JCS: An Explainable COVID-19 Diagnosis System by Joint Classification and Segmentation

Yu-Huan Wu, Shang-Hua Gao, Jie Mei, Jun Xu, *Member, IEEE*, Deng-Ping Fan, *Member, IEEE*,

Rong-Guo Zhang, and Ming-Ming Cheng, *Senior Member, IEEE*

Abstract—Recently, the coronavirus disease 2019 (COVID-19) has caused a pandemic disease in over 200 countries, influencing billions of humans. To control the infection, identifying and separating the infected people is the most crucial step. The main diagnostic tool is the Reverse Transcription Polymerase Chain Reaction (RT-PCR) test. Still, the sensitivity of the RT-PCR test is not high enough to effectively prevent the pandemic. The chest CT scan test provides a valuable complementary tool to the RT-PCR test, and it can identify the patients in the early-stage with high sensitivity. However, the chest CT scan test is usually time-consuming, requiring about 21.5 minutes per case. This paper develops a novel Joint Classification and Segmentation (*JCS*) system to perform real-time and explainable COVID-19 chest CT diagnosis. To train our *JCS* system, we construct a large scale COVID-19 Classification and Segmentation (*COVID-CS*) dataset, with 144,167 chest CT images of 400 COVID-19 patients and 350 uninfected cases. 3,855 chest CT images of 200 patients are annotated with fine-grained pixel-level labels of opacifications, which are increased attenuation of the lung parenchyma. We also have annotated lesion counts, opacification areas, and locations and thus benefit various diagnosis aspects. Extensive experiments demonstrate that the proposed *JCS* diagnosis system is very efficient for COVID-19 classification and segmentation. It obtains an average sensitivity of 95.0% and a specificity of 93.0% on the classification test set, and 78.5% Dice score on the segmentation test set of our *COVID-CS* dataset. The COVID-CS dataset and code are available at https://github.com/yuhuan-wu/JCS.

Index Terms—COVID-19, joint diagnosis, CT classification, CT segmentation, COVID-19 dataset.

I. INTRODUCTION

CORONAVIRUS disease 2019, or COVID-19, is an epidemic disease caused by the Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2). It outbreaks around the world in a short period and has caused 1,914,916 confirmed cases and 123,010 confirmed deaths as of April 15th, 2020. COVID-19 pushes the health systems of over 200 countries to the brink of collapse due to the lack of medical supplies and staff and thus has been declared as a pandemic by the World Health Organization [1]. The current main diagnostic tool for COVID-19 is via the Reverse Transcription Polymerase Chain Reaction (RT-PCR) test [2]. However, the sensitivity of the RT-PCR test is not high enough to effectively prevent the pandemic [3], [4]. So the false-negative cases of RT-PCR tests are a potential threat to public wellness, and missing any COVID-19 cases will probably cause secondary infections of large areas.

To hinder the terrific infection of COVID-19, medical radiology imaging is employed as a complementary tool for the RT-PCR test [5]. This is based on the fact that the clinical signs of chest X-rays for most COVID-19 patients indicate lung infection [6]. The works of [3], [4] show that CT scan tests are with high sensitivity. Besides, a CT scan test is necessary for monitoring the severity of the illness [7]. However, the diagnosis duration is the major limitation of CT scan tests: even experienced radiologists need about 21.5 minutes [8] to analyze the test results of each case. The experienced radiologists are severely lack during the pandemic outbreak, posting difficulties identifying potentially infected patients in time. Thus, automatic diagnosis systems are highly desired.

Thanks to the powerful discriminative ability of deep convolutional neural networks (CNNs), artificial intelligence (AI) technologies are reforming the medical imaging community. Deep CNNs are usually trained on large scale datasets to demonstrate their capability. However, most of the existing CT scan datasets for COVID-19 [9]–[12] could not meet this demand, as they contain at most hundreds of CT images from tens of cases. Besides, most of the current COVID-19 datasets only provide the patient-level labels (*i.e.*, class labels) indicating whether the person is infected and lacks fine-grained pixel-level annotations. Thus, CNN models trained with these datasets usually neglect the valuable information for explaining the final predictions. Despite several CT scan diagnosis systems [4], [13]–[17] have been established for testing the suspected COVID-19 cases, most of them suffer from two drawbacks: 1) they are trained on small scale datasets and thus not robust enough for versatile COVID-19 infections; 2) they perform classification based on the black box CNNs while lacking the explainable transparency to assist the doctors during the medical diagnosis.
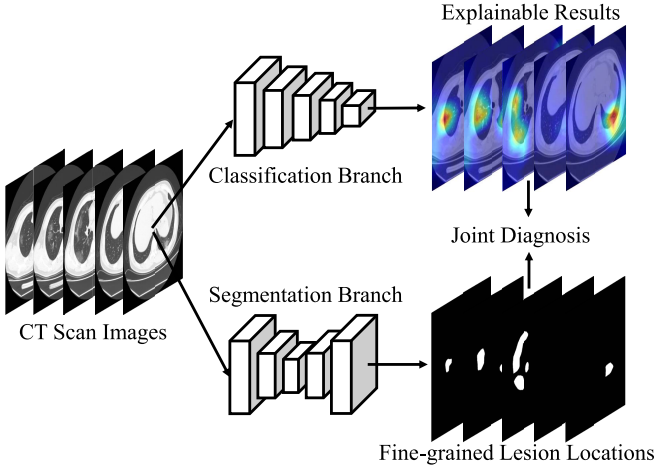
Fig. 1. Illustration of our JCS diagnosis system for COVID-19. Our JCS system will perform the segmentation diagnosis only if the classification branch reports positive COVID-19 predictions.

To alleviate the drawbacks mentioned above, in this work, we 1) construct a large scale *COVID-CS* dataset with both patient-level and pixel-level annotations and 2) propose a *J*oint *C*lassification and *S*egmentation (*JCS*) based diagnosis system to provide explainable diagnosis results for medical staffs fighting with COVID-19. Specifically, we utilize the collected *COVID-CS* dataset that contains thousands of CT images from hundreds of COVID-19 cases to train our *JCS* system for better diagnosis performance. As illustrated in Figure 1, our *JCS* diagnosis system first identifies the suspected COVID-19 patients by a classification branch and provides diagnosis explanations via activation mapping techniques [20]. Our system is then feasible to discover the locations and areas of the COVID-19 infection in lung radiography via fine-grained image segmentation techniques. With the explainable classification results and corresponding fine-grained lesion segmentation, our *JCS* system largely simplifies and accelerates the diagnosis process for radiologists or other medical experts. As shown in Table II, our *JCS* system needs only 22.0 seconds for each infected case or 1 second for each uninfected case, much faster than the RT-PCR tests and CT scan analysis by experienced radiologists. With the assistance of our *JCS* system, experienced radiologists only cost 54.4 (32.4 for radiologists and 22.0 for *JCS*) seconds for each infected case or 1.0 second for each uninfected case, keeping the same high specificity and sensitivity. Hence, the speed and effectiveness of assistance have shown the superiority of our *JCS* system.

In summary, our contributions are mainly three-fold:

- **We construct a new large scale COVID-19 dataset**, called *COVID-CS*, which contains 3,855 fine-grained pixel-level labeled CT images from 200 COVID-19 patients, 64,771 patient-level annotated CT images from 200 other COVID-19 patients, and 75,541 CT images of 350 uninfected cases.
- **We develop a novel COVID-19 diagnosis system** to perform explainable Joint Classification and accurate

TABLE I
SUMMARY OF DIFFERENT DATASETS (UPDATED ON 2020/4/10)

Dataset	Date	Link	Type	#Images	#Cases
PLXR [11]	2020/03/23	Link	X-rays	98	70
8023Dataset [9]	2020/03/25	Link	X-rays	229*	-
CTSeg [12]	2020/03/28	Link	CT	110	60
COVID-CT [10]	2020/03/30	Link	CT	746*	-
COVID-CS (**Ours**)	2020/04/12	-	CT	>**144K**†	**750**

*: The number is reported from the authors' GitHub repository.
†: Among our dataset, 3,855 images of 200 positive cases are pixel-level annotated, 64,771 images of the other 200 positive cases are patient-level annotated, and the rest 75,541 images are from the 350 negative cases.

TABLE II
AVERAGE TIME OF COVID-19 DIAGNOSIS BY DIFFERENT METHODS. "CT R." INDICATES CT RADIOLOGIST AND "CT R. + *JCS*" IS CT RADIOLOGIST DIAGNOSIS WITH THE ASSISTANCE OF *JCS*

Method	RT-PCR	CT R.	CT R. + *JCS*	*JCS*
Time	∼4h [21]	21.5min [8]	1s†/54.4s	1s†/22.0s

†: diagnose uninfected cases.

lesion Segmentation (*JCS*), showing clear superiority over previous systems.

- On our *COVID-CS* dataset, **our *JCS* system achieves 95.0% sensitivity and 93.0% specificity on COVID-19 classification, and 78.5% Dice score on segmentation**, surpassing previous state-of-the-art segmentation methods.

The remaining paper is organized as follows. In §II, we briefly summarize the related works. In §III, we introduce the developed diagnosis system for recognizing and analyzing the COVID-19 cases. In §IV, we present our *COVID-CS* dataset with our labeling procedures in detail. Extensive experiments are conducted in §V to evaluate the performance of our system on COVID-19 recognition, with in-depth analysis. §VI concludes this work.

II. RELATED WORKS

A. Existing Accessible COVID-19 Datasets

As of April 15th, 1,914,916 people are infected by COVID-19. But their CT scans are usually private and could not be publicly accessed. To facilitate the development of diagnostic systems, several COVID-19 related datasets are publicly released by researchers around the world. A summary of these datasets is provided in Table I.

One X-ray dataset from Cohen *et al.* [9] contains overall 122 frontal view X-rays, including 100 images of COVID-19 cases, 11 SARS images, and 11 other pneumonia images. The COVID-CT dataset from [10] has 746 CT scan images, 349 images from COVID-19 patients and 397 from non-COVID-19 cases. All the images in these datasets are collected from public websites and/or COVID-19 related papers on medRxiv, bioRxiv, and journals, *etc*. CTs containing COVID-19 abnormalities are selected by reading the figure captions in the papers. Some other resources of the COVID-19 dataset are PLXR [11] and CTSeg [12], which

contains 98 and 110 CT scan images cases, respectively. These datasets are on a small scale and lack diversity since they often contain less than hundreds of images from tens of cases. To fully exploit the power of deep CNNs, it is essential to construct a large scale dataset to train deep CNNs in accurate and robust COVID-19 systems.

