^t\right| & 0 \\ 0 & 0 & 0 & \left|O_4^t\right| \end{bmatrix} \tag{13}$$

where we can observe that $\mathbf{E^t}(i, j) = 0, \forall i \neq j$, indicating there is no prediction mistakes for the TCM of an ideal AI model. For this proposed AVNC model, the performance per class (PPC) will be calculated. For each class $k = 1, \cdots, 4$, that class k is set to positive, and other three classes $[1, 2, 3, 4] - k$ are negative.

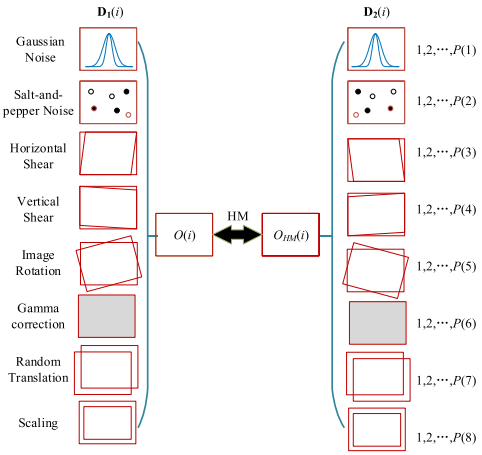


Fig. 8. Diagram of proposed IMDA method.

TABLE VI
EFFECT OF PROPOSED IMDA

Model	C	Sen	Prc	F1	Model	C	Sen	Prc	F1
IMDA (Ours)	1	97.19	96.18	96.68	No-DA	1	94.91	95.58	95.25
	2	97.68	96.64	97.16		2	92.68	91.86	92.27
	3	95.25	97.57	96.40		3	93.9	93.42	93.66
	4	97.38	97.06	97.22		4	94.59	95.21	94.9
	F_1^μ			96.87		F_1^μ			94.03

The PPCs of sensitivity (S), precision (P), and F1 score (F_1) are calculated as: $S(k) = TP(k)/[TP(k) + FN(k)]$, $P(k) = TP(k)/[TP(k) + FP(k)]$, $F_1(k) = 2 \times P(k) \times S(k)/[P(k) + S(k)]$.

There are two types of overall F1 scores that can measure over all categories: micro-averaged and macro-averaged. This study chooses the micro-averaged F1 (F_1^μ) because our dataset is slightly unbalanced $F_1^\mu = \frac{2 \times P^\mu \times S^\mu}{P^\mu + S^\mu}$, where $S^\mu = \frac{\sum_{k=1}^{4} TP(k)}{\sum_{k=1}^{4} TP(k)+FN(k)} \& P^\mu = \frac{\sum_{k=1}^{4} TP(k)}{\sum_{k=1}^{4} TP(k)+FP(k)}$.

IV. Results and Discussions

A. Effectiveness of IMDA

Figure 9 shows the $\mathbf{D_1}$ result of proposed IMDA. Due to the page limit, the o_{HM} and $\mathbf{D_2}$ are not displayed here. The original preprocessed image is shown in Figure 12(a). Afterwards, we show the test results of using IMDA and not using IMDA in Table VI, where "No-DA" means not using any DA technique.

We can observe from Table VI that the performance will decrease if we remove the IMDA from our algorithm. The micro-average F1 score over four classes using IMDA is $F_1^\mu = 96.87\%$, which will crease to only 94.03% if no DA method is used. This comparison shows the effectiveness of IMDA.

B. Configuration of Attention

We compare different configuration of attention here. We test five configurations: (i) No attention (NA), i.e., proposed VSBN (ii) squeeze-and-excitation (SE) [24] module; (iii) CAM and SAM in parallel (CSP); (iv) First SAM Second CAM (FSSC); and (iv) First CAM Second SAM, viz., CBAM used in this proposed AVNC. The results are presented in Table VII.

