Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Pattern Recognition Letters 152 (2021) 311–319

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

Two-Stage Deep Learning Framework for Discrimination between COVID-19 and Community-Acquired Pneumonia from Chest CT scans

Mohamed Abdel-Basset[a], Hossam Hawash[a,*], Nour Moustafa[b], Osama M. Elkomy[a]

[a] Faculty of Computers and Informatics, Zagazig University, Zagazig, Sharqiyah, 44519, Egypt
[b] School of Engineering and Information Technology, University of New South Wales @ ADFA, Canberra, ACT 2600, Australia

ARTICLE INFO

Article history:
Received 31 May 2021
Revised 1 September 2021
Accepted 25 October 2021
Available online 29 October 2021

Edited by Maria De Marsico

ABSTRACT

COVID-19 stay threatening the health infrastructure worldwide. Computed tomography (CT) was demonstrated as an informative tool for the recognition, quantification, and diagnosis of this kind of disease. It is urgent to design efficient deep learning (DL) approach to automatically localize and discriminate COVID-19 from other comparable pneumonia on lung CT scans. Thus, this study introduces a novel two-stage DL framework for discriminating COVID-19 from community-acquired pneumonia (CAP) depending on the detected infection region within CT slices. Firstly, a novel U-shaped network is presented to segment the lung area where the infection appears. Then, the concept of transfer learning is applied to the feature extraction network to empower the network capabilities in learning the disease patterns. After that, multi-scale information is captured and pooled via an attention mechanism for powerful classification performance. Thirdly, we propose an infection prediction module that use the infection location to guide the classification decision and hence provides interpretable classification decision. Finally, the proposed model was evaluated on public datasets and achieved great segmentation and classification performance outperforming the cutting-edge studies.

1. Introduction

Recently, the human being around the world has been conquered with emergent Coronaviridae species known as severe acute respiratory syndrome coronavirus 2 (SARS-COV-2) [1]. The already acknowledged coronaviruses might be considered as a tip of the iceberg, with hypothetically extra unrevealed zoonotic effects to be discovered. Therefore, it is vitally important to early recognize infected persons for performing preventive containment procedures and medical treatment processes. Although many criteria enable successful diagnosis of COVID-19 for individuals, the clinical laboratory tools that depend on virus nucleic acid sequencing and Reverse-transcription-polymerase chain reaction (RT-PCR) suffers from many shortcomings and deficiencies; for instance, nucleic acid check has been reliant on numerous rate-bounding aspects, involving accessibility and magnitude of the testing apparatus in the affected area [2]. More significantly, the superiority, constancy, and reproducibility of the examination apparatus are controversial [3]. Like many pneumonia diseases, the diagnosis of COVID-19 can be performed based on computed tomography (CT) scan representing the lungs and soft tissues. Even though distinctive CT scans might assist early examination of suspicious COVID-19 cases, viral types of pneumonia have almost identical images that interfere with other contagious and blazing lung diseases [4]. Hence, it is troublesome for radiologists to discriminate COVID-19 from different viral types of pneumonia, and even from Community-acquired pneumonia (CAP) [5]. For example, as presented in Fig. 1, the apparition of COVID-19 exhibits high similarity with CAP in lung CT scan, which in turn sophisticate the process of CVOID-19 diagnosis [6]. The present clinical environment requires an intelligent and consistent diagnosis approach for COVID-19 to minimize the clinicians' burdens and finetune the disease diagnosis efficiency. Nevertheless, it is challenging to develop such an approach, due to wide differences in the sizes, shapes, and locations of infection in the lung CT scan, as depicted in Fig. 1. It looks problematic to build an efficient technique to learn from the complex features of pneumonia infections utilizing just the conventional techniques of computer vision [7].

Recent advances of convolutional neural networks (CNNs) have come up with a sequence of innovations in the area of natural image analysis [8] and other computer vision tasks [9]. The CNNs can extract and learn an improved visual feature representation, eliminating the necessity for manual descriptions. Such advances provide extra confirmation that better performance could be achieved

* Corresponding author.
E-mail address: hossamreda@zu.edu.eg (H. Hawash).

https://doi.org/10.1016/j.patrec.2021.10.027
0167-8655/

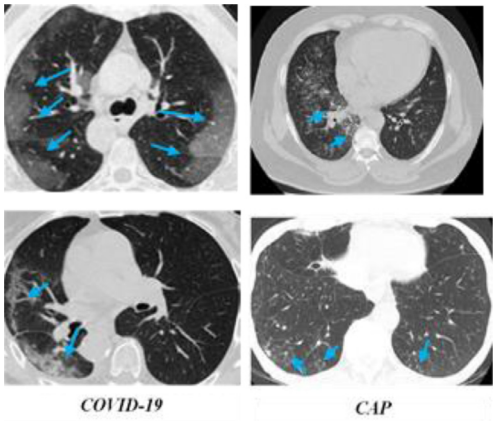

Fig. 1. Samples of COVID-19 and CAP as presented in the left and the right column, respectively. The main infection regions are specified with blue arrow.

with deeper architecture. Thus, it is reasonable that exploiting deeper CNN networks to realize such improvements in the diagnosis COVID-19. Currently, it is possible to use residual learning blocks to develop superior deep models with a number of layers higher than 100 [10].

1.1. Challenges and Research gaps

The design of efficient and reliable computer-aided-diagnosis (CAD) for COVID-19 screening still facing multiple challenges and research gaps that can be summarized as follow:

It is hard and challenging for doctors to identify a various form of pneumonia from a large number of CT images, especially during the COVID-19 outbreak. This confusing issue stems from the resemblance and difference of inter-class of infection regions, as shown in Fig. 1.

The pneumonia lung image still encompasses a huge fragment of regions irrelevant to infection, that have a complex discrepancy of tissues. Noticeably, the infection of irrelevant regions has an excessively negative effect on the model's performance. This is more sophisticated than the recognition of substances with natural images.

1.2. Novelty and Contributions

To tackle the beforementioned issues, this work presents a novel infection-aware approach for diagnosing the infection of COVID-19. It comprises the next building blocks:

At first stage, A novel U-shaped architecture (GR-U-Net) is proposed for segmenting the lung regions, where the encoder path is built using pre-trained Efficient-Net, the bottleneck path is build using a densely connected network that provides a collective learning paradigm and prevents gradient vanishing.

The proposed GR-U-Net redesigns the skip connection a time-efficient bidirectional convolutional gated recurrent unit (BConvGRU) to capture Spatial-Semantic features from the encoder and pass it to the decoder path. while the squeeze and excite operation is applied to perform spatial and channel feature recalibration in the decoding layer.