B. Manual COVID-19 Diagnosis

The most crucial step of preventing the spread of the COVID-19 is immediately identifying every patient from normal people. Missing any patient will probably cause secondary COVID-19 infections in large areas. Currently, the main manual diagnostic tool is the RT-PCR test [22]. However, the sensitivity of the RT-PCR test is not high enough to effectively prevent the pandemic [3], [4]. As widely available in many hospitals, the CT scan provides a valuable complementary tool to the RT-PCR test. However, some special cases with the RT-PCR test confirmed positive have normal CTs [15], [23], [24]. Such a combination allows maximally to identify potentially infected people, as it can identify COVID-19 patients in the early-stage with high sensitivity [3], [4], [25]. The CT scan is also necessary for monitoring the severity of the illness [7]. During the pandemic outbreak, experienced medical staff is severely lacking, posting difficulties identifying potentially infected patients in time. Thus, automatic diagnosis systems are highly desired.

C. Automatic COVID-19 Diagnosis Systems

Most current medical imaging systems are developed for common diseases that exist for many years, *e.g.*, tuberculosis [26]. These developed systems can be directly modified to attenuate the COVID-19 outbreak. The doctors find that the chest X-rays of COVID-19 patients exhibiting certain abnormalities in the radiography. Based on ResNet-50 [27], COVID-ResNet [28] is proposed to differentiate three types of COVID-19 infections from normal pneumonia individuals. On 1531 chest X-ray images, Zhang *et al.* proposed a deep anomaly detection system for COVID-19 screening, achieving 96.0% sensitivity and 70.65% specificity. Yang *et al.* [29] proposed a system to evaluate the images of 102 volunteers, with a sensitivity of 83.3% and specificity of 94.0%. The system developed by Li *et al.* [30] identifies 78 COVID-19 patients with a sensitivity of 82.6% and a specificity of 100.0% by using the axial and coronal-view of lung CT severity index (CTSI). Chung *et al.* [14] confirmed via collected from 21 patients that visual inspection helps identify the COVID-19 cases and predict the severity via the overall lung total severity score (LTSS). Bernheim *et al.* [15] analyzed the 121 COVID-19 patients and carried on a visual check by the experienced radiologist to categorize them as early, intermediate and late cases. Wang *et al.* [16] found that the COVID-19 disease will be severe during 6-11 days from the infection, based on a study on 366 CT scans of 90 patients. Shi *et al.* [17] developed an imaging-assisted diagnosis procedure to detect the COVID-19 caused pneumonia. Fang *et al.* [4] examined 81 patients by a procedure based on the CTSI and obtained a sensitivity of 98.0%, in contrast

to the sensitivity of 71.0% by RT-PCR. Zhou *et al.* [31] implemented the examination using the non-contrast CTSI of 62 COVID-19 patients, confirming that the CT-assisted evaluation shows better detection accuracy in the progressive stage confirmed to the early-stage. Despite their success on a small set of samples, these COVID-19 diagnosis systems have not been tested by large scale samples. They could not provide useful diagnostic evidence during their diagnostic inference. More works can refer to the reviews of [18], [19], [32].

As far as we know, only two works extract infected regions via pixel-level segmentation. Rajinikanth *et al.* [33] performed the segmentation via the watershed transform techniques [34] with coarse results and limited accuracy. Zhou *et al.* [35] developed a U-Net with an attention mechanism and obtained a Dice score of 69.1% on CTSeg [12] dataset, but its training and test split have only 88 and 22 images. In this work, we propose a diagnosis system by integrating learning-based classification and segmentation networks to provide explainable diagnostic evidence for doctors and improve the user-interactive aspects of the diagnosis process.

D. Deep Classification and Segmentation Methods

Ever since the release of the ImageNet dataset [36], deep convolutional neural networks (CNNs) have become the workhorse for image classification tasks with improving performance. Representative deep classifiers, *e.g.*, AlexNet [37], VGGNet [38], ResNet [27], DenseNet [39], and Res2Net [40], have been widely employed as the feature extractors for other computer vision tasks, such as image segmentation [41]–[43], salient object detection [44], face recognition [45], aerial images analysis [46], style transfer [47], feature matching [48], crowd counting [49], and image restoration [50], *etc*. Despite impressive representation ability of these classifiers, the classification process does not explain clearly the predicted results.

Image segmentation tackles the problem of pixel-level predictions. Semantic segmentation aims to classify the semantic label for each pixel on a natural image [51]. Representative works in this area include FCN [52] and DeepLab [53]. Instance segmentation focuses on discriminating each semantic instance with a unique instance label and pixel-level mask in the image [54]–[56]. Panoptic segmentation [57] integrates semantic segmentation and instance segmentation, and it does semantic segmentation on non-objects (sky, water, grass, etc.) and instance segmentation on objects (cat, dog, bus, etc.). U-Net [58] is a widely employed network for medical image segmentation analysis. It is further extended to 3D U-Net [59], TernausNet [60], and U-Net++ [61] with promising performance on versatile image segmentation tasks. In this work, we develop a novel COVID-19 diagnosis system by integrating deep image classification and segmentation techniques.

III. OUR COVID-19 DIAGNOSIS SYSTEM

The opacification is the basic CT feature of COVID-19 patients [62], and it is defined as the increased attenuation of the lung parenchyma [63]. Our *JCS* system consists of an explainable classification branch to identify the COVID-19

opacifications and a segmentation branch to discover the opacification areas. The classifier is trained on many images with low-cost patient-level annotations and some images with pixel-level annotations for better activation mapping. And the segmentation branch is trained with accurately annotated CT images, performing fine-grained lesion segmentation. By integrating the two models, our *JCS* system provides informative diagnosis results for COVID-19.

A. Explainable Classification

Owing to the strong representation ability of CNNs, the COVID-19 infections can be predicted through only patient-level supervised training. To this end, we build a classification branch that consists of the proposed classification model to endow our *JCS* diagnosis system with the capability of discriminating the COVID-19 patients.

1) Diagnosing COVID-19 via Classification: Predicting whether the suspected patient is COVID-19 positive or not is a binary classification task based on his/her CT scan images. Since designing the novel classification model is not our focus, we build our classifier based on the Res2Net network [40]. As a powerful network, Res2Net has a stronger multi-scale representation ability than ResNet [40]. The last layer is modified as a fully-connected layer with two channels to output the probability of COVID-19 infection or not. If the probability of the infected channel is larger than that of the uninfected one, the patient is diagnosed as COVID-19 positive, or vice versa. For each patient, the CT images are sent to the classification model one by one. If the number of infected CT images is above a threshold, the patient is diagnosed as COVID-19 positive.

2) Explanation by Activation Mapping: As the diagnosis process of CNN classification is in a black box, we employ the activation mapping [20] to increase the explainable transparency of our COVID-19 diagnosis system on its predictions. The last convolutional layer of the classification network is followed by a global average pooling (GAP) layer and a fully-connected layer. Through the GAP layer, our classification model down-samples the feature size from (H, W) to $(1, 1)$, and thus lost the spatial representation ability. Through activation mapping [20], our system finds the response region of the prediction result. The hypothesis is that the gradient of regions in features before the GAP layer is consistent with the prediction evidence. The feature map before the GAP layer contains both high-level semantic and location information. Each channel corresponds to the activation of different semantic cues. The activation mapping is obtained through the gradients of the predicted probability of the feature map. Specifically, given the prediction of COVID-19 branch y_p and the feature map X before GAP, the weight for the k-th channel of X is calculated as:

$$w_k = \frac{1}{HW} \sum_{i=1}^{H} \sum_{j=1}^{W} \frac{\partial y_p}{\partial X_{i,j}^k}, \tag{1}$$

where $X_{i,j}^k$ is the value at position (i, j) in the k-th channel of feature map X. Larger gradients in Eqn. (1) produce a larger

weight of the activation mapping for a certain channel. The activation mapping for a COVID-19 case is computed as:

$$AM_p = \sum_k ReLU(w_k X^k). \tag{2}$$

As shown in Fig. 9, the activation mapping accurately locates the opacification areas of COVID-19 patients, providing explainable results for the prediction of our *JCS* system.

3) Alleviating Data Bias by Image Mixing: By utilizing our explainable classification model, our system can be trained only with patient-level annotation. However, since CT images are from multiple sources, the classifier may be trained to overfit unwanted areas (e.g., the area outside the lesion), as observed via the activation mapping. Therefore, we propose to utilize the image *mixing* technique [64] and help the classifier focus on the lesion areas of COVID-19 cases. The CT images from different sources and the corresponding patient-level annotations are mixed during training. Specifically, for two randomly sampled CT images x_i and x_j ($i \neq j$) and corresponding labels $\hat{y}_i$ and $\hat{y}_j$, the newly mixed sample and the corresponding label are written as:

$$\begin{aligned} x_{ij}^m &= \lambda x_i + (1 - \lambda)x_j, \\ \hat{y}_{ij}^m &= \lambda \hat{y}_i + (1 - \lambda)\hat{y}_j, \end{aligned} \tag{3}$$

where $\lambda \in [0, 1]$ is a random number generated in Beta distribution, *i.e.*, $\lambda \sim \text{Beta}(\alpha, \alpha)$. With mixed samples, our classification model is trained to focus more on the decisive lesion areas of COVID-19 cases, rather than the bias in the data source. Also, the mixing process weakens the confidence of labels, and thus alleviating our system from overfitting.

4) Pixel-Level Supervision for Activation Mapping: Traditional classification models only utilize image labels for training. The activation mapping of them may be inaccurate as these models automatically learn the differences of images of different classes. In our proposed dataset, there are thousands of images with pixel-level annotations for the specific opacification areas, and they can be the direct supervision of the activation mapping. Motivated by the above observations and the work of [65], during the training network, we apply a segmentation loss L_{seg} for the activation mapping of the COVID-19 class channel:

$$L_{seg} = \frac{1}{HW} \| AM_{p,c}^{norm} - S \|_2, \tag{4}$$

where $AM_{p,c}^{norm}$ is the activation mapping of the COVID-19 class channel normalized to $(0, 1)$, S is the binary ground truth pixel-level annotation map, $\| \cdot \|_2$ denotes the ℓ_2 norms. L_{seg} will not be computed if images have no ground truth pixel-level annotations. After applying the segmentation loss L_{seg}, Fig. 9 shows that the activation mapping significantly improves in locating opacifications.

