BCS-Net: Boundary, Context, and Semantic for Automatic COVID-19 Lung Infection Segmentation From CT Images

Runmin Cong, Member, IEEE, Haowei Yang, Qiuping Jiang, Member, IEEE, Wei Gao, Senior Member, IEEE, Haisheng Li, Cong Wang, Yao Zhao, Senior Member, IEEE, and Sam Kwong, Fellow, IEEE

Abstract—The spread of COVID-19 has brought a huge disaster to the world, and the automatic segmentation of infection has become an urgent task. However, there are several challenges for the accurate and complete segmentation, such as the scattered infection area distribution, complex background noises, and blurred segmentation boundaries. To this end, in this article, we propose a novel network for automatic COVID-19 lung infection segmentation from computed tomography (CT) images, named BCS-Net, which considers the boundary, context, and semantic attributes. The BCS-Net follows an encoder-decoder architecture, and more designs focus on the decoder stage that includes three progressively boundary-context-semantic reconstruction (BCSR) blocks. In each BCSR block, the attention-guided global context (AGGC) module is designed to learn the most valuable encoder features for decoder by highlighting the important spatial and boundary locations and modeling the global context dependence. Besides, a semantic guidance (SG) unit generates the SG map to refine the decoder features by aggregating multiscale high-level features at the intermediate resolution. Extensive experiments demonstrate that our proposed framework outperforms the existing competitors both qualitatively and quantitatively.

Index Terms—Boundary-context-semantic reconstruction (BCSR), COVID-19, infection segmentation, lung computed tomography (CT) image.

I. INTRODUCTION

GLOBALLY, as of March 2022, more than 452 million confirmed cases of COVID-19 have been reported to the World Health Organization (WHO), including more than six million deaths. Especially, the new wave of epidemics caused by the Delta and Omicron variants of COVID that broke out in India, South Korea, and Hong Kong from 2021 is more contagious, and the global epidemic situation still cannot be relaxed.

As a rapid and large-scale COVID-19 testing method, the reverse transcription-polymerase chain reaction (RT-PCR) testing has been widely adopted worldwide, but its false negative rate is as high as 17%–25.5%, which is only suitable for preliminary screening. In clinical practice, to make a definite diagnosis of suspected cases and determine an appropriate treatment plan, lung imaging interpretation by ultrasound [1], [2], X-rays [3], [4], [5], or computed tomography (CT) [6], [7], [8], [9], [10], [11], [12], [13] is an indispensable link. However, different imaging devices have their own advantages. Ultrasound can propagate in a certain direction and penetrate objects. Based on the principle that ultrasonic waves generate echoes, we can collect and display such echoes.
In recent years, the vigorous development of deep learning has greatly promoted the development of computer vision-related fields [14], [15], [16], [17], [18], [19], [20], [21], [22], [23], [24]. Among them, medical image segmentation algorithms have made great progress and achieved a qualitative leap in performance, such as brain image synthesis [25], polyp segmentation [26], COVID-19 infection detection [27], COVID-19 forecasting [28], and COVID-19 infection segmentation [8], [12]. However, differences in imaging equipment and disease characteristics make it difficult to use a unified segmentation model for different diseases. Researchers often need to design some unique modules to better achieve lesion area segmentation. Observing an example shown in Fig. 1, we can summarize three difficult problems that the COVID-19 lung infection segmentation models need to solve.

1) The infection regions of COVID-19 are very scattered, with many isolated areas of various sizes, which are very challenging for complete detection. Moreover, the sizes of the infected areas vary greatly, which makes it more difficult to accurately detect the infected areas of different scales. In addition, the boundary of the infection region is not easy to segment accurately and sharply. To this end, we can seek a solution from the perspective of encoder features containing relatively rich and effective information, which can be used to guide the feature learning in the decoder stage. Concretely, we propose an attention-guided global context (AGGC) module to select the most valuable information from encoder features for decoder, where the spatial attention (SA) and boundary attention (BA) are used to provide more accurate important spatial and boundary guidance information, and the global context modeling (GCM) unit is used to model the global context dependence and constrain the generation of more complete segmentation result.

2) The infected areas have many detailed boundaries, and more attention needs to be paid to the clearness and sharpness of the boundaries in the segmentation results. From a clinical point of view, the boundary of the infected area is of great significance for diagnosis. However, due to the pooling or upsampling/downsampling operation, the CNN-based method may blur the boundaries of the segmentation result. To generate a clearer and sharper boundary, we introduce boundary guidance in the proposed network. In addition to introducing the supervised learning of the boundary map, we also treat the generated boundary map as an attention weight to highlight the encoder features in the AGGC module and allow the boundary constraint to play a greater role in the network.

3) For the COVID-19-infected lung CT image, although the processed data are only limited to the image of the lung area, the background regions are relatively complex, and there are still a lot of interferences, such as inflamed areas that are not infected by COVID-19. Therefore, it is crucial to effectively suppress background noises. Considering that the high-level features, including more
semantic and category attributes, can be used to suppress complex backgrounds, we design a semantic guidance (SG) unit to aggregate multiscale high-level features and formulate an SG map to refine the decoder features. Moreover, to alleviate the information loss from multiple sampling operations and maintain the correlation between feature maps of adjacent scales, the multiscale feature fusion is unified on the intermediate resolution to generate the SG map.