At the second stage, advanced pre-trained EfficientNet-B7 is employed as a robust feature extractor for extracting the disease features from the received lung images. multi-scale information is extracted at a different layer of the extractor and then pooled to effectively capture different sizes of infection.

Motivate by the recent breakthrough in deep attention mechanisms (AM), particularly by the self-attention mechanism in various lesion classification task [11], we develop an interactive AM in that utilize the infection ground truth (GT) to enhance network attention, and hence provide extra concentration on the infection areas to improve model interpretability.

1.3. Related Works

As previously stated, Deep Learning approaches have been performing a significant role in facilitating the recognition of COVID-19. For example, Kang et al [12], introduce a novel multi-view representation learning approach to discriminate between COVID-19 and CAP using a group of features captured from patients' CT scans. It learns an integrated latent representation to encode information from diverse aspects of features and thus improve diagnosis accuracy. Ouyang et al. [13] proposed a dual sampling model that classifies the COVID-19 patients and the CAP patients in chest CT using 3D CNN accompanied with an attention module to concentrate on the lung infection areas to determine the final classification decision. Besides, Wang et al [14] employed two residual convolution modules for discrimination and localization of infection in x-ray images with the main aim to efficiently discriminate COVID-19 from CAP infections. Additionally, Marques et al [15] applied EfficientNet architecture to recognize the COVID-19 case from normal or pneumonia cases. In a similar way, the authors of [16] employed a convolutional network based on ResNet50 for classifying the COVID-19 and CAP from chest CT scans. Furthermore, in [17], DeepCOVID-XR was introduced as an ensemble of convolution models for detecting COVID-19 features from frontal chest radiographs. Bai et al [18] introduced Efficient Net B4 network for classifying COVID-19 and other pneumonia for each patient, whereas CT slices were introduced for lung segmentation, then passed to two fully connected layers to pool the slices.

1.4. Study Organization

The remaining of this work is presented as follows: Detailed explanations and information corresponding to our proposed framework are presented in Section 2. The proposed experimental conditions of this work are debated in Section 3, the results, the comparisons, and the analysis of outcomes are debated in Section 4. Section 5 argues the main limitation of the current work. To end, the conclusions and future research direction are explained in Section 6.

2. Proposed Approach

This section provides a detailed explanation of the proposed model for diagnosis of COVID-19 from CT image in two stages.

2.1. Lung Segmentation

Despite the great efforts to propose a significant image segmentation schema as in SENet [19], BConvLSTM [20], and dense convolutions [21,22], the complicated nature of COVID-19 require further performance improvement on these existing schema related to their accuracy, effectiveness, and efficiency.

In some of U-shaped segmentation networks, the mined feature maps in the skip connection are passed to a handling phase (convolution, attention layers, etc.) and later combined. The major downside of these models is that the handling phase is done independently for the encoder and decoder feature maps, and these maps are followingly combined. Accordingly, we proposed a GR-U-Net that redesign the skip connection using BConv-GRU to capture high resolution information from the feature maps of the encoder

and semantic information from feature maps of decoder. The structural design of GR-U-Net with advanced skip connection is shown in Fig. 3.

2.1.1. Encoding Track

Traditional U-Net architecture involves a shrinking path to capture contextual representation from the input images hierarchically. To gain further performance improvement over U- net, we created our encoder using a pre-trained Efficient-Net based on the notion of transfer learning (TL).

By considering the shortage of the amount of data for training sophisticated models, the complexity of gathering a vast number of labeled instances from images, and the isolated learning nature of DL models (focusing on a specific task) is eliminated. We adopted the TL paradigm to leverage pre-trained ResNet18 experience and exploit it to solve our issue with less amount of data. We build the encoding direction similar to the first four layers of Efficient-Net.

In typical U-Net architecture, at the end of encoding direction, we have a stack of convolutional layers to learn various features. However, such a sequence of layers leads to redundant feature learning. To tackle this problem, we adopt densely linked convolutions by concatenating other layers' feature maps with the current maps, which are followingly passed to the next convolution as an input. Additionally, the dense connection quickly sending the gradients to empowers features propagation and reusability, which in turn enriches the network's representational power. Up to this, at the end of the encoding direction, we introduce a sequence of N densely connected consecutive blocks where each block consists of two successive convolutions. We assume that the output of i_{th} convolution block denoted as $X_e^i \in R^{F_l \times H_l \times W_l}$, and the input i_{th} $(i \in \{1, \cdots, N\})$ of the block is the concatenation of all prior layers maps $[X_e^1, X_e^2, \cdots, X_e^N] \in R^{(i-1)\ F_l \times H_l \times W_l}$.

2.1.2. Decoding Track

In this part, each stage is an up-sampling operation over the preceding layer output. For further improvement of network representation power, we augmented the conventional U-Net with two intelligent modules, namely Squeeze and Excitation (SE) module and BConvGRU. Merged SE modules help networks utilize global patterns to empathize with relevant significant features and ignore less informative ones. These blocks take up-sampling output feature maps and promote them to be more instructive using either spatial weight in case of spatial SE (sSE) and channel weights in case of channel SE (cSE) depending on spatial and channel interdependencies respectively. The output of the sSE module and cSE are then concatenated and passed to the next-sampling function. In the typical U-Net, the feature maps in the encoding track are merged with their corresponding produced up-sampling maps. Both kinds of feature maps are integrated using the Bi-ConvGRU layer that produces combined maps to followingly fed into two convolutional layers, concatenated S-SE and C-SE output, and later convolutional layer. Given previous layer output feature maps as $X_d \in \mathrm{R}^{F_{l+1}+H_{l+1}+W_{l+1}}$ where layer l has F_l of feature maps with the size of $W_l \times H_l$. In the next layer, we got $F_{l+1} = 2 \times F_l$, $W_{l+1} = \frac{1}{2} \times W_l$, and $H_{l+1} = \frac{1}{2} \times H_l$, which means that the decoding track is halving the number of the feature maps and doubling their dimensions in each stage to obtain the original size of the input image at the ending layer.