B. Accurate Segmentation

Our segmentation branch aims at discovering the exact lesion areas from the CT images of COVID-19 patients. Fig. 2 shows the architecture of our segmentation branch with or without the combination of the segmentation and classification models. The details of such a combination are illustrated in Fig. 4.
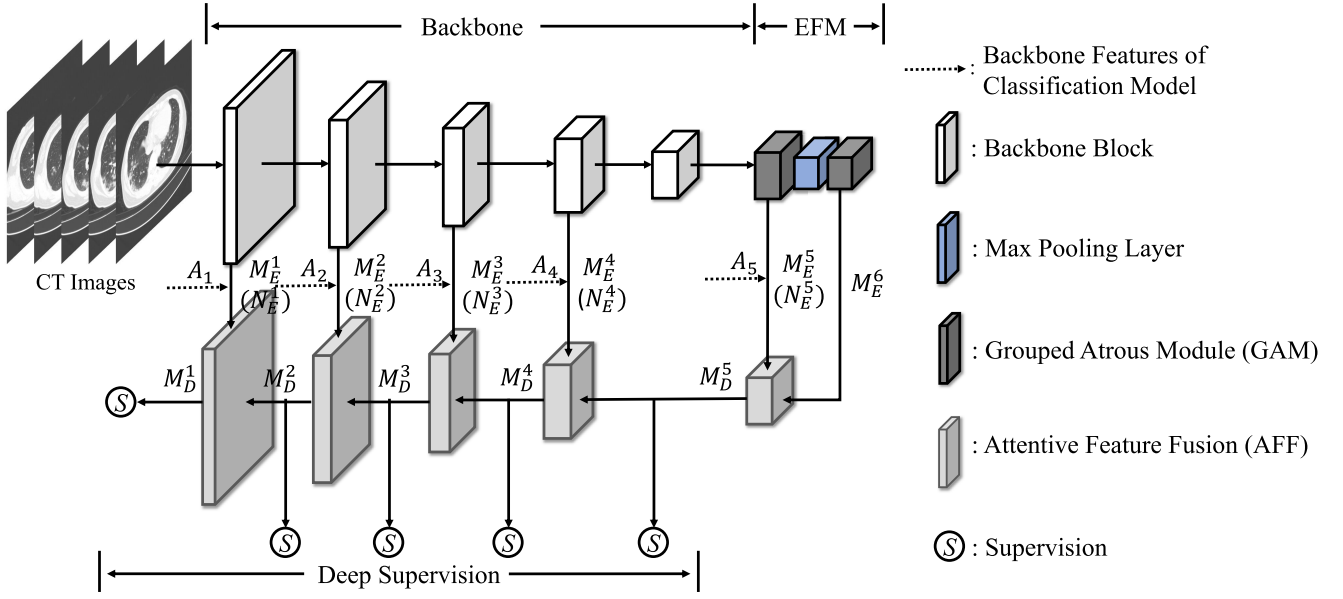
Fig. 2. The architecture of our segmentation branch. EFM indicates the Enhanced Feature Module (§III-B.2). AFF refers to the Attentive Feature Fusion strategy (§III-B.3). If not combined with the classification model, $M_E^1 \sim M_E^5$ will be fed into the decoder; otherwise, the combined $N_E^1 \sim N_E^5$ will be fed into the decoder (Fig. 4, §III-B.4). We apply deep supervision to train our segmentation branch (§III-B.5).

1) Encoder-Decoder Architecture: Our segmentation model consists of an encoder and a decoder.

Encoder. The encoder is based on the VGG-16 [38] backbone, without the last two fully-connected layers. It has five VGG blocks defined as $\{E_1, E_2, E_3, E_4, E_5\}$, respectively. The VGG-16 backbone is first fed with the CT images and produces multi-scale feature maps from the last layers of the five VGG blocks. To downsize the input feature map by half, the front of each block (except the first one) is a *max pooling* function with a stride of 2. The feature map produced by the block E_1 contains the finest features with the highest resolution, while the feature map by the block E_5 is coarsest with the lowest resolution. To achieve better performance, we propose an Enhanced Feature Module (EFM, which will be introduced in §III-B.2) for our encoder to improve its representational power. The EFM module is added after the last layer $conv5\_3$ in the block E_5. It consists of two Grouped Atrous Modules (GAM) to extract stronger feature maps with larger receptive fields. The GAM module generates an extra smaller feature map, half size compared to the coarsest feature map of the VGG-16 backbone. It also enhances the representational power of the feature map produced by the block E_5. Hence, our encoder produces six levels of feature maps $\{M_E^1, M_E^2, M_E^3, M_E^4, M_E^5, M_E^6\}$, with strides of $\{1, 2, 4, 8, 16, 32\}$, respectively. As we employ a U-shape encoder-decoder architecture [58], all these six feature maps will be used in the decoder, as will be introduced later.

Decoder. Our decoder has five side-outputs with 5 different sizes. Here, we do not predict the side-output from the coarsest feature map with a stride of 32, and thus no side-output matches the size of the coarsest feature map M_E^6. In our decoder, we propose an Attentive Feature Fusion (AFF, which will be introduced in §III-B.3) strategy to aggregate the feature maps from different stages and predict the side-output
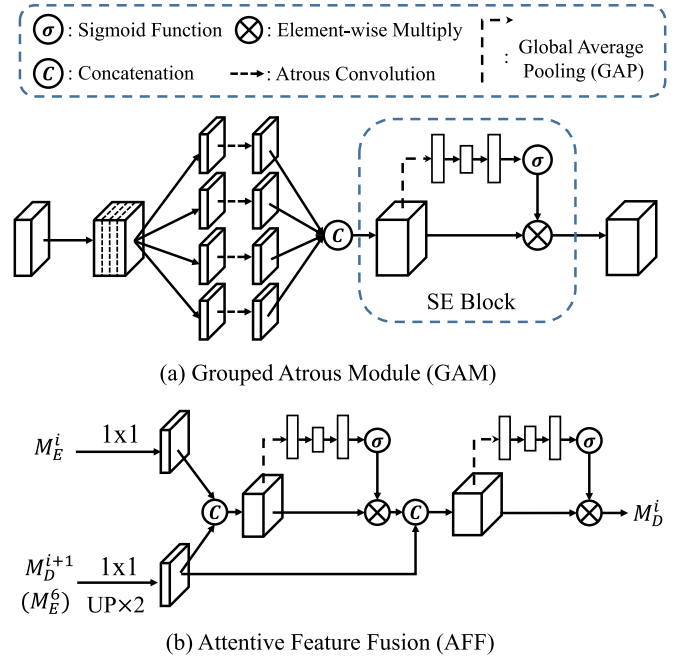

(a) Grouped Atrous Module (GAM)

(b) Attentive Feature Fusion (AFF)

Fig. 3. Proposed (a) GAM and (b) AFF for the segmentation network. In AFF, M_D^{i+1} will be replaced with M_E^6 if $i = 5$. Cubes represent three-dimensional feature maps, while rectangles mean feature vectors.

of each stage. Our AFF emphasizes the significance of the top-level feature map and utilizes the attention mechanism to filter useful features from the bottom feature map. The last output with the same resolution of the CT image input will be chosen as the final prediction.

2) Enhanced Feature Module: The proposed EFM module is added after the last layer of E_5 in the VGG-16 encoder. It consists of two sequential GAM modules and a *max pooling* function between them. As shown in Fig. 3 (a),

the first layer of the GAM module is a 1×1 convolution layer to expand the channels of the feature map. Then the feature map is equally divided into 4 groups. Unlike the trivial group convolution, we deploy atrous convolution [53] with different atrous rates to the 4 groups to derive a more abundant feature map with various receptive fields. Atrous convolution can greatly enlarge the perceptive field of convolutional filters and keep the same computational cost with normal convolution. In 2D cases, atrous convolution with 3×3 kernel size can be simply formulated as below:

$$q[i, j] = bias + \sum_{k=-1}^{+1} \sum_{l=-1}^{+1} (x[i + k \cdot n, j + l \cdot n] \cdot w[k + 1, l + 1]), \quad (5)$$

where n indicates the atrous rate, w is the convolution weight of which the size is 3×3, q and x are output and input feature map, respectively, i and j are the feature map location. Note that $n = 1$ is the special case for normal convolution. To fully exploit useful features, we adopt the Squeeze-Excitation (SE) block [66] in our network, that is, using the attention mechanism for re-calibrating channel-wise convolutional feature responses. More specifically, each channel of the input feature map will be multiplied by a channel factor calculated by a SE block. The SE block consists of two linear layers, followed by a sigmoid function. The input feature map after global average pooling will be fed into this block and we can derive a channel factor ranging $(0, 1)$ for each input feature channel. We set the reduction rate in the SE block as 4, which means we set the output number of the first linear layer as the $1/4$ number of the input channels. To reduce the output channels by half, we add a 1×1 convolution layer after the SE block.

At last, we use a 3×3 convolution layer, in which the number of channels equals that of the input feature map, as the transition layer to the next module.

3) Attentive Feature Fusion: Traditional fusion strategy of top-down decoders [58], [67] treats the input feature maps equally. To better aggregate the feature maps, we propose an Attentive Feature Fusion (AFF) strategy. In our AFF fusion strategy, the feature map with a smaller size is more valued. As shown in Fig. 3 (b), the input feature maps M_E^i and M_D^{i+1} in the current stage are reduced to half size via 1×1 convolution layers. Then the reduced M_D^{i+1} is up-sampled by bilinear interpolation to output a double-sized feature map. We concatenate the two outputs and apply the SE block (also used in GAM) to produce an enhanced feature map. This feature map is then concatenated with the feature map of doubly up-sampled output in the previous stage. After the concatenation, we use another SE block to enhance the feature map again. After each SE block, we use a 3×3 convolution layer, with the same number of channels as the input, as the transition layer. A 1×1 convolution layer with a single neuron will be used to predict one feature map as the side-output of the current stage.