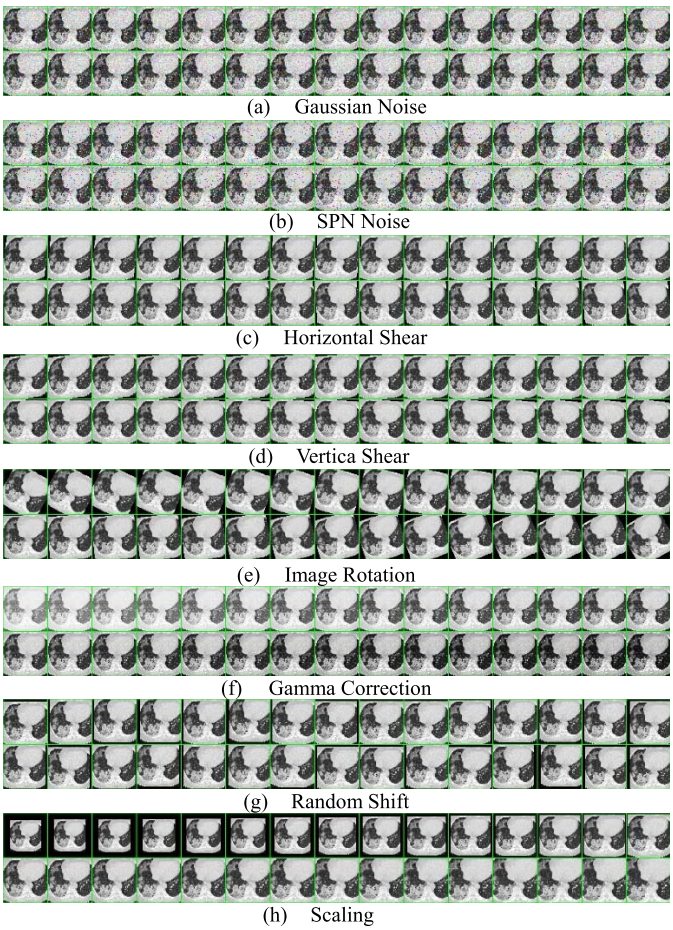


Fig. 9. D_1 result of IMDA.

TABLE VII
COMPARISON OF DIFFERENT CONFIGURATIONS OF ATTENTION

Model	C	Sen	Prc	F1	Model	C	Sen	Prc	F1
NA	1	91.40	93.04	92.21	SE [24]	1	95.26	95.60	95.43
	2	92.32	94.17	93.24		2	96.07	94.72	95.39
	3	94.07	93.91	93.99		3	95.08	94.44	94.76
	4	95.90	92.86	94.35		4	93.61	95.17	94.38
	F_1^μ			93.48		F_1^μ			94.98
CSP	1	95.44	96.97	96.20	FSSC	1	95.09	96.79	95.93
	2	95.89	95.72	95.81		2	96.25	96.77	96.51
	3	97.80	95.53	96.65		3	97.29	96.15	96.71
	4	94.92	95.86	95.39		4	96.89	95.94	96.41
	F_1^μ			96.01		F_1^μ			96.39
CBAM (Ours)	1	97.19	96.18	96.68					
	2	97.68	96.64	97.16					
	3	95.25	97.57	96.40					
	4	97.38	97.06	97.22					
	F_1^μ			96.87					

From Table VII, we can observe that NA method achieves the worst result of only $F_1^\mu = 93.48$, indicating that the VSBN model without any attention mechanism come across somewhat misclassification over the test set. Then we can see SE [24] obtains the second worst result of $F_1^\mu = 94.98$. The reason is SE uses global average-pooled features which are suboptimal to infer fine channel attention. The CSP method get the mediocre result of $F_1^\mu = 96.01$, which is better than SE [24]. The reason is (i) SE [24] misses the spatial attention;