To empower convolution operation learning performance, we adopt modified spatial squeezing and channel excitation operations [20] after each block to overcome channel dependencies issues. Since each of F filters convolves along the corresponding receptive field, which prevents calculated output U from exploiting relevant information outside of this region as depicted in Fig. 4 (a). Accordingly, we adopted a squeeze operation to calculate channel statistics $Z \in R^c$ using global average pooling by partitioning U along with its spatial scopes $H \times W$. At the same time, the c-th component of z is expressed using equation (1).

$$Z_c = \mathrm{F}_{sc}(\mathrm{U}_c) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{w} \mathrm{U}_c \mathrm{c}\ (\mathrm{i},\ \mathrm{j}) \tag{1}$$

Then, to exploit the combined features in the channel squeezing operation, we apply an excitation operation to detect channels' nonlinear interaction and also capture a non-mutually exclusive association, and to do that we implemented a simple gating operation with a sigmoid activation as formulated in equation (2).

$$s = F_{ex}\ (z,\ W) = \sigma\,(g(z, W)) = \sigma\ (W_2 \& (W_1 z)) \tag{2}$$

where & represents *Relu* activation, $W_1 \in R^{C \times \frac{C}{r}}$, and $W_2 \in R^{C \times \frac{C}{r}}$. And for more generalization, we adopted three FCL layers as a dimensionality-reduction layer where r denotes reduction threshold; we got higher results with $r = 2$. Finally, the output U is computed and rescaled activations s according to equation (3).

$$X_c = F_{scale}\ (u_c,\ s_c) = s_c\ u_c \tag{3}$$

while $X = [x_1,\ x_2 x_c]$ and $F_{scale}\ (u_c,\ s_c)$ represents channel product of feature maps $u^c \in \mathrm{R}^{H \times W}$ with value s_c. Besides, we adopt the channel squeeze and spatial excitation mechanism (sSE) proposed in [19], in which the feature map U squeezed along the convolutional channels and excited spatially, which consequently finetune image segmentation process; the structure of is shown *Fig. 4* (b). The sliced input tensor $U = [u^{1,1}, u^{1,2}, u^{i,j} u^{H,W}] with u^{i,j} \in R^{1 \times 1 \times C}$ denote the spatial coordinate (i, j) where $i \in [1,\ 2 \ldots\ldots\ H]$ and $j \in [1,\ 2 \ldots\ldots\ W]$. The convolution operation $*$ perform spatial squeezing $q = W_{sq} * U$ using weight $W_{sq} \in \mathrm{R}^{1 \times 1 \times C \times 1}$ where the tensor $q \in \mathrm{R}^{H \times W}$. Then Each $q_{i,j}$ denoted linear combination of all cannel C for each spatial point (i, j) that following rescaled into the interval of $[0, 1]$ using *sigmoid* function σ for recalibrating or exciting U spatially $\widehat{U} = [\sigma(q_{i,j}) u^{1,1}, \sigma(q_{i,j}) u^{1,2}, \sigma(q_{i,j}) u^{i,j}, \sigma(q_{i,j}) u^{H,\ W}]$where $\sigma(q_{i,j}) u^{i,j}$ reflects the relative significance of spatial point (i, j). in our model, we utilize both the spatial and channel replication from both modules concurrently via the concatenation of their output as vector $\tilde{X}_d^{up}$.

The output maps of up-sampling, $\tilde{X}_d^{up}$ fed into Batch Normalization (BN) function that generates $\overleftrightarrow{X}_d^{up}$ to overcome the dilemma of variation in the distributions of activations methods slowing the training process as a result of the period spent to acclimate for the next iteration. BN [23] is exploited to improve network stability by standardizing each layer's inputs, which effectively speeds up the training process.

The batch-normalized output $\overleftrightarrow{X}_d^{up} \in \mathrm{R}^{F_l \times H_l \times W_l}$, is now forwarded to a Bi-ConvGRU layer. Meanwhile, the typical LSTM network does not consider the spatial relationship since it primarily utilizes full connections between input and state and also for state-to-state transference. This problem is tackled with ConvLSTM in [20] the convolutional operation exploited to replace the standard full relationship. Inspired by this, we propose ConvGRU to utilize the time efficiency of GRU over LSTM. Accordingly, the spatio- semantic attributes extracted convolutional maps could be successfully acquired by ConvGRU, and its units can be calculated with equations (4-7).

$$Z_t = \sigma\,(W_{xz} * X_t + W_{hz} * h_{t-1}) \tag{4}$$

$$r_t = \sigma\,(W_{xr} * X_t + W_{hr} * h_{t-1}) \tag{5}$$

$$\widehat{h}_t = f(W_{xh} * X_t + r_t \odot (W_{hh} * h_{t-1})) \tag{6}$$

$$h_t = (1 - z_t) \odot \hat{h}_t + z_t * h_{t-1} \tag{7}$$

Where the $\odot$, σ, and $*$ respectively symbolize the Hadamard product, sigmoid operation, and convolution layer. The $\mathrm{Z_t}$ and $\mathrm{r_t}$

and respectively denote the update and reset gates. Having X_t symbolizing input tensor as a combination of tensors X_e and $\overleftrightarrow{X}_d^{up}$, the gated cell initially computes the production values of the update and reset gates according to the (4) and (5), correspondingly. The sigmoid activation guarantee that the output values of z_t and r_t remain within a range from 0 to 1. Next, equation (6) is used to compute the hidden state $\hat{h}_t$ according to the present input value and the hidden state h_{t-1}at previous timestep. In equation (7), the final hidden state h_t is computed by linearly combining the value of hidden state $\hat{h}_t$ at current time step and value of hidden state at previous time step h_{t-1}. The before-mentioned ConvGRU only considers the semantic dependencies in forwarding directions. Nevertheless, the backward dependency information is still unused; thus, learning the semantic dependencies in a bidirectional manner is likely to be more beneficial because of the demonstrated efficacy of combing forward and backward learning in a single network [24]. Therefore, the proposed GR-U-Net introduces an improved skip connection based on Bi-directional-ConvGRU (Bi-ConvGRU) that effectuate two ConvGRU layers to convolve the received convolutional maps from encoders in a bidirectional manner then calculate the new feature maps at the decoding track. Where these generated maps contain Spatio-semantic representations calculated as $Y_t = tanh(W_t^{\vec{H}} * \vec{H}_t + W_t^{\overleftarrow{H}} * \overleftarrow{H}_t)$. In which, $Y \in \mathrm{R}^{F_l \times H_l \times W_l}$, *tanh* indicates the tangent line function, and the of the forward hidden state and the backward hidden state symbolized as $\vec{H}_t$ and $\overleftarrow{H}_t$.

2.2. Feature Extraction Module

In this subsection, the segmented image is processed to capture disease-relevant features. In computer vision, DL models perform learning using one of two strategies, namely, learning from scratch and transfer learning (TL) from pre-trained models. Since learning from scratch strategy requires a large amount of data, we decide it will be ineffective for the underlying problem; in order to tackle this problem, we make use of TL idea by employing two advanced pre-trained architectures. In particular, we adopt the most modern advanced pre-trained architectures.