4) Combination With the Classification Model: As described above, we have designed two models, one for COVID-19 classification and the other one for COVID-19 opacification segmentation. However, they are separately
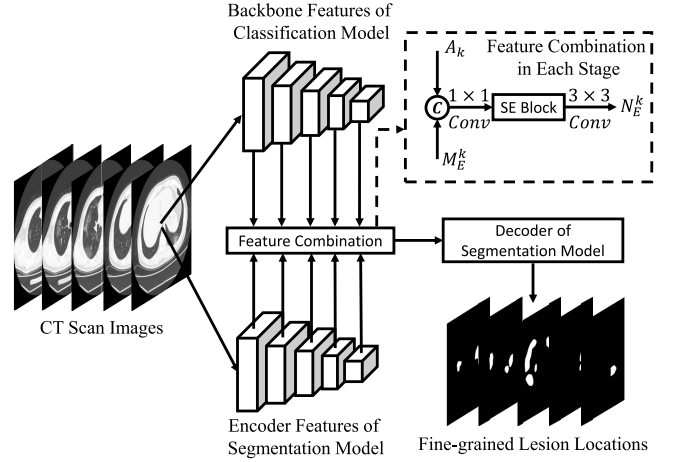

Fig. 4. Combination of the segmentation and classification models. We combine the encoder features of the segmentation model with the backbone features of the classification model.

working on the diagnosis system, and there might be a way to combine them together for better performance. Inspired by this, we leverage the features of the classification model to enhance the features of the segmentation model. As shown in Fig. 4, we merge the feature maps of each stage from the encoder of the segmentation model and the backbone of the classification model together. The feature maps of the encoder of the segmentation model are $M_E^1, M_E^2, M_E^3, M_E^4, M_E^5$ as defined in §III-B.1. The Res2Net [40] backbone of the classification model has five stages and we use the last feature maps A_k of stage $k \in [1, 5]$ for the feature combination. In merging the features of stage k, we have two feature maps A_k, M_E^k for the merge. We first resize the smaller one A_k, making it the same size as the larger one M_E^k, and concatenate them together. Then, we apply a simple 1×1 convolution layer for the feature channel reduction, making the output feature maps the same number of channels as M_E^k. Such 1×1 convolution layer is followed by a SE block with a reduction rate of 4. At last we use a 3×3 convolution layer of the same number of input and output channels as the transition layer. The output N_E^k will be regarded as the enhanced encoder features and be fed into the decoder of the segmentation model (Fig. 2). Then results are predicted as introduced in §III-B.1.

5) Deep Supervision Loss: Although the final prediction is only from the last side-output, we apply the deep supervision strategy [68] to all side-outputs with different sizes. For each side-output, we up-sample it to the size of the ground-truth map, and compute the sum of the standard binary cross-entropy loss and the Dice loss [69] as follows:

$$L = BCE(\mathbf{P}, \mathbf{G}) + 1 - \frac{\mathbf{P} \cdot \mathbf{G}}{\|\mathbf{P}\|_1 + \|\mathbf{G}\|_1}, \quad (6)$$

where the binary cross-entropy (BCE) loss is averaged among all $H \times W$ pixels, $p_{i,j}$ is the confidence score at pixel (i, j) calculated by a *sigmoid* function, and "·" means the dot product. $\mathbf{P}$ and $\mathbf{G}$ are predicted map and ground-truth map,
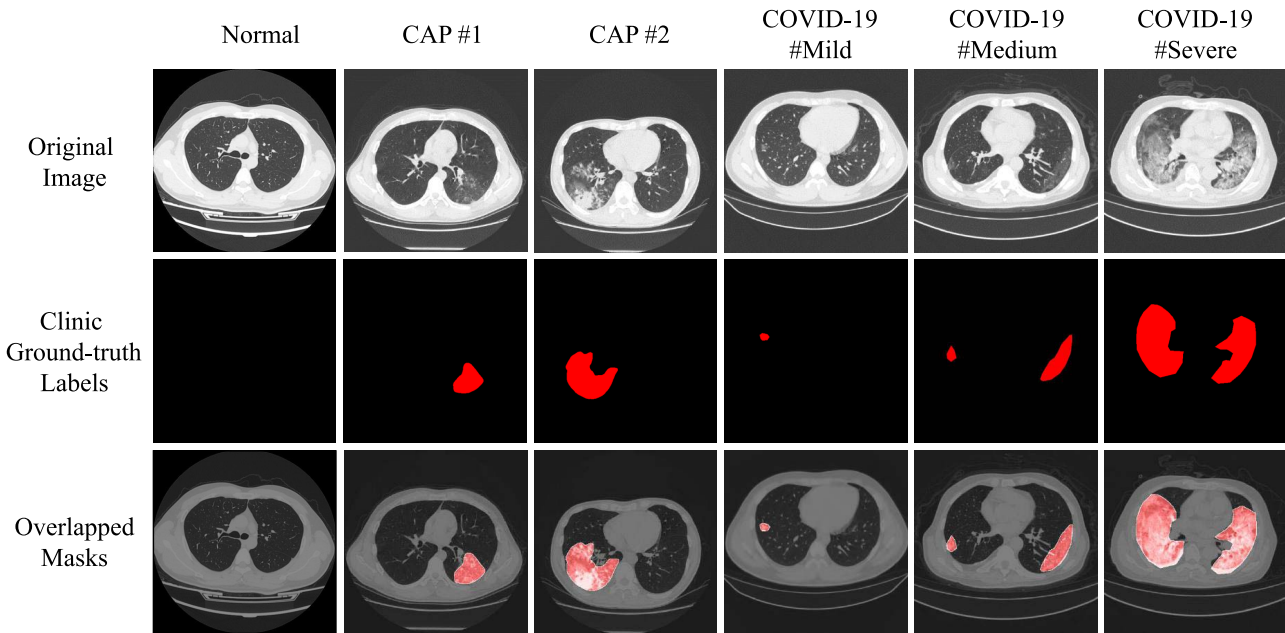
Fig. 5. Examples of our *COVID-CS* dataset, including CT scan images and labels of a normal person (1st column), two community-acquired pneumonia (CAP) cases (2nd and 3rd columns), and three COVID-19 patients from mild to severe ($4^{th} \sim 6^{th}$ columns).

respectively, while $\|\mathbf{P}\|_1$ and $\|\mathbf{G}\|_1$ denote the corresponding ℓ_1 norms.

C. Joint Diagnosis

An explainable classifier or accurate segmentation model itself could not fully implement comprehensive functions for COVID-19 diagnosis. Comparing to the segmentation model, our classifier is trained with CT images from both COVID-19 infected and uninfected cases, benefiting from more training data with lower annotation costs. Although our classifier can provide explainable lesion location of COVID-19 through activation mapping techniques, it cannot perform accurate and complete lesion segmentation. To this end, our segmentation model further provides complementary analysis by discovering the complete lesions in the lung and estimate the severity of the COVID-19 patients. But annotating vast segmentation labels by experienced radiologists is prohibitively expensive. To integrate their advantages for better application, we develop a diagnosis system for COVID-19 via joint explainable classification and segmentation models. In practice, our classification model will first predict whether the CT images of a suspected case to be COVID-19 positive or not. If the prediction is positive, the suspected case is very likely to be infected by COVID-19. Our segmentation model will then be performed on the CT images for in-depth analysis and to discover the whole opacification areas in each CT image.

IV. OUR *COVID-CS* DATASET

Data plays an essential role in the deep learning-based AI diagnosis systems. Currently, there are few publicly available COVID-19 datasets with either large scale samples or fine-grained pixel-level labeling. To fill in this gap, we construct a new COVID-19 Classification and Segmentation
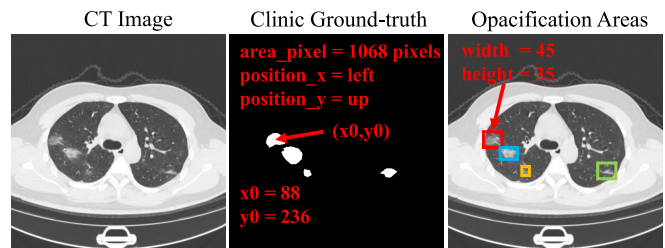

Fig. 6. Illustration of diverse information about opacification areas (in pixels), location (x0,y0), position (left, up), and width/height of opacification areas in our *COVID-CS* dataset.

(*COVID-CS*) dataset. In this section, we present the data collection, professional labeling, and statistics of our dataset. Fig. 5 shows some examples of our *COVID-CS* dataset (normal case and COVID-19 cases) and examples of CAP patients. Fig. 6 presents diverse information in the segmentation set of our *COVID-CS* dataset.