TABLE VIII
COMPARISON TO STATE-OF-THE-ART ALGORITHMS

Model	C	Sen	Prc	F1	Model	C	Sen	Prc	F1
ELMBA [5]	1	62.63	67.61	65.03	RCBBO [6]	1	71.93	84.19	77.58
	2	64.29	65.10	64.69		2	72.86	72.73	72.79
	3	71.86	66.77	69.22		3	73.56	76.41	74.96
	4	63.93	63.52	63.73		4	80.66	68.91	74.32
	F_1^μ			65.71		F_1^μ			74.85
GN [7]	1	81.75	83.07	82.40	RN18 [8]	1	82.81	82.66	82.73
	2	86.07	82.39	84.19		2	81.07	74.43	77.61
	3	80.51	84.07	82.25		3	74.24	76.98	75.58
	4	81.31	80.13	80.72		4	82.13	86.38	84.20
	F_1^μ			82.36		F_1^μ			80.04
5LSP [9]	1	93.16	91.39	92.27	7LSP [10]	1	89.47	93.58	91.48
	2	93.21	91.10	92.14		2	93.93	92.44	93.18
	3	91.53	91.53	91.53		3	93.73	95.18	94.45
	4	86.56	90.10	88.29		4	95.08	91.34	93.17
	F_1^μ			91.03		F_1^μ			93.09
FCONet [11]	1	92.28	95.64	93.93	DLA [12]	1	87.89	91.59	89.70
	2	96.79	94.43	95.59		2	80.89	85.47	83.12
	3	94.75	95.88	95.31		3	83.22	82.11	82.66
	4	94.92	92.94	93.92		4	92.30	85.95	89.01
	F_1^μ			94.68		F_1^μ			86.18
CSS [13]	1	94.04	92.25	93.14	DeCov Net [14]	1	91.05	90.58	90.81
	2	93.75	95.11	94.42		2	93.75	90.99	92.35
	3	91.36	93.58	92.45		3	90.51	86.97	88.70
	4	94.43	92.75	93.58		4	88.69	95.58	92.01
	F_1^μ			93.39		F_1^μ			90.94
SMO [15]	1	97.02	92.63	94.77	AVNC (Ours)	1	97.19	96.18	96.68
	2	89.11	95.23	92.07		2	97.68	96.64	97.16
	3	94.92	94.92	94.92		3	95.25	97.57	96.40
	4	94.26	92.89	93.57		4	97.38	97.06	97.22
	F_1^μ			93.86		F_1^μ			96.87

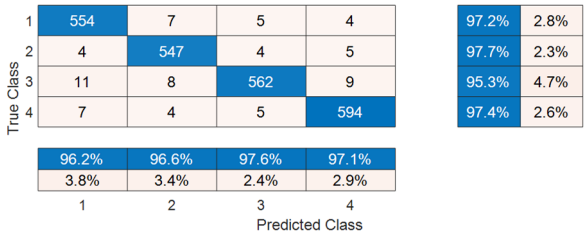


Fig. 10. Confusion matrix of proposed AVNC approach.

(ii) CSP uses SAM and CAM parallelly, and the CAM uses both max pooling and average pooling.

Finally, our CBAM and the FSSC methods achieves the best two results, because both use CAM and SAM in sequence. Hence, compared to CSP, we can conclude the sequential link can get better results than parallel link (CSP). The difference of CBAM and FSSC lies in that CBAM is "First CAM Second SAM", and FSSC is "First SAM Second CAM". This experiment shows the CBAM method yields better result than FSSC method, which is in line with Ref. [21].