2.2.1. EfficientNet

According to the notion that imposing stability among all networks, dimensions lead to a significant improvement of accuracy and effectiveness. Tan et al. [10] recently proposed a novel EfficientNet architecture aiming to enhance the CNN architecture performance by performing three-dimensional scaling specifically for width (w), depth (d), and resolution (r). Unlike the traditional process of arbitrary scaling, EfficientNet introduces an effective compound scaling strategy that enables uniform balanced scaling of network dimension and amazingly achieved such balance by just scaling each of them with a constant ratio. Particularly, uniform dimensions scaling can be computed using the compound coefficient, as shown in equation (8).

$$d = \propto^{\varphi}$$

$$w = \beta \propto^{\varphi}$$

$$r = \gamma^{\varphi}$$

$$\propto \geq 1,\ \beta \geq 1,\ \gamma > 1 \tag{8}$$

Where φ denotes the user-quantified coefficient for determining the number of resources available for model scaling, the constants $\propto$, β, γ are that can be calculated from small grid search to determine how network width, depth, and resolution will respectively be allocated such resources.

The architecture of EfficientNet is based on mobile size net layers network with nine stages of convolutional layers with a kernel of size 3 × 3 or 5 × 5 with FCL at the end and called EfficientNet-B0 which have $\propto = 1.2$, $\beta = 1.1$, $\gamma = 1.15$ under constraint $\propto . \beta^2 . \gamma^2 \approx 2$ with fixed $\varphi = 1$, then additional seven architectures from EfficientNet-B1 to EfficientNet-B7 constructed by fixing α, β, γ value while scaling up beforementioned baseline network with a different value of φ.

2.2.2. Multi-Scale Feature Fusion (MsFF) module

It is broadly known that multi-scale features are beneficial in a variety of tasks of natural and medical image analysis tasks. For example, in the segmentation task, the incorporation of multi-scale information has shown amazing performance [25–28]. Motivated by these works, we propose to learn and capture CT features at various scales, aiming to capture and learn both local and global semantic information. Particularly, we introduce multi-scale feature fusion (MsFF) based on the recently proposed multi-scaling mechanism. The multi-scales features are denoted as *Fs*, with s representing the network level (see Fig. 2). Meanwhile, features fusion performed at diverse resolutions corresponding to every level *s*, bilinear interpolation is used to upsample them into the same resolution; hence generate enlarged feature maps *F′s*. Follow, *F′s* from altogether scales are merged creating a tensor that fed into convolution layer to generate a joint multiscale feature map, $F_{MS} = conv([F_0'; F_0'; F_0'; F_0'])$. Thus, F_{MS} capture low-level representation from earlier layers, and high-level contextual information from the later layers. After that, these aggregated feature maps fed into the infection prediction path with AM to calculate the attention features that are used to guide the classification decision as discussed in the next section.

2.2.3. Infection Guidance (GuInf) module

Given lung-segmented CT slices as an input, the channel of infection predictions generates a map that represents the infection regions related to such that they are diagnostically relevant. As presented in Fig. 2, This path of the network is trained with infection GT annotated by VB-Net toolkit[28]. Branching from the feature extraction path at *bp* (bifurcation position) layer, this path comprises an extra convolutional layer with the kernel of 1 × 1. After that, the generated feature map is standardized via *SoftMax* function according to equation (9).

$$y_{i,j} = P\left(cx_{i,j}\right) = \frac{1}{1 + \exp\left(-x_{i,j}\right)} \tag{9}$$

where $x_{i,j}$, $i \in \{1, 2, \cdots, M\}$, $j \in \{1, 2, \cdots, N\}$ symbolizes the pixels of size $M \times N$, and $P(cx_{i,j})$ denotes the probability value that $x_{i,j}$belonging to the ct-h class, while c =1 indicates that the pixel is diagnostically relevant while other pixels have C=0. This path is trained to predict infection ROI by reducing the dice loss [29] for the estimated infection map and the corresponding GT mask as formulated in equation (10).

$$y_{i,j} = \frac{2 \times \left\{\sum_{I=1}^{M} \sum_{J=1}^{N} y_{i,j} \cdot \widehat{y}_{i,j} + \epsilon\right\}}{\left\{\sum_{I=1}^{M} \sum_{J=1}^{N} \left(y_{i,j} + \widehat{y}_{i,j} + \epsilon\right)\right\}} \tag{10}$$

The output map with the probability scores $y_{i,j}$ indicating the importance of the positional information for the diagnosis. This final map exploited to guide the attention of the model for pneumonia classification.

2.2.4. Output layer

In this layer, the output of multi-stream fusion layers concatenated for final disease classification. Where the configuration of the final dense layers includes four FCL layers with (512, 256, 15, 56) separated with 0.5 dropouts, *ReLU* activations, and batch normalization layers. The slices were pooled using two FCL layers to

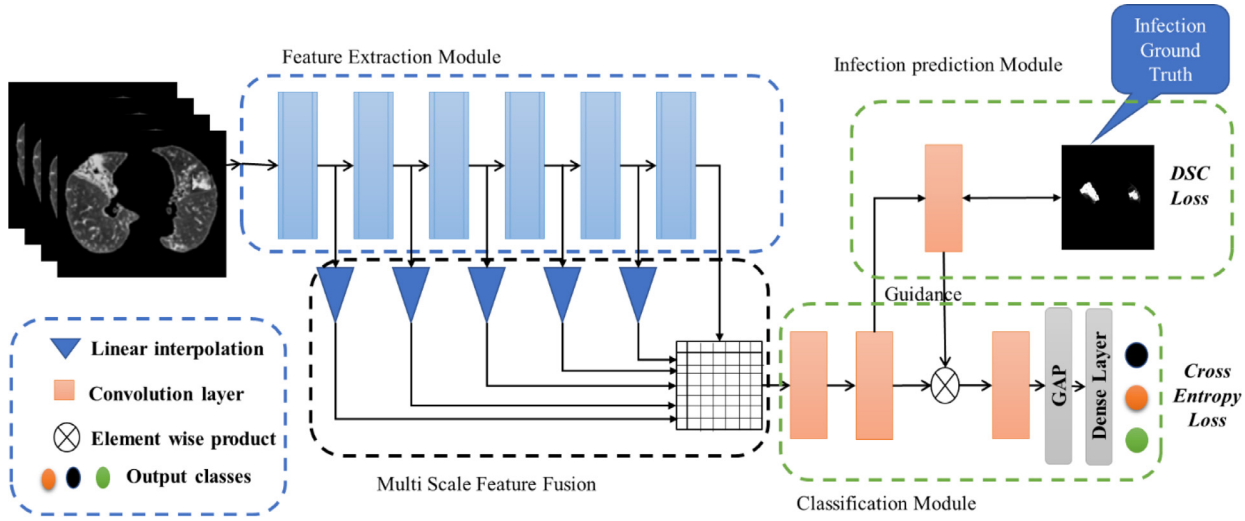

Fig. 2. Architecture of Proposed Classification network.