A. Data Collection

To protect the patients' privacy, we omit their personal information in our dataset construction. We collected 144,167 CT scan images from 750 cases, 400 of which are positive cases of COVID-19, and the other 350 cases are negative, all confirmed by RT-PCR tests. As previous studies [70] did, we do not take community-acquired pneumonia (CAP) patients (see Fig. 5) into consideration. Although CAP patients may be diagnosed as COVID-19 positive with our proposed diagnosis system since CT images of CAP patients also have similar opacifications, the threat of CAP is much less than that of COVID-19. And our purpose is to quickly develop an automatic diagnosis system and diagnose suspected cases as soon as possible. Besides, CAP patients can be simply diagnosed as COVID-19
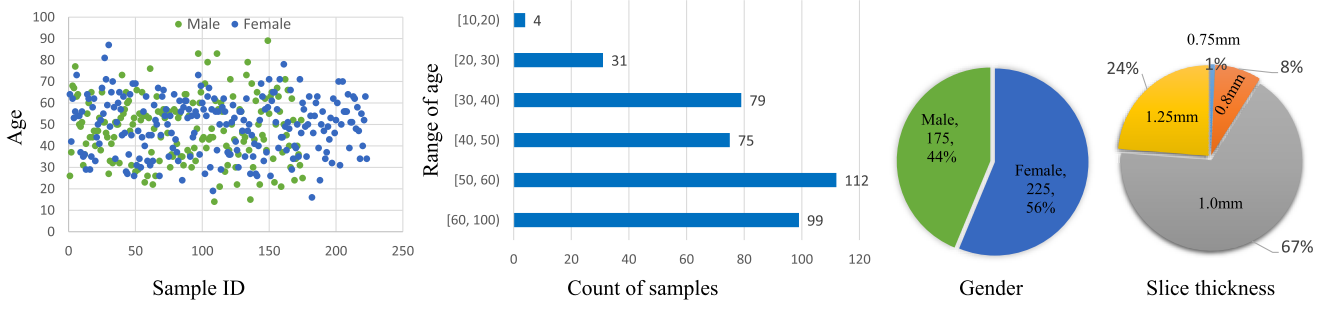
Fig. 7. The age, gender, and slice thickness distribution of the COVID-19 patients in our *COVID-CS* dataset. Zoom in for details.
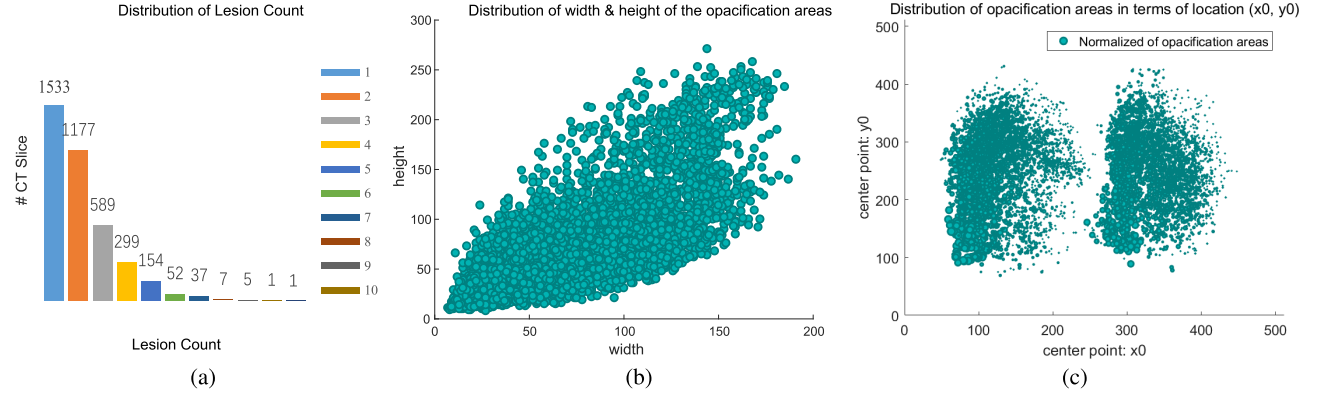

Fig. 8. Statistics of the segmentation set (200 COVID-19 cases) in our *COVID-CS* dataset. (a) Lesion count distribution. (b) The distribution of width & height of the opacification areas. (c) The relationship between the opacification areas and their locations.

TABLE III
THE CT SCANNERS AND NUMBERS OF POSITIVE CASES

Manufacturer	Product Name	#Cases
GE Medical Systems	Revolution CT	1
GE Medical Systems	LightSpeed VCT	6
GE Medical Systems	Discovery CT750 HD	12
GE Medical Systems	BrightSpeed	12
Toshiba	Aquilion ONE	33
GE Medical Systems	LightSpeed16	64
United Imaging Healthcare	uCT 780	272

negative with the help of the CAP/COVID-19 classifier [70], RT-PCR test, and the experience of doctors.

All involved patients underwent standard chest CT scans. Each case has $250 \sim 400$ CT images, and the number of CT images in each case is only determined by the type of the CT scanner and its scan settings. The CT scanners include BrightSpeed, Discovery CT750 HD, LightSpeed VCT, LightSpeed16, Revolution CT from GE Medical Systems, Aquilion ONE from Toshiba, and uCT 780 from United Imaging Healthcare. The numbers of cases from different scanners are summarized in Table III. The thickness of reconstructed CT slices ranges from 0.75mm to 1.25mm (Fig. 7 for more details).

B. Professional Labeling

We provide two aspects of labels for the collected CT scan images in our *COVID-CS* dataset, so as to implement joint classification and segmentation tasks. As mentioned above, our dataset is divided into 400 COVID-19 cases and 350 uninfected cases. For the segmentation task, we perform professional labeling through the following strategies:

- In order to save their labeling time, the radiologists only select at most 30 discrete CT scan images for each patient, in which the infections are observed for further annotation. In this step, our goal is to label every opacification area with pixel-level annotations.
- To generate high-quality annotations, we first invite a radiologist to mark as many opacification areas as possible based on his/her clinical experience. Then we invite another senior radiologist to refine the labeling marks several times for cross-validation. Some inaccurate labels are fixed after this step.

By implementing the above annotation procedures, we finally obtain 3,855 pixel-level labeled CT scan images of 200 COVID-19 patients with a resolution of 512×512. 64,771 CT images of the other 200 COVID-19 patients are without pixel-level annotation due to the shortage of radiologists, but such data will be used in classification tests. As can be seen in the last three columns of Fig. 5, our *COVID-CS* dataset covers different levels, *i.e.*, mild, medium, and severe, of COVID-19 cases.

C. Dataset Statistics

Age. The 400 COVID-19 patients (175 males and 225 females) range from 14 to 89 years, with an average age of 48.9 years. Fig. 7 shows the distribution of ages, the counts of samples in age ranges, and the gender percentages.

Lesion count. As shown in Fig. 8 (a), we illustrate the distribution of lesion counts. We observe that the lesion count distributes from 1 to 10 in each CT scan image.

Opacification areas. We plot the widths and heights of the opacification areas in Fig. 8 (b). The ranges of width and height are $7 \sim 191$ and $8 \sim 271$, respectively, showing diverse distributions.

Location. We also show the relationship between each opacification area and the corresponding central location (x0, y0) in Fig. 8 (c). As can be seen, the normalized opacification areas range from the smallest size (35/28452 pixels) to the largest size (28452/28452 pixels). It also shows that, in our *COVID-CS* dataset, the opacification areas are evenly distributed in diverse locations, which are also evenly distributed in the lungs.

V. EXPERIMENTS

A. Experimental Settings

Training/Test Protocol. For the segmentation task, our training set contains 2,794 images from 150 COVID-19 patients and the test set has 1,061 images from the other 50 COVID-19 cases. For the classification task, the training set contains the 2,794 images from the 150 COVID-19 infected cases in the segmentation set. In addition, we randomly pick 150 uninfected cases with 7,500 CT images as negative cases for training. The test set contains the 64,711 images of the other randomly selected 200 infected cases and the 68,041 images from 200 uninfected cases.

Evaluation Metrics. For the classification task, we adopt the widely used metrics of specificity and sensitivity as suggested by [26]. For the segmentation task, we use two standard metrics, *i.e.*, Dice score [71] and Intersection over Union (IoU). To provide a more comprehensive evaluation, we further use the widely used metric enhanced alignment measure (E_ϕ) [72].

Comparison methods. On the classification task, we compare our classification model with or without the image mixing technique [64]. On the segmentation task, to provide an in-depth evaluation of our *JCS* model, we compare it with versatile cutting-edge models, *i.e.*, the U-Net [58] for medical imaging and the DSS [73], PoolNet [74], and EGNet [75] for saliency detection.

B. Implementation Details

In our *JCS* system, the classification and segmentation models are trained separately. For the classification model, we train it with a batch-size of 256 on 4 GPUs. The CT images are resized into 224×224 for computational efficiency. We adopt the SGD optimizer with the initial learning rate of 0.1, divided by 10 in every 30 epochs. The classifier is trained with 100 epochs. For data augmentation, we use the random horizontal flip and random crop, and the image mixing technique [64] to alleviate the data bias. The α in the Beta distribution of image mixing is set as 0.5.

For the segmentation model, the number of CT images in each mini-batch is always 4, and the size of the input CT images is unchanged as 512×512. The backbone of our segmentation model is pretrained on ImageNet [36].

The atrous rates of four atrous convolutions in two sequential GAMs are {1, 3, 6, 9} and {1, 2, 3, 4}, respectively. The initial learning rate is 2.5×10^{-5}. We adopt the *poly* learning rate policy that the actual learning rate will be multiplied by a factor $(1 - \frac{cur\_iter}{max\_iter})^{power}$, where the power is 0.9. The segmentation model is trained with 21000 iterations. We employ the Adam [76] optimizer and set β_1, β_2 as 0.9 and 0.999, respectively. For data augmentation, we use random horizontal flip and random crop. When combined with the classification model, the classification model has been pretrained on our classification training set with pixel-level annotations.

C. Results

Activation mapping on explainable classification. Fig. 9 shows the visualization of activation mapping (AM) of our classification branch trained with or without image mixing [64]. At first, we train our classification model and achieve good performance in terms of sensitivity and specificity. But we find that The AM of our classification model initially trained with random horizontal flip and random crop (Fig. 9 (a)) not only covers the lesion areas, but also presents unrelated areas. If this problem is not solved, an automatic diagnosis system with an overfitted classification network is very harmful to clinical diagnosis. To solve this problem, we investigated and identified that the image mixing technique could solve this problem. By introducing the image mixing technique [64], the AM of our classification model provides more accurate locations of the opacification areas as shown in Fig. 9 (b). Moreover, Fig. 9 (c) indicates the AM of models trained with the help of pixel-level supervision (segmentation loss L_{seg} as introduced in §III-A.4). The AM of models becomes more accurate and specific in locating the opacifications. However, the improvements of adding segmentation loss L_{seg} in classification performance can be ignored, potentially due to saturated classification accuracy (No.3, Table IV).

Performance on explainable classification. During the inference, AM assists the medical experts using our *JCS* system to judge whether the prediction is correct or not. For each patient, opacifications can be found in some of its CT images and many images may have no opacifications. So we set a threshold for the classification. When the number of CT images from a suspected patient is larger than a threshold, the patient is diagnosed as COVID-19 positive. Changing the threshold enables our model to achieve a trade-off between sensitivity and specificity. Table IV shows the results of the classification model under different thresholds on the test set of our *COVID-CS* dataset. One can see that our model is very robust to the changing of thresholds, and achieves a sensitivity of 95.0% and a specificity of 93.0% when the threshold is 25. However, AM could not provide accurate segmentation of opacification areas (if any exist). Subsequently, we further employ our segmentation model to discover the opacification areas in the CT images of COVID-19 patients.