C. Comparison to State-of-the-Art Diagnosis Methods

We finally compare this proposed AVNC approach with state-of-the-art approaches. First, the confusion matrix of proposed AVNC is displayed in Figure 10. Remember in Table V and Equation (13), the ideal confusion matrix is a diagonal matrix in which the elements along diagonal line is $10 \times [57, 56, 59, 61]$. Figure 10 shows that in total 554 COVID-19 cases are classified correctly; nevertheless, 7, 5, and 4 cases are wrongly misclassified to CAP, SPT, and HC,

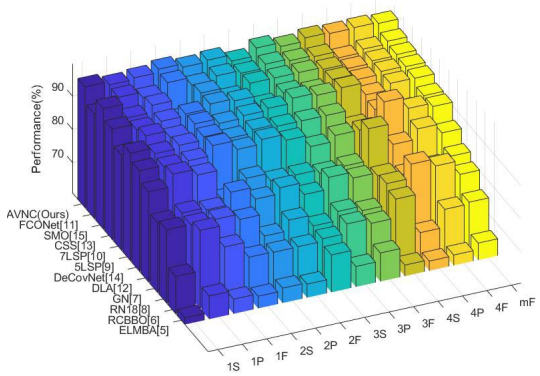


Fig. 11. Comparison (kS: sensitivity of class k, kP: precision of class k, kF: F1 score of class k, mF: micro-averaged F1 score).

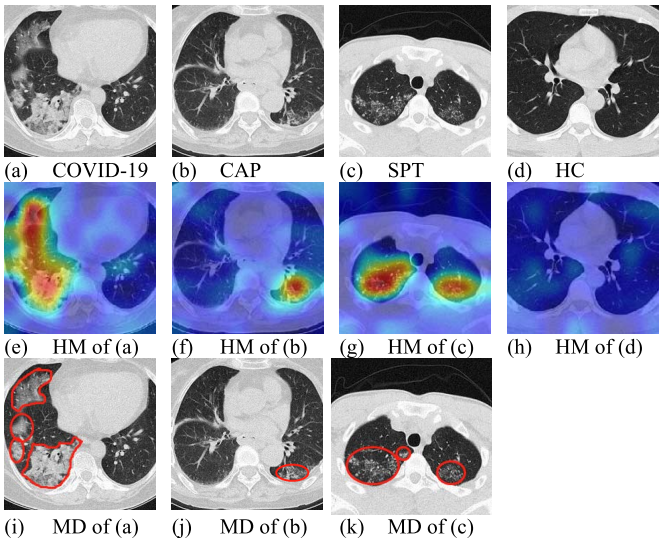


Fig. 12. Heatmap generated by Grad-CAM and our AVNC model (HM: heatmap).

respectively. Similarly, we can get the case-by-case prediction results for 2nd, 3rd, and 4th categories.

This proposed AVNC method was compared with 11 state-of-the-art approaches: ELMBA [5], RCBBO [6], GN [7], RN18 [8], 5LSP [9], 7LSP [10], FCONet [11], DLA [12], CSS [13], DeCovNet [14], SMO [15]. The detailed comparison results are shown in Table VIII.

Figure 11 shows the comparison results of all the 12 algorithms in terms of the overall performance F_1^μ. It shows this proposed AVNC achieves the highest value of $F_1^\mu = 96.87$, which again indicates the superiority of our algorithm.

D. Explainability of Proposed Model

Figure 12(a-d) displays the four samples of COVID-19, CAP, SPT, and HC, respectively. Their corresponding heatmaps (HMs) generated by Grad-CAM are displayed in Figure 12(e-h). Meanwhile, we showed the manual delineation (MD) results in Figure 12(i-k).

V. CONCLUSION

In this paper, a novel AVNC model was proposed, which utilizes the proposed VSBN as backbone and integrates the attention mechanism and an improved multiple-way data augmentation.

Experimental results in Section IV showed that this proposed AVNC achieved the sensitivity/precision/F1 per class all above 95%. Particularly, AVNC yielded a micro-averaged F1 score of 96.87%, which is higher than 11 state-of-the-art approaches.

We shall attempt to (i) expand our dataset to include more samples and to include more categories; (ii) integrate our AVNC with IoT [26], [27], cloud computing, edge computing [28], and online web services.