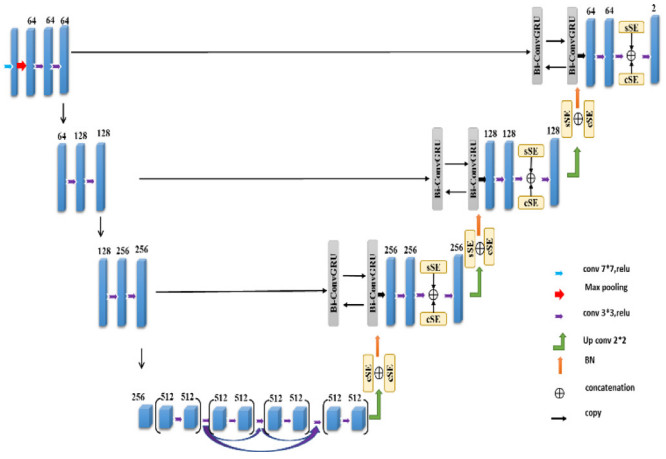

Fig. 3. The construction of GR-U-Net for lung segmentation.

make predictions at the patient level. Then, the SoftMax function is utilized to calculate the probability score concerning each disease class, while the class that gains a higher probability score is considered as the correct disease class as formulated in equation (15-16).

Sooner or later, to training or evaluate our network, we aim to calculate and reduce the loss value. herein, we seek to minimize Cross-Entropy loss (CEL) to calculate classification output according to equation (17).

Where *yi* is the truthful disease label while (y_i) $\tilde{i}$s the model predicted classThe model architecture with the highest performance has been chosen. Furthermore, we conducted a grid-search for various hyper-parameters and got the most top performance with training epochs within range (60-80), batch size with a value of (16,32), and with learning rate in between (0.0001 – 0.001).

3. Experimental Design

3.1. Dataset's Description

To assess the effectiveness of the proposed GR-U-Net, we adopt a lung segmentation dataset publicly available for evaluating the proposed segmentation technique, we adopt a lung segmentation dataset publicly available on The Cancer Imaging Archive (TCIA) Public Access (https://wiki.cancerimagingarchive.net/display/Public/Lung+CT+Segmentation+Challenge+2017). The data was captured from 60 patients with a total of 9,593 images. We split the data into five folds for training purposes and the test set constituted 20% of the data. To analyze the classification performance, the COVID-CT-MD dataset [30] is employed for training and evaluating the proposed model. The dataset consists of chest CT scans of 169 persons confirmed as positive COVID-19 patients (Feb-2020: Apr2020), 60 CAP patients (Apr-2018: Dec-2019), and 76 non-infected individuals (Jan-2019: May-2020). The data was collected Babak Imaging Center, Tehran, Iran. Three main criteria are considered by three radiologists for classifying the cases. The labeled portion of data comprises 18,392 slices with no infection (NIF) and 4,957 slices showing infection.

3.2. Performance metrics

For evaluating the lung segmentation performance, three common performance measures are employed i.e., accuracy, Dice similarity coefficient (DSC), Jaccard index (JI). While in the second stage of the framework, five popular performance measures are considered for evaluating the classification performance. The computation of these metrics can be calculated as follow:

$$DSC = \frac{2.TP}{2.TP + FN + FP} \tag{11}$$

$$JI = \frac{TP}{TP + FP + TN} \tag{12}$$

$$Accuracy = \frac{TP + TN}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \tag{13}$$

$$p = SoftMax(X) = \frac{exp(X)}{\sum_1^c exp(X)} \tag{14}$$

$$\tilde{y} = argmax(p) \tag{15}$$

$$L_{Entopy} = \sum (y_i \log(\tilde{y}_i) + (1 - y_i) \log(1 - \tilde{y}_i)) \tag{16}$$

$$Precision = \frac{TP}{\mathrm{TP} + \mathrm{FP}} \tag{17}$$

$$Recall = \frac{TP}{\mathrm{TP} + \mathrm{FN}} \tag{18}$$

$$F1 - measure = 2 * \frac{Recall \times Precision}{Recall + Precision} \tag{19}$$

Area under the receiver operating characteristic (ROC) curve (AUC), which represents the relation between true positive rate (TPR) and the false positive rate (FPR).

Table 1
Performance comparison of lung dataset.

Methods	Accuracy(%)	DSC(%)	JI(%)	AUC(%)
R2U-net [31]	92.13±13.1	88.98±8.64	94.96±5.41	94.22±8.93
CE-Net [32]	95.08±11.4	91.38±4.25	95.26±2.93	96.34±10.2
CPFNet [33]	95.61±9.18	93.06±6.13	94.21±3.45	95.62±7.55
GR-U-Net	97.93±9.81	93.97±3.21	97.28±2.61	97.88±6.12

Table 2
The results of paired t-test experiments

R2U-net [31]	0.030581	0.030492	0.10078
CE-Net [32]	0.042270	0.032882	0.052341
CPFNet [33]	0.028069	0.042838	0.068629

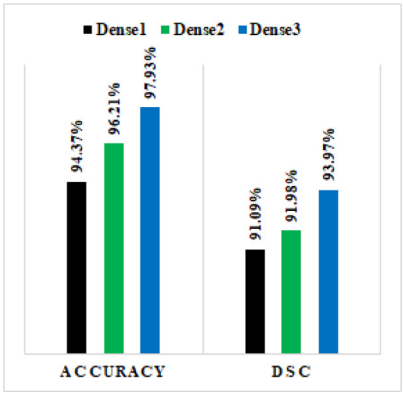

Fig. 4. GR-U-net performance with different number of dense blocks.