Ablation study on our EFM and AFF in the segmentation branch. In §III-B we introduced two novel modules named EFM and AFF for the segmentation. EFM is designed to
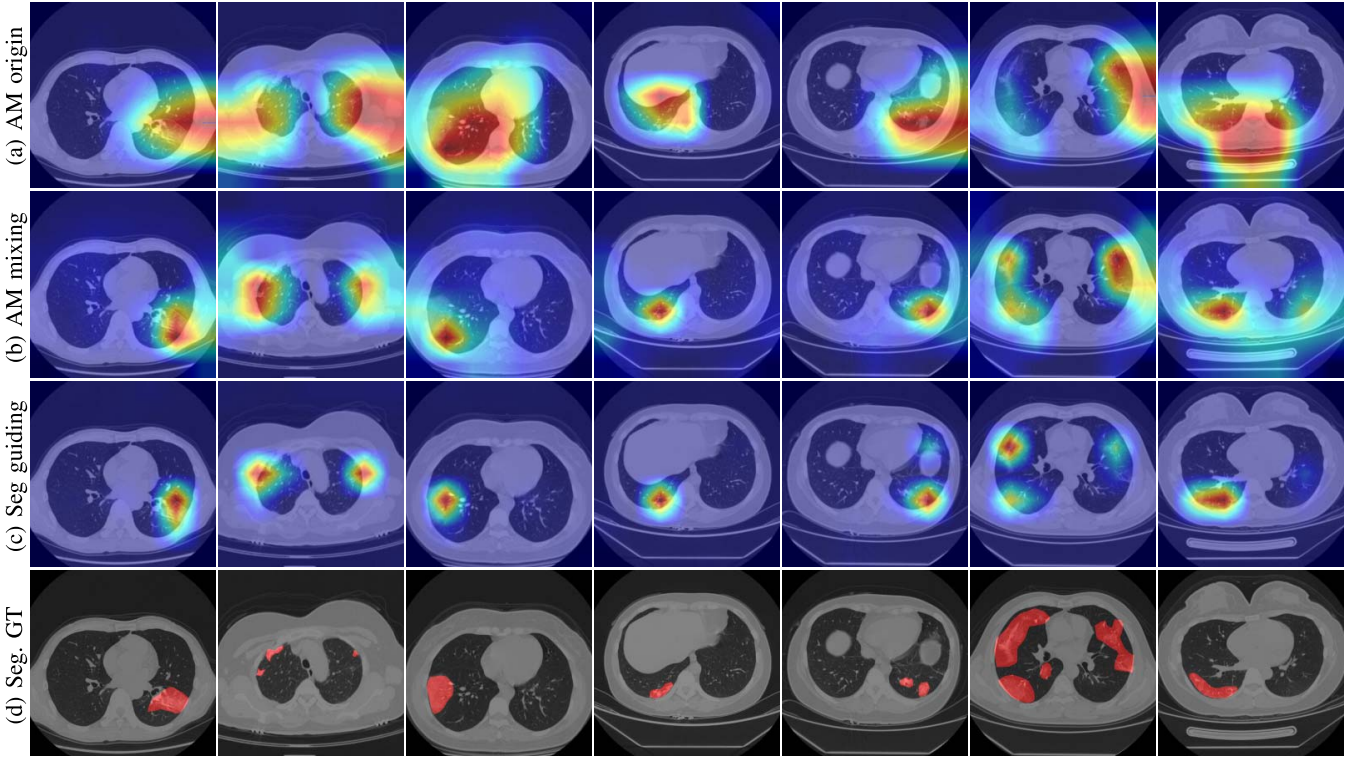
Fig. 9. Visualizations of activation mapping (AM). AM origin (mixing) means the AM of models trained without (with) image mixing technique [64]. Seg guiding means the AM of models trained with the segmentation loss L_{seg}.

TABLE IV

SENSITIVITY AND SPECIFICITY OF OUR CLASSIFICATION MODEL UNDER DIFFERENT THRESHOLDS. WE SET THE THRESHOLD AS 25 (THE GRAY ROW) IN THE FINAL SETTING

No.	Threshold	Sensitivity	Specificity
1	15	96.0%	91.5%
2	20	95.0%	92.0%
3	25	95.0%	93.0%
4	30	94.5%	93.5%

TABLE V

ABLATION STUDY FOR THE PROPOSED EFM AND AFF IN THE SEGMENTATION MODEL. THE BASELINE IS THE VGG16-BASED SEGMENTATION MODEL WITHOUT EFM&AFF (NO. 1). WE ADD EFM AND AFF SEPARATELY AND SHOW THE EFFECTIVENESS OF THEM (NO. 2 AND NO. 3). THE NO. 4 RESULT IS THE COMPLETE VERSION OF THE SEGMENTATION MODEL

No.	EFM	AFF	Dice	IoU	E_ϕ
1			71.0%	57.7%	88.0%
2	✔		74.3%	61.4%	88.9%
3		✔	75.9%	63.4%	90.9%
4	✔	✔	77.5%	65.4%	92.0%

enhance the representation power of our encoder in the segmentation branch. In the feature fusion stage, AFF is applied and the feature map with a smaller size is more valued while the traditional fusion strategy treats the input feature maps equally. The ablation studies for the proposed EFM and AFF are shown in Table V. The No.1 result is the baseline performance without EFM and AFM. After applying the proposed

TABLE VI

ABLATION STUDY FOR THE COMBINATION BETWEEN THE SEGMENTATION MODEL AND THE CLASSIFICATION MODEL. THE BASELINE SEGMENTATION RESULTS ARE GENERATED USING THE SEGMENTATION MODEL ONLY (NO.1). AFTER ADDITIONALLY ADDING FEATURES FROM THE CLASSIFICATION MODEL, WE ACHIEVE 1.0% IMPROVEMENT IN TERMS OF THE DICE METRIC (NO.2)

No.	SEG	+CLS	Dice	IoU	E_ϕ
1	✔		77.5%	65.4%	92.0%
2	✔	✔	78.5%	66.4%	92.7%

EFM and AFF separately to the baseline, the performance has 3.3% and 4.9% improvement in terms of the Dice metric. So both EFM and AFF are very helpful for the segmentation branch. When combining EFM with AFF, we achieve 6.5% higher results in terms of the Dice metric. The improvement in terms of the IoU and E-measure [72] metric is similar to that of the Dice metric. Hence, the proposed EFM and AFF are very beneficial for the segmentation model.

Ablation study on the combination between the segmentation model and classification model. As introduced in §III-B.4, we combine the classification model with the segmentation model for deriving more abundant features. To verify such a choice, we run the experiments as shown in Table VI. The baseline is the single segmentation model (No.1, Table VI). But we also observe that the choice of the combination of the classification model and segmentation model (No.2, Table VI) has 1.0% improvement in terms of the Dice metric, and shows
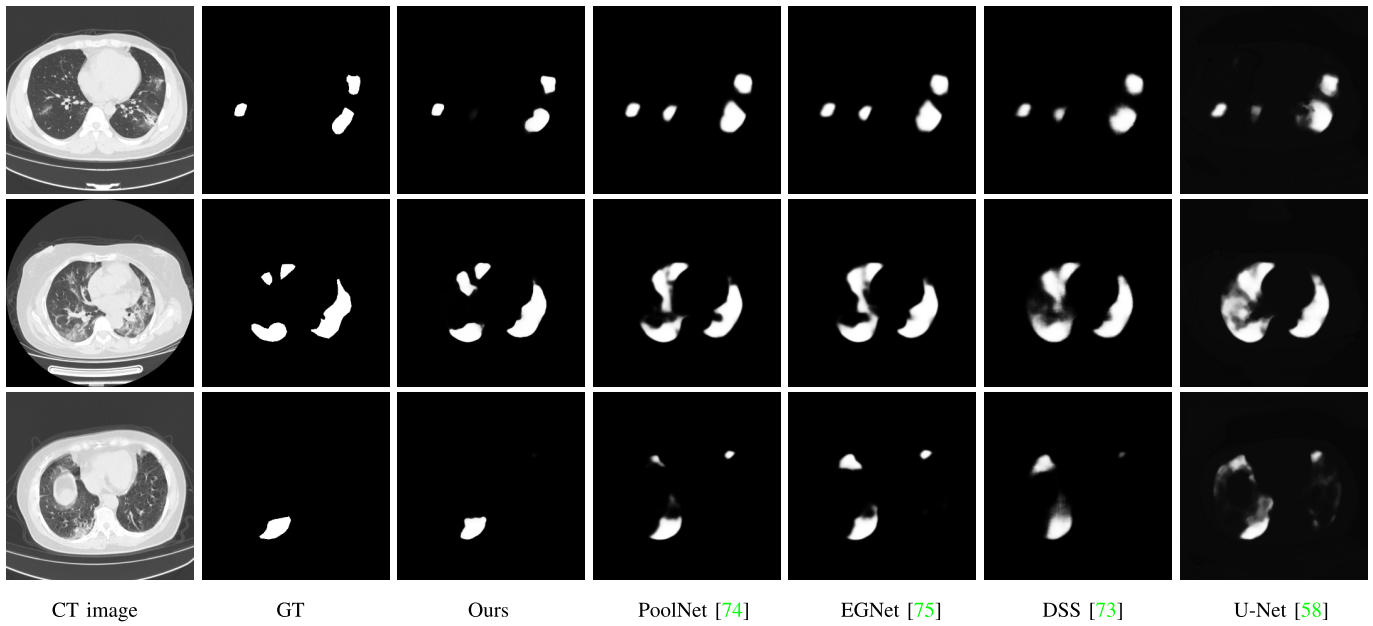
Fig. 10. Qualitative comparisons of different methods on our segmentation test set. The first, second, and third rows show the comparison results on CT images with different lesion areas from the mild, medium, and severe COVID-19 patients, respectively.

TABLE VII

QUANTITATIVE RESULTS ON OUR SEGMENTATION TEST SET

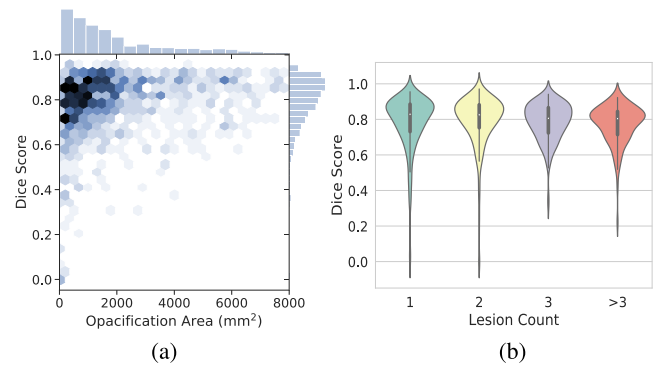| Methods	Publication	Dice	IoU	E_ϕ
U-Net [58]	MICCAI'15	65.1%	54.1%	79.7%
DSS [73]	TPAMI'19	65.7%	51.7%	79.9%
EGNet [75]	ICCV'19	69.3%	55.4%	83.6%
PoolNet [74]	CVPR'19	69.7%	55.9%	83.9%
JCS (Ours)	Submit'20	**78.5%**	**66.4%**	**92.7%**

Fig. 11. Statistical analysis for our segmentation model on our segmentation test set. (a) The relationship between the opacification area of each CT image and the corresponding Dice score. (b) The relationship between the lesion count and the corresponding probability distribution of the Dice score.

features of the classification model can certainly help the segmentation model predict better results.