REFERENCES

[1] A. G. Hadi, M. Kadhom, N. Hairunisa, E. Yousif, and S. A. Mohammed, "A review on COVID-19: Origin, spread, symptoms, treatment, and prevention," *Biointerface Res. Appl. Chem.*, vol. 10, pp. 7234–7242, Dec. 2020.

[2] G. S. Campos *et al.*, "Ion torrent-based nasopharyngeal swab metatranscriptomics in COVID-19," *J. Virolog. Methods*, vol. 282, Aug. 2020, Art. no. 113888.

[3] C. Brammer *et al.*, "Qualitative review of early experiences of off-site COVID-19 testing centers and associated considerations," *Healthcare-J. Del. Sci. Innov.*, vol. 8, no. 3, Sep. 2020, Art. no. 100449.

[4] Y. Li and L. Xia, "Coronavirus disease 2019 (COVID-19): Role of chest CT in diagnosis and management," *Amer. J. Roentgenol.*, vol. 214, no. 6, pp. 1280–1286, Jun. 2020.

[5] S. Lu *et al.*, "A pathological brain detection system based on extreme learning machine optimized by bat algorithm," *CNS Neurolog. Disorders-Drug Targets*, vol. 16, no. 1, pp. 23–29, Jan. 2017.

[6] S. Wang *et al.*, "Pathological brain detection via wavelet packet Tsallis entropy and real-coded biogeography-based optimization," *Fundamenta Informaticae*, vol. 151, nos. 1–4, pp. 275–291, Mar. 2017.

[7] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Boston, MA, USA, Jun. 2015, pp. 1–9.

[8] M. Guo and Y. Du, "Classification of thyroid ultrasound standard plane images using ResNet-18 networks," in *Proc. IEEE 13th Int. Conf. Anti-Counterfeiting, Secur., Identificat. (ASID)*, Xiamen, China, Oct. 2019, pp. 324–328.

[9] Y.-D. Zhang, S. C. Satapathy, S. Liu, and G.-R. Li, "A five-layer deep convolutional neural network with stochastic pooling for chest CT-based COVID-19 diagnosis," *Mach. Vis. Appl.*, vol. 32, no. 1, p. 14, Jan. 2021.

[10] Y.-D. Zhang, S. C. Satapathy, L.-Y. Zhu, J. M. Gorriz, and S.-H. Wang, "A seven-layer convolutional neural network for chest CT based COVID-19 diagnosis using stochastic pooling," *IEEE Sensors J.*, early access, Sep. 22, 2020, doi: 10.1109/JSEN.2020.3025855.

[11] H. Ko *et al.*, "COVID-19 pneumonia diagnosis using a simple 2D deep learning framework with a single chest CT image: Model development and validation," *J. Med. Internet Res.*, vol. 22, p. 13, Jun. 2020.

[12] Q. Q. Ni *et al.*, "A deep learning approach to characterize 2019 coronavirus disease (COVID-19) pneumonia in chest CT images," *Eur. Radiol.*, vol. 30, no. 12, pp. 6517–6527, 2020.

[13] J. P. Cohen *et al.*, "Predicting COVID-19 pneumonia severity on chest X-ray with deep learning," *Cureus*, vol. 12, p. e9448, Jul. 2020.

[14] 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.

[15] M. Togacar, B. Ergen, and Z. Comert, "COVID-19 detection using deep learning models to exploit social mimic optimization and structured chest X-ray images using fuzzy color and stacking approaches," *Comput. Biol. Med.*, vol. 121, Jun. 2020, Art. no. 103805.

[16] S.-H. Wang, V. V. Govindaraj, J. M. Górriz, X. Zhang, and Y.-D. Zhang, "COVID-19 classification by FGCNet with deep feature fusion from graph convolutional network and convolutional neural network," *Inf. Fusion*, vol. 67, pp. 208–229, Mar. 2021.