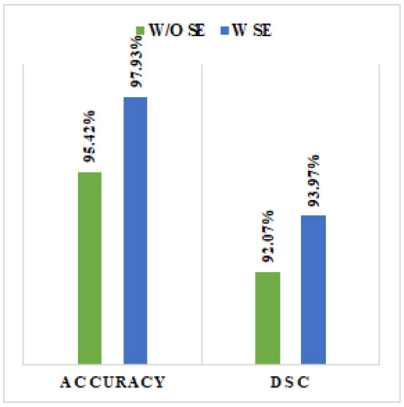

Fig. 5. GR-U-Net performance with and without SE blocks.

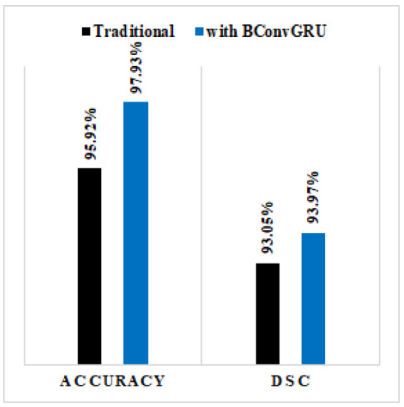

Fig. 6. GR-U-Net performance with and without BConvGRU.

4. Results and discussion

4.1. Segmentation performance

4.1.1. Comparative analysis

In our experiments, for the purpose of demonstrating the proposed segmentation technique, we conduct comparative experiments against other studies R2U-net [31], CE-Net [32], CPFNet [33] and the obtained quantitative results shown in Table 1. It could be observed that the proposed GR-U-Net achieves the highest accuracy of 97.93%, DSC of 93.97%, JI of 97.28, and AUC of 97.28 which outperforms the competing cutting-edge segmentation networks

4.1.2. Statistical Analysis

Beyond and above, the statistical significance of the segmentation outcomes of proposed GR-U-Net compared to the results attained by the competing segmentation networks, a paired t-test experiment is introduced, and the calculated p-values are presented in Table 2. Where the p-value < 0.05 implies that the findings of the proposed model statistically vary from those of the competent methods. It could be seen that all the p-values are less than 0.05 which further validate the efficiency of the proposed GR-U-Net in segmenting the lungs from Chest CT scans.

4.1.3. Ablation Analysis

In Fig. 4. We compare the impact Number of dense blocks on the proposed GR-U-Net return. We could observe that the three dense blocks yield higher performance with 97.93% of accuracy and 93.97% of DSC compared to utilizing one block that gives 94.37% of accuracy and 91.09% of DSC or two blocks that result in 96.21% of accuracy and 91.98% of F1-measure. We do not leverage more than three blocks to avoid enlarging several network parameters. Also, to approve the effectiveness of both SE blocks on the decoding track, we compare our model with and without these blocks in Fig. 5, and it could be observed that enrolling SE blocks within the decoding path achieve performance improvements with 2.51% and 1.9% on accuracy and F1-measure, respectively. Moreover, to demonstrate the impact of adopting BConvGRU as a skip connection compared to the traditional U-net skip connection, we analyze the proposed GR-U-Net performance using both kinds of connections, as shown in Fig. 6. It could be noted that capturing Spatio-semantic characteristics within BConvGRU raises the model's accuracy and F1-measure with 2.01% and 0.92%, respectively.

For extra validation of the efficiency of the proposed, ROC analysis is performed by plotting the ROC of the proposed GR-U-Net as presented in Fig. 7.

4.2. Classification Performance

4.2.1. Comparative analysis

In this experiment, the proposed model is compared against cutting-edge COVID-19 screening approaches namely CAD [14], COVNet [16], AFS-DF [34], and EfficientNet [15], as presented in Table 3. It is notable that show the lowest AFS-DF classification performance (accuracy:87.12%, f1-measure: 87.69%, recall: 92.82%, precision:83.09%. AUC: 92.71%) because of reliance on suboptimal feature engineering. Comparatively, deep learning (COVNet [16], AFS-DF [34], EfficientNet [15]) shows great performance improvement because of their ability to perform robust feature extraction automatically throughout learning. More significantly, the proposed two-stage model overcomes the competent methods with large margins (accuracy: 2.15%, f1-measure: 1.75%, AUC: 1.25%).

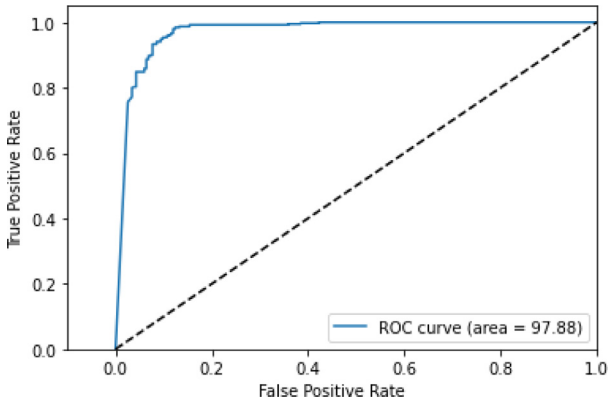

Fig. 7. GR-U-Net ROC analysis.

Table 3
Performance comparison of lung dataset.

Methods	A	F1	R	P	AUC
AFS-DF [34]	87.12%	87.69%	92.82%	83.09%	92.71%
CAD [14]	90.24%	89.58%	93.34%	86.11%	95.33%
COVNet [16]	94.65%	94.31%	96.51%	92.21%	97.12%
EfficientNet [15]	94.32%	94.68%	95.66%	93.72%	97.61%
proposed	96.80%	96.43%	96.50%	96.37%	98.86%

Table 4
the results of paired t-test experiments

AFS-DF [3]	0.01393	0.04792	0.02845
CAD [1]	0.03090	0.04333	0.00150
COVNet [2]	0.01364	0.03397	0.04286
EfficientNet [4]	0.02818	0.01361	0.06293

4.2.2. Statistical Analysis

Beyond and above, the statistical significance of the results of proposed two-stage frameworks compared to those obtained from competing approaches, a paired t-test experiment is introduced, and the calculated p-values are presented in Table 4. It could be seen that all the p-values are less than 0.05 which further validates the efficiency of the proposed model in discriminating COVID-19 from CAP in computer-aided diagnosis.

Table 5
ablation studies for the proposed classification model

	A	F1	AUC
Baseline (B)	90.43%	91.22%	90.14
B(TL)	93.16%	93.22%	95.11
B(TL)+ MsFF	94.82%	94.83%	98.01
B (TL)+ GuInf	95.19%	95.51%	98.36
Proposed	96.80%	96.43%	98.86%

Table 6
confusion matrix of proposed classification model.