Comparison of segmentation performance. Table VII lists the quantitative comparisons of 4 cutting-edge methods and our model on segmentation. It can be seen that the proposed model achieves the best result on all three metrics. It obtains improvements of 8.8%, 10.5%, and 8.8% on Dice score, IoU, and E_ϕ over the second-best PoolNet [74], respectively. Besides, PoolNet [74] and EGNet [75] obtain comparable results on the three metrics. U-Net [58] performs better than DSS [73] in terms of IoU, though they are comparable on the Dice score. Fig. 10 shows the qualitative results of the comparison methods. One can see that the other competitors produce inaccurate or even wrong predictions of the lesion areas in the CT images of mild, medium, and severe COVID-19 infections. But our segmentation model correctly discovers the whole lesion areas on all levels of COVID-19 infections.

To further study its stability, we perform a statistical analysis of our segmentation model on our segmentation test set. Fig. 11 (a) shows the correlation between the Dice score of our model and the opacification areas of CT images. Note that the CT images with the opacification area $\geq 8000mm^2$ are not plotted in Fig. 11 (a) since they only occupy 1.0% of all CT images in terms of quantity. We observe that 95.9% of CT

images have the Dice scores in [0.6, 1], while the other 3.3% of CT images are with Dice scores between [0.1, 0.6] and recognized as bad cases. Only 0.8% of CT images suffer from Dice scores of less than 0.1, and they are taken as failure cases. We also explore the relationship between the lesion count of each slice and the Dice score from a different perspective. As shown in Fig. 11 (b), the probability distribution of the Dice score is little affected by the number of lesion counts in a CT image. The medium dice score is above 0.8 for 4 different cases of lesion counts, and the 95.0% confidence interval lies in [0.5, 1]. We also observe that the lesion count of failure cases is ≤ 2. The consistently promising performance on segmenting lesion areas and the low probability (0.8%) of failure confirm the stability of our segmentation model.

Diagnosis of time. The speed test of the *JCS* system is on a single RTX 2080Ti. Assuming each suspected case has 300 CT images, the classification model in *JCS* only costs

about 1.0 second to ensure whether infected. If infected, The segmentation model will spend 21.0 seconds on fine-grained lesion segmentation. Hence, the *JCS* system costs 22.0 seconds for each infected case or 1.0 second for each uninfected case. Note that the complete RT-PCR test and radiologist CT diagnosis cost about 4 hours and 21.5 minutes, respectively, no matter the cases are infected or not.

VI. Conclusion

To facilitate the training of strong CNN models for COVID-19 diagnosis, in this paper, we systematically constructed a large scale COVID-19 Classification and Segmentation (*COVID-CS*) dataset. We also developed a Joint Classification and Segmentation (*JCS*) system for COVID-19 diagnosis. In our system, the classification model identified whether the suspected patient is COVID-19 positive or not, along with convincing visual explanations. It obtained a 95.0% sensitivity and 93.0% specificity on the classification test set of our *COVID-CS* dataset. To provide complementary pixel-level prediction, we implemented a segmentation model to discover fine-grained lesion areas in the CT images of COVID-19 patients. Comparing to the competing methods, e.g., Pool-Net [74], our segmentation model achieved an improvement of 8.8% on the Dice metric. Our *JCS* system is also very stable. On our segmentation test set, it failed only on 0.8% images and obtained Dice scores between [0.6, 1] for 95.9% of images. The online demo of our *JCS* diagnosis system for COVID-19 will be available soon.

References

[1] WHO. (2020). *Coronavirus Disease (COVID-19) Outbreak Situation*. [Online]. Available: https://www.who.int/emergencies/diseases/novel-coronavirus-2019

[2] H. B. Jenssen. *China National Health Commission Diagnosis and Treatment of Pneumonitis Caused by New Coronavirus (Trial Version 6)*. Accessed: Sep. 8, 2020. [Online]. Available: http://www.nhc.gov.cn/yzygj/s7653p/202002/8334a8326dd94d329df351d7da8ae%fc2.shtml

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

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

[5] Y. Wang, H. Hou, W. Wang, and W. Wang, "Combination of CT and RT-PCR in the screening or diagnosis of COVID-19," *J. Global Health*, vol. 10, no. 1, pp. 1–4, Jun. 2020.

[6] J. Zhang, Y. Xie, Y. Li, C. Shen, and Y. Xia, "COVID-19 screening on chest X-Ray images using deep learning based anomaly detection," 2020, *arXiv:2003.12338*. [Online]. Available: https://arxiv.org/abs/2003.12338

[7] S. Inui *et al.*, "Chest ct findings in cases from the cruise ship 'diamond princess' with coronavirus disease 2019 (covid-19)," *Radiol., Cardiothoracic Imag.*, vol. 2, no. 2, 2020, Art. no. e200110.

[8] Z. Huang *et al.*, "The battle against coronavirus disease 2019 (COVID-19): Emergency management and infection control in a radiology department," *J. Amer. College Radiol.*, vol. 17, no. 6, pp. 710–716, Jun. 2020.

[9] J. Paul Cohen, P. Morrison, and L. Dao, "COVID-19 image data collection," 2020, *arXiv:2003.11597*. [Online]. Available: http://arxiv.org/abs/2003.11597

[10] X. Yang, X. He, J. Zhao, Y. Zhang, S. Zhang, and P. Xie, "COVID-CT-dataset: A CT scan dataset about COVID-19," 2020, *arXiv:2003.13865*. [Online]. Available: http://arxiv.org/abs/2003.13865

[11] N. Sajid. *COVID-19 Patients Lungs X-Ray Images 10000*. Accessed: Oct. 4, 2020. [Online]. Available: https://www.kaggle.com/nabeelsajid917/covid-19-x-ray-10000-images

[12] H. B. Jenssen. *COVID-19 Ct Segmentation Dataset*. Accessed: Oct. 4, 2020. [Online]. Available: http://medicalsegmentation.com/covid19/

[13] L. Li *et al.*, "Artificial intelligence distinguishes COVID-19 from community acquired pneumonia on chest ct," *Radiology*, vol. 296, no. 2, pp. E65–E71, Aug. 2020, doi: 10.1148/radiol.2020200905.

[14] M. Chung *et al.*, "CT imaging features of 2019 novel coronavirus (2019-nCoV)," *Radiology*, vol. 295, no. 1, pp. 202–207, Apr. 2020.

[15] A. Bernheim *et al.*, "Chest CT findings in coronavirus disease-19 (COVID-19): Relationship to duration of infection," *Radiology*, vol. 295, no. 3, Jun. 2020, Art. no. 200463, doi: 10.1148/radiol.2020200463.

[16] Y. Wang *et al.*, "Temporal changes of CT findings in 90 patients with COVID-19 pneumonia: A longitudinal study," *Radiology*, vol. 296, no. 2, pp. E55–E64, Aug. 2020.

[17] H. Shi *et al.*, "Radiological findings from 81 patients with COVID-19 pneumonia in Wuhan, China: A descriptive study," *Lancet Infectious Diseases*, vol. 20, no. 4, pp. 425–434, Apr. 2020.

[18] Y. Qiu, Y. Liu, S. Li, and J. Xu, "MiniSeg: An extremely minimum network for efficient COVID-19 segmentation," 2020, *arXiv:2004.09750*. [Online]. Available: http://arxiv.org/abs/2004.09750

[19] 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.

[20] 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," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 618–626.

[21] J. Won *et al.*, "Development of a laboratory-safe and low-cost detection protocol for SARS-CoV-2 of the coronavirus disease 2019 (COVID-19)," *Experim. Neurobiol.*, vol. 29, no. 5, p. 402, Oct. 2020.

[22] B. Udugama *et al.*, "Diagnosing COVID-19: The disease and tools for detection," *ACS Nano*, vol. 14, no. 4, pp. 3822–3835, Apr. 2020.

[23] W. Yang and F. Yan, "Patients with RT-PCR-confirmed COVID-19 and normal chest CT," *Radiology*, vol. 295, no. 2, p. E3, May 2020.

[24] Z. Hu *et al.*, "Clinical characteristics of 24 asymptomatic infections with COVID-19 screened among close contacts in nanjing, China," *Sci. China Life Sci.*, vol. 63, no. 5, pp. 706–711, May 2020.

[25] C. Long *et al.*, "Diagnosis of the coronavirus disease (COVID-19): RRT-PCR or CT?" *Eur. J. Radiol.*, vol. 126, May 2020, Art. no. 108961.

[26] Y. Liu, Y.-H. Wu, Y. Ban, H. Wang, and M.-M. Cheng, "Rethinking computer-aided tuberculosis diagnosis," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 2646–2655.

[27] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[28] M. Farooq and A. Hafeez, "COVID-ResNet: A deep learning framework for screening of COVID19 from radiographs," 2020, *arXiv:2003.14395*. [Online]. Available: http://arxiv.org/abs/2003.14395

[29] R. Yang *et al.*, "Chest CT severity score: An imaging tool for assessing severe COVID-19," *Radiol., Cardiothoracic Imag.*, vol. 2, no. 2, Apr. 2020, Art. no. e200047.

[30] K. Li *et al.*, "CT image visual quantitative evaluation and clinical classification of coronavirus disease (COVID-19)," *Eur. Radiol.*, vol. 30, no. 8, pp. 4407–4416, Aug. 2020.

[31] Z. Zhou *et al.*, "Coronavirus disease 2019: Initial chest CT findings," *Eur. Radiol.*, vol. 30, no. 8, pp. 4398–4406, Aug. 2020.

[32] L. Wynants *et al.*, "Prediction models for diagnosis and prognosis of COVID-19: Systematic review and critical appraisal," *Brit. Med. J.*, vol. 369, Apr. 2020, Art. no. m1328.