[17] S.-H. Wang, D. R. Nayak, D. S. Guttery, X. Zhang, and Y.-D. Zhang, "COVID-19 classification by CCSHNet with deep fusion using transfer learning and discriminant correlation analysis," *Inf. Fusion*, vol. 68, pp. 131–148, Apr. 2021.

[18] B. S. Rao, "Accurate Leukocoria predictor based on deep VGG-Net CNN technique," *IET Image Process.*, vol. 14, no. 10, pp. 2241–2248, Aug. 2020.

[19] S. Tridgell *et al.*, "Unrolling ternary neural networks," *ACM Trans. Reconfigurable Technol. Syst.*, vol. 12, p. 23, Nov. 2019.

[20] M. Ahmad, M. Abdullah, and D. Han, "A novel encoding scheme for complex neural architecture search," in *Proc. 34th Int. Tech. Conf. Circuits/Syst., Comput. Commun. (ITC-CSCC)*. JeJu, South Korea: IEEE, Jun. 2019, pp. 1–4.

[21] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Munich, Germany, 2018, pp. 3–19.

[22] C. Lucero *et al.*, "Unconscious number discrimination in the human visual system," *Cerebral Cortex*, vol. 30, no. 11, pp. 5821–5829, Oct. 2020.

[23] H. Larochelle and G. E. Hinton, "Learning to combine foveal glimpses with a third-order Boltzmann machine," in *Proc. Neural Inf. Process. Syst. (NeurIPS)*. Vancouver, BC, Canada: Hyatt Regency, 2010, pp. 1243–1251.

[24] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze- and-excitation networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 8, pp. 2011–2023, Aug. 2020.

[25] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," *Int. J. Comput. Vis.*, vol. 128, no. 2, pp. 336–359, Feb. 2020.

[26] Y. Zhang, R. Wang, M. S. Hossain, M. F. Alhamid, and M. Guizani, "Heterogeneous information network-based content caching in the Internet of vehicles," *IEEE Trans. Veh. Technol.*, vol. 68, no. 10, pp. 10216–10226, Oct. 2019.

[27] Y. Zhang, Y. Li, R. Wang, M. S. Hossain, and H. Lu, "Multi-aspect aware session-based recommendation for intelligent transportation services," *IEEE Trans. Intell. Transp. Syst.*, early access, May 14, 2020, doi: 10.1109/TITS.2020.2990214.

[28] Y. Zhang, Y. Li, R. Wang, J. Lu, X. Ma, and M. Qiu, "PSAC: Proactive sequence-aware content caching via deep learning at the network edge," *IEEE Trans. Netw. Sci. Eng.*, vol. 7, no. 4, pp. 2145–2154, Oct. 2020, doi: 10.1109/TNSE.2020.2990963.

Shui-Hua Wang (Senior Member, IEEE) received the bachelor's degree in information sciences from Southeast University in 2008, the master's degree in electrical engineering from the City College of New York in 2012, and the Ph.D. degree in electrical engineering from Nanjing University in 2017. She is working as a Research Associate with the University of Leicester, U.K.

Steven Lawrence Fernandes (Senior Member, IEEE) began his postdoctoral research with the University of Alabama at Birmingham. He also conducted postdoctoral research with the University of Central Florida. His current area of research is focused on using artificial intelligence techniques to extract useful patterns from big data.

Ziquan Zhu received the B.E. degree in engineering from Jilin University, Changchun, China, in 2018, and the M.E. degree in construction from the University of Florida, Gainesville, FL, USA, in 2020.

Yu-Dong Zhang (Senior Member, IEEE) received the B.E. degree in information sciences and the M.Phil. degree in communication and information engineering from the Nanjing University of Aeronautics and Astronautics in 2004 and 2007, respectively, and the Ph.D. degree in signal and information processing from Southeast University in 2010. He serves as a Professor with the University of Leicester.