		Actual classes			
Predicted classes		COVID-19	CAP	NIF	Recall
	COVID-19	1291	46	11	95.77%
	CAP	41	906	6	95.07%
	NIF	9	13	1612	98.65%
	Precision	96.27%	93.89%	98.96%	
	F1-measure	96.02%	94.47%	98.80%	

4.2.3. Ablation studies

In this section, ablation experiments are performed to evaluate the contribution of different building blocks and the obtained results are given in Table 5. in this experiment, the standard EfficientNet is employed as baseline architecture. It could be noted that applying the TL to the proposed model show great performance improvement. Besides, the inclusion of MsFF improves the performance because it enables the model to learn disease features with different sizes of infection. The inclusion of GuInf module greatly contributes to improving the classification performance as it guides the network to take classification decisions based on infection areas. Combining all the beforementioned modules in the proposed model gives us the best classification performance

4.2.4. Confusion Matrix and Failure analysis

Moreover, to provide detailed analysis, the confusion matrix of the proposed model is presented in Table 6. It could be noted that the proposed model correctly classified 1291 slices out of 1341 (PR: 96.27%) as COVID-19, and correctly classified 906 out of 965 slices (PR:93.89%) as CAP. Thus, most of the misclassification occurs be-

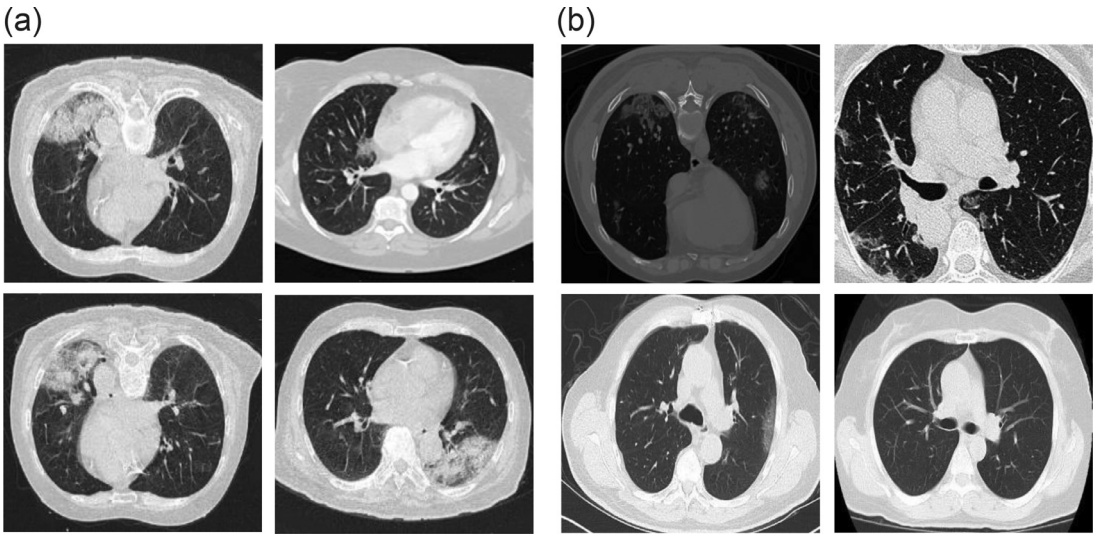

Fig. 8. (a) CAP samples incorrectly classified as COVID-19 with the proposed model (b) COVID-19 samples incorrectly classified as CAP with the proposed model.

tween COVID 19 and CAP; this can be explained as co-infected patients. The misclassification between the COVID-19 and NIF classes can be explained by the early stage of infection where the infection is still slightly present in the received CT slice. Based on misclassification witnessed in confusion matrices, we present eight failure cases in Fig. 8. Particularly, in part (a), we provide four cases of CAP that are classified as COVID-19 infection. While part (b) visualizes some of the incorrectly classified COVID-19 as CAP. This means that the generated attention map from both models erroneously gets activated on various regions irrelevant to pneumonia. This potentially happens due to the lack of volumetric information in our models since the learning performed at slice level.

5. Limitations

Multiple limitations are observed for the proposed framework. Firstly, imaged characteristics of COVID-19 is shown higher similarity with other viral pneumonia. However, owing to the nonexistence of laboratory etiological confirmation of such cases, we could not choose different viral types of pneumonia for classification in this paper. Secondly, we noticed another disadvantage for the deep learning paradigm that is the absence of results' uncertainty; in other words, specifying the value of confidence that the patient belongs to a certain class. Thirdly, there is a considerable amount of overlap in lung presentation to various viral as well as generated lung reactions, which make it impossible to discriminate all lung diseases from chest CT with one approach. Which inevitably requires a multidisciplinary approach. Finally, this paper emphases on detection of COVID-19 and distinguish it from CAP; yet has not considered categorizing the disease severities.

6. Conclusions and Future Work

This study introduces a two-stage DL framework for distinguishing COVID-19 infection from CAP in CT scans of patients to assist doctors and researchers to effectively detect SARS-COV-2 infection from other cases of pneumonia. Firstly, A U-shaped network is introduced in the first stage for lung segmentation where the Bi-convGRU is introduced to capture spatial-semantic features from the encoding track and the output maps of the former block in the decoding track, and also make use of spatial and channel features recalibration in decoding track. Then, we exploit the recent advance in TL techniques by independently employing the EfficientNet-b7 as pre-trained feature extractors while attention modules are introduced to learn multi-scale features necessary for lesion localization purposes.

Up to this, in future work, we intend to accumulate extra CT scans from several centers, demonstrating our model performance, and publish it as a free application. As a subsequent stage, it will be necessary to predict not only the presence of infection but also estimate the degree of severity to enable continuous monitoring of patients during the treatment period. Also, we aim to investigate the hierarchical characteristics of CT images along with other aspects such as RT-PCR, epidemiological, and clinical symptoms for faster and improved diagnosis.

Declaration of Competing Interest

The authors declare that there is no conflict of interest regarding the publication of the paper.

Funding

This research has no funding source.