[33] V. Rajinikanth, N. Dey, A. N. J. Raj, A. E. Hassanien, K. C. Santosh, and N. S. M. Raja, "Harmony-search and otsu based system for coronavirus disease (COVID-19) detection using lung CT scan images," 2020, *arXiv:2004.03431*. [Online]. Available: http://arxiv.org/abs/2004.03431

[34] J. B. T. M. Roerdink and A. Meijster, "The watershed transform: Definitions, algorithms and parallelization strategies," *Fundamenta Informaticae*, vol. 41, no. 1,2, pp. 187–228, 2000.

[35] T. Zhou, S. Canu, and S. Ruan, "An automatic COVID-19 CT segmentation network using spatial and channel attention mechanism," 2020, *arXiv:2004.06673*. [Online]. Available: http://arxiv.org/abs/2004.06673

[36] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[37] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inform. Process. Syst.*, 2012, pp. 1097–1105.

[38] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Represent.*, 2015, pp. 1–14.

[39] 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.

[40] S.-H. Gao, M.-M. Cheng, K. Zhao, X.-Y. Zhang, M.-H. Yang, and P. Torr, "Res2Net: A new multi-scale backbone architecture," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 2, pp. 652–662, Feb. 2021.

[41] M.-M. Cheng *et al.*, "HFS: Hierarchical feature selection for efficient image segmentation," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 867–882.

[42] Y. Liu *et al.*, "DEL: Deep embedding learning for efficient image segmentation," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, Jul. 2018, p. 870.

[43] J. Zhao, R. Bo, Q. Hou, M.-M. Cheng, and P. Rosin, "FLIC: Fast linear iterative clustering with active search," *Comput. Vis. Media*, vol. 4, no. 4, pp. 333–348, Dec. 2018.

[44] Y. Liu, Y.-C. Gu, X.-Y. Zhang, W. Wang, and M.-M. Cheng, "Light-weight salient object detection via hierarchical visual perception learning," *IEEE Trans. Cybern.*, early access, Dec. 7, 2020, doi: 10.1109/TCYB.2020.3035613.

[45] K. Zhao, J. Xu, and M.-M. Cheng, "RegularFace: Deep face recognition via exclusive regularization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1136–1144.

[46] Y.-Q. Tan, S.-H. Gao, X.-Y. Li, M.-M. Cheng, and B. Ren, "VecRoad: Point-based iterative graph exploration for road graphs extraction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 8910–8918.

[47] M.-M. Cheng, X.-C. Liu, J. Wang, S.-P. Lu, Y.-K. Lai, and P. L. Rosin, "Structure-preserving neural style transfer," *IEEE Trans. Image Process.*, vol. 29, pp. 909–920, 2020.

[48] J.-W. Bian *et al.*, "An evaluation of feature matchers for fundamental matrix estimation," in *Proc. Brit. Mach. Vis. Conf.*, 2019, pp. 1–4.

[49] L. Zhang *et al.*, "Nonlinear regression via deep negative correlation learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 3, pp. 982–998, Mar. 2021.

[50] J. Xu *et al.*, "Noisy-as-clean: Learning self-supervised denoising from corrupted image," *IEEE Trans. Image Process.*, vol. 29, pp. 9316–9329, 2020.

[51] Q. Wang, J. Gao, and X. Li, "Weakly supervised adversarial domain adaptation for semantic segmentation in urban scenes," *IEEE Trans. Image Process.*, vol. 28, no. 9, pp. 4376–4386, Sep. 2019.

[52] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.

[53] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.

[54] Y. Liu, Y.-H. Wu, P.-S. Wen, Y.-J. Shi, Y. Qiu, and M.-M. Cheng, "Leveraging instance-, image- and dataset-level information for weakly supervised instance segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Sep. 10, 2020, doi: 10.1109/TPAMI.2020.3023152.

[55] Y.-H. Wu, Y. Liu, L. Zhang, W. Gao, and M.-M. Cheng, "Regularized densely-connected pyramid network for salient instance segmentation," 2020, *arXiv:2008.12416*. [Online]. Available: http://arxiv.org/abs/2008.12416

[56] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 2, pp. 386–397, Feb. 2020.

[57] K. Sofiiuk, K. Sofiyuk, O. Barinova, A. Konushin, and O. Barinova, "AdaptIS: Adaptive instance selection network," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 7355–7363.

[58] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention—MICCAI*, N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, Eds. Cham, Switzerland: Springer, 2015, pp. 234–241.

[59] O. Çiçek, A. Abdulkadir, S. Lienkamp, T. Brox, and O. Ronneberger, "3D U-Net: Learning dense volumetric segmentation from sparse annotation," in *Medical Image Computing and Computer-Assisted Intervention—MICCAI*. Cham, Switzerland: Springer, Oct. 2016, pp. 424–432.

[60] V. Iglovikov and A. Shvets, "TernausNet: U-Net with VGG11 encoder pre-trained on ImageNet for image segmentation," 2018, *arXiv:1801.05746*. [Online]. Available: http://arxiv.org/abs/1801.05746

[61] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: Redesigning skip connections to exploit multiscale features in image segmentation," *IEEE Trans. Med. Imag.*, vol. 39, no. 6, pp. 1856–1867, Jun. 2020.

[62] Y. Xiong *et al.*, "Clinical and high-resolution ct features of the COVID-19 infection: Comparison of the initial and follow-up changes," *Invest. Radiol.*, vol. 55, no. 6, pp. 332–339, Jun. 2020, doi: 10.1097/RLI.0000000000000674.

[63] A. N. Leung, R. R. Miller, and N. L. Müller, "Parenchymal opacification in chronic infiltrative lung diseases: CT-pathologic correlation.," *Radiology*, vol. 188, no. 1, pp. 209–214, Jul. 1993.

[64] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopezpaz, "Mixup: Beyond empirical risk minimization," in *Proc. Int. Conf. Learn. Represent.*, 2018, pp. 1–13.

[65] K. Li, Z. Wu, K.-C. Peng, J. Ernst, and Y. Fu, "Guided attention inference network," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 12, pp. 2996–3010, Dec. 2020.

[66] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.* Cham, Switzerland: Springer, Jun. 2018, pp. 7132–7141.

[67] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2117–2125.

[68] C.-Y. Lee, S. Xie, P. Gallagher, Z. Zhang, and Z. Tu, "Deeply-supervised nets," in *Artificial Intelligence and Statistics*, 2015, pp. 562–570.

[69] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-Net: Fully convolutional neural networks for volumetric medical image segmentation," in *Proc. 4th Int. Conf. 3D Vis. (3DV)*. Cham, Switzerland: Springer, Oct. 2016, pp. 565–571.

[70] F. Shi *et al.*, "Large-scale screening of COVID-19 from community acquired pneumonia using infection size-aware classification," 2020, *arXiv:2003.09860*. [Online]. Available: http://arxiv.org/abs/2003.09860

[71] F. Shan *et al.*, "Lung infection quantification of COVID-19 in CT images with deep learning," 2020, *arXiv:2003.04655*. [Online]. Available: http://arxiv.org/abs/2003.04655

[72] D.-P. Fan, C. Gong, Y. Cao, B. Ren, M.-M. Cheng, and A. Borji, "Enhanced-alignment measure for binary foreground map evaluation," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, Jul. 2018, pp. 698–704.

[73] Q. Hou, M.-M. Cheng, X. Hu, A. Borji, Z. Tu, and P. H. S. Torr, "Deeply supervised salient object detection with short connections," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 4, pp. 815–828, Apr. 2019.

[74] J.-J. Liu, Q. Hou, M.-M. Cheng, J. Feng, and J. Jiang, "A simple pooling-based design for real-time salient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3917–3926.

[75] J. Zhao, J.-J. Liu, D.-P. Fan, Y. Cao, J. Yang, and M.-M. Cheng, "EGNet: Edge guidance network for salient object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 8779–8788.

[76] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Represent.*, 2015, pp. 1–15.
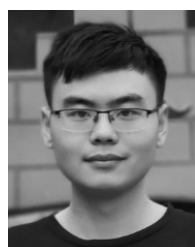
Yu-Huan Wu received the bachelor's degree from Xidian University, in 2018. He is currently pursuing the Ph.D. degree with the College of Computer Science, Nankai University, supervised by Prof. M.-M. Cheng. His research interests include computer vision and machine learning.

Shang-Hua Gao is currently pursuing the Ph.D. degree with the Media Computing Laboratory, Nankai University. He is also supervised via Prof. M.-M. Cheng. His research interests include computer vision, machine learning, and radio vortex wireless communications.

Jie Mei is currently pursuing the Ph.D. degree with the College of Computer Science, Nankai University, Tianjin, China. His research interests include computer vision, machine learning, and remote sensing image processing.

Jun Xu (Member, IEEE) received the B.Sc. and M.Sc. degrees from the School of Mathematics Science, Nankai University, Tianjin, China, in 2011 and 2014, respectively, and the Ph.D. degree from the Department of Computing, Hong Kong Polytechnic University, in 2018. He worked as a Research Scientist with IIAI, Abu Dhabi, United Arab Emirates. He is currently a Lecturer with the School of Statistics and Data Science, Nankai University.

Deng-Ping Fan (Member, IEEE) received the Ph.D. degree from Nankai University, in 2019. He joined the Inception Institute of Artificial Intelligence (IIAI), in 2019. He has published about 20 top journal and conference papers, such as CVPR and ICCV. His research interests include computer vision, deep learning, saliency detection, especially on co-salient object detection, RGB salient object detection, RGB-D salient object detection, and video salient object detection.

Rong-Guo Zhang received the Ph.D. degree majoring in pattern recognition and intelligent systems from the Institute of Automation, Chinese Academy of Sciences, in 2012. He currently serves as the Head of the Institute of Advanced Research, Beijing Infervision Technology Company, Ltd. His research interests include computer vision, deep learning, and medical image processing.

Ming-Ming Cheng (Senior Member, IEEE) received the Ph.D. degree from Tsinghua University in 2012. Then, he did two years Research Fellow with Prof. Philip Torr in Oxford. He is currently a Professor with Nankai University, leading the Media Computing Laboratory. His research interests include computer graphics, computer vision, and image processing. He received research awards, including the ACM China Rising Star Award, the IBM Global SUR Award, and the CCF-Intel Young Faculty Researcher Program. He is on the editorial boards of IEEE TRANSACTIONS ON IMAGE PROCESSING (TIP).