Ethical approval

This article does not contain any studies with human participants or animals performed by any of the authors

References

[1] C.A. Devaux, J.M. Rolain, P. Colson, D. Raoult, New insights on the antiviral effects of chloroquine against coronavirus: what to expect for COVID-19? Int. J. Antimicrob. Agents (2020), doi:10.1016/j.ijantimicag.2020.105938.
[2] F. Song, et al., Emerging 2019 novel coronavirus (2019-NCoV) pneumonia, Radiology (2020), doi:10.1148/radiol.2020200274.
[3] W. Choi, et al., Performance of radiologists in differentiating COVID-19 from viral pneumonia on chest CT, Radiology (2020).
[4] A. Kumar, M. Fulham, D. Feng, J. Kim, Co-Learning Feature Fusion Maps from PET-CT Images of Lung Cancer, IEEE Trans. Med. Imaging (2020), doi:10.1109/TMI.2019.2923601.
[5] H. Wang, H. Jia, L. Lu, Y. Xia, Thorax-Net: An Attention Regularized Deep Neural Network for Classification of Thoracic Diseases on Chest Radiography, IEEE J. Biomed. Heal. Informatics (2020), doi:10.1109/JBHI.2019.2928369.
[6] T. Ai, et al., Correlation of Chest CT and RT-PCR Testing for Coronavirus Disease 2019 (COVID-19) in China: A Report of 1014 Cases, Radiology (2020), doi:10.1148/radiol.2020200642.
[7] F. Shi, et al., Review of Artificial Intelligence Techniques in Imaging Data Acquisition, Segmentation, and Diagnosis for COVID-19, IEEE Reviews in Biomedical Engineering (2021), doi:10.1109/RBME.2020.2987975.
[8] A. Narin, C. Kaya, Z. Pamuk, Automatic detection of coronavirus disease (COVID-19) using X-ray images and deep convolutional neural networks, Pattern Anal. Appl. (2021), doi:10.1007/s10044-021-00984-y.
[9] Y. Jin, D. Han, H. Ko, TrSeg: transformer for semantic segmentation, Pattern Recognit. Lett. (2021), doi:10.1016/j.patrec.2021.04.024.
[10] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," 2019.
[11] R.R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization, Int. J. Comput. Vis. (2020), doi:10.1007/s11263-019-01228-7.
[12] H. Kang, et al., Diagnosis of Coronavirus Disease 2019 (COVID-19) with Structured Latent Multi-View Representation Learning, IEEE Trans. Med. Imaging (2020), doi:10.1109/TMI.2020.2992546.
[13] X. Ouyang, et al., Dual-Sampling Attention Network for Diagnosis of COVID-19 from Community Acquired Pneumonia, IEEE Trans. Med. Imaging (2020), doi:10.1109/TMI.2020.2995508.
[14] Z. Wang, et al., Automatically discriminating and localizing COVID-19 from community-acquired pneumonia on chest X-rays, Pattern Recognit (2021), doi:10.1016/j.patcog.2020.107613.
[15] G. Marques, D. Agarwal, I. de la, Torre Díez, Automated medical diagnosis of COVID-19 through EfficientNet convolutional neural network, Appl. Soft Comput. J. (2020), doi:10.1016/j.asoc.2020.106691.
[16] L. Li, et al., Using Artificial Intelligence to Detect COVID-19 and Community-acquired Pneumonia Based on Pulmonary CT: Evaluation of the Diagnostic Accuracy, Radiology (2020), doi:10.1148/radiol.2020200905.
[17] R.M. Wehbe, et al., DeepCOVID-XR: An artificial intelligence algorithm to detect COVID-19 on chest radiographs trained and tested on a large U.S. Clinical data set, Radiology (2021), doi:10.1148/RADIOL.2020203511.
[18] Z. Xiong, et al., Artificial Intelligence Augmentation of Radiologist Performance in Distinguishing COVID-19 from Pneumonia of Other Origin at Chest CT, Radiology (2020), doi:10.1148/radiol.2020201491.
[19] J. Hu, L. Shen, S. Albanie, G. Sun, E. Wu, Squeeze-and-Excitation Networks, IEEE Trans. Pattern Anal. Mach. Intell. (2020), doi:10.1109/TPAMI.2019.2913372.
[20] M. Asadi-Aghbolaghi, R. Azad, M. Fathy, and S. Escalera, "Multi-level Context Gating of Embedded Collective Knowledge for Medical Image Segmentation," Mar. 2020, [Online]. Available: http://arxiv.org/abs/2003.05056.
[21] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," 2017, doi: 10.1109/CVPR.2017.243.
[22] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 2016, doi: 10.1109/CVPR.2016.90.
[23] S. Ioffe, Christian Szegedy, Batch Normalization: Accelerating Deep Network Training by Reducing, J. Mol. Struct. (2015).
[24] L. Tian, X. Li, Y. Ye, P. Xie, Y. Li, A Generative Adversarial Gated Recurrent Unit Model for Precipitation Nowcasting, IEEE Geosci. Remote Sens. Lett. (2020), doi:10.1109/LGRS.2019.2926776.
[25] H. Yang, J.Y. Kim, H. Kim, S.P. Adhikari, Guided Soft Attention Network for Classification of Breast Cancer Histopathology Images, IEEE Trans. Med. Imaging (2020), doi:10.1109/TMI.2019.2948026.
[26] S. Jetley, N. A. Lord, N. Lee, and P. H. S. Torr, "Learn to pay attention," 2018.
[27] A. Sinha, J. Dolz, Multi-Scale Self-Guided Attention for Medical Image Segmentation, IEEE J. Biomed. Heal. Informatics (2021), doi:10.1109/JBHI.2020.2986926.
[28] F. Shan, et al., Abnormal lung quantification in chest CT images of COVID-19 patients with deep learning and its application to severity prediction, Med. Phys. 48 (4) (Apr. 2021) 1633–1645, doi:10.1002/mp.14609.
[29] C. H. Sudre, W. Li, T. Vercauteren, S. Ourselin, and M. Jorge Cardoso, "Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations," 2017, doi: 10.1007/978-3-319-67558-9_28.

[30] P. Afshar, et al., COVID-CT-MD, COVID-19 computed tomography scan dataset applicable in machine learning and deep learning, Sci. Data (2021), doi:10.1038/s41597-021-00900-3.
[31] M.Z. Alom, C. Yakopcic, M. Hasan, T.M. Taha, V.K. Asari, Recurrent residual U-Net for medical image segmentation, J. Med. Imaging (2019), doi:10.1117/1.jmi.6.1.014006.
[32] Z. Gu, et al., CE-Net: Context Encoder Network for 2D Medical Image Segmentation, IEEE Trans. Med. Imaging (2019), doi:10.1109/TMI.2019.2903562.
[33] S. Feng, et al., Cpfnet: Context pyramid fusion network for medical image segmentation, IEEE Trans. Med. Imaging (2020), doi:10.1109/TMI.2020.2983721.
[34] L. Sun, et al., Adaptive Feature Selection Guided Deep Forest for COVID-19 Classification with Chest CT, IEEE J. Biomed. Heal. Informatics (2020), doi:10.1109/JBHI.2020.3019505.