In summary, we propose an automatic COVID-19 lung infection segmentation network with the encoder–decoder structure, equipped with three boundary–context–semantic reconstruction (BCSR) blocks. Although the encoder–decoder architecture is a common structure in medical image segmentation, we make a delicate design in it to cope with the special characteristics of the COVID-19 infection segmentation task, such as relatively scattered infected regions, more boundary details, complex backgrounds, and noisy interferences. First, to address the problem of inaccurate and incomplete detection caused by the scattered infected regions, we design an AGGC module to select the most valuable information from encoder features for decoder, which highlights the important spatial and boundary locations in an attention manner and models the global context dependence for the complete segmentation. Second, to achieve the sharp boundaries of the final segmentation result, we introduce the BA unit by adding the boundary supervised learning and boundary map refinement. Third, to suppress the background interferences, an SG unit is designed to aggregate multiscale high-level features and formulate an SG map to refine the decoder features. All modules cooperate with each other to make our network achieve the competitive performance. As shown in Fig. 1(c), our proposed method has advantages in detection accuracy, completeness, clarity, and sharpness. The main contributions of this article are as follows.

1) An end-to-end network, named BCS-Net, is proposed to achieve automatic COVID-19 lung infection segmentation from CT images, which models the boundary constraint, context relationship, and SG. Moreover, our network achieves the competitive performance on the publicly available dataset both qualitatively and quantitatively.

2) An AGGC module is designed to filter the most valuable encoder information for decoder, which highlights the important spatial and boundary locations in an attention manner and models the global context dependence for the complete segmentation.

3) An SG unit is proposed to aggregate multiscale high-level features and generate the SG map to refine the decoder features, in which the multiscale fusion is unified on the intermediate resolution to alleviate the information loss and maintain the correlation between adjacent scales.

II. RELATED WORK

In this section, we will briefly introduce some related works in COVID-19 diagnosis based on deep learning for CXR and CT images, such as classification and segmentation. There exist lots of deep learning-based segmentation methods, such as Mask regions with convolutional neural network (R-CNN) [29], You Only Look At CoefficienTs (YOLACT) [30], DeepLab [31], and fully convolutional network (FCN) [32], which have made great achievements in segmentation tasks of various scenes and can be used for the medical segmentation task. But, there are some differences between medical images and traditional ordinary images. Taking the COVID-19-infected lung CT image as an example, the areas to be segmented are usually scattered, with more details, complex backgrounds, and more interfering noises. Therefore, after the emergence of the U-Net [33], it has gradually become the infrastructure for medical image segmentation and even general image segmentation tasks. Its success lies in the fact that a symmetric encoder–decoder structure with skip connections can better and more comprehensively utilize features at different levels, thereby generating more task-discriminative representations and improving performance.

Due to the high sensitivity and clarity of CT images, most of the current diagnosis of COVID-19 is based on CT images, including segmentation [6], [7], [8], [9], [10], [11]. In practice, segmenting the key area from chest CT can provide useful information for medical staffs to diagnose the COVID-19, including ground-glass opacity (GGO) and consolidation. So far, many COVID-19 lung infection segmentation methods from CT images based on deep learning have been proposed, and promising performance has been obtained. Zhou et al. [6] proposed an automated segmentation network by integrating the spatial and channel attention mechanisms. Fan et al. [8] designed the parallel partial decoder, reverse attention, and edge attention to boost the segmentation accuracy and also provided a semi-supervised framework to alleviate the shortage of labeled data. Paluru et al. [9] proposed an anamorphic depth embedding-based lightweight CNN to segment anomalies in COVID-19 chest CT images, which incorporates fully convolutional anamorphic depth blocks with depth-wise squeezing and stretching after the downsampling and upsampling operations. Wang et al. [10] proposed a noise-robust learning framework based on self-ensembling of CNNs. Han et al. [11] proposed an attention-based deep 3-D multiple instance learning (AD3D-MIL) to achieve the goal of accurate and interpretable screening of COVID-19 from chest CT images.

In these previous works, some ingenious and effective structures were designed, which achieved good performance and brought us a lot of inspiration. However, the existing work does not fully consider the following two points: 1) the loss of semantic information caused by oversampling in multilayer feature fusion and 2) the correlation between contextual information and semantic information plays an important role in complete segmentation. For the first issue, we use the middle scale as the output criterion in the SG unit to generate the SG mask, which avoids excessive upsampling and downsampling of a single layer in the multilayer feature fusion. As for the second issue, we design an AGGC module to select the most valuable information from encoder features for decoder, which can not only highlight the important spatial and boundary locations, but also perceive the global correlation between different areas.
Fig. 2. Illustration of the overall framework of the proposed network. The input CT slice image is first embedded into the backbone with four layers to extract the multilevel features; then, we utilize the stacked BCSR blocks to progressively reconstruct the segmentation results. The BCSR block consists of AGGC module, SG unit, and BA unit. The framework finally produces five prediction maps, in which the $S_2$ map is the final segmentation result.

Fig. 3. Illustration of the proposed AGGC module, integrating the important locations, boundary details, and global context. “SA” is the spatial attention unit, “$S_b$” denotes BA map generated by the BA unit, “GCM” is the global context modeling unit that models the global context dependence, and “G” is the global context-aware feature map.

III. PROPOSED METHOD

A. Overview

Fig. 2 illustrates the overall framework of our proposed BCS-Net to achieve the COVID-19 lung infection segmentation from CT images. Our network follows an encoder–decoder architecture in an end-to-end manner, in which the backbone extractor [34] in the encoder stage aims to extract the top-down multilevel features, and the decoder stage includes three progressively BCSR blocks to learn the segmentation-related features and generate the segmentation mask. In each BCSR block, we jointly consider the global SG, context dependence modeling, and BA refinement to provide more sufficient supplementary information for feature decoding. Specifically, to deal with the problem of scattered lesion location and irregular shape, we design an AGGC module, which selects the most valuable encoder propagation features by performing attention and context modeling on the encoder features of the corresponding decoder layer. As shown in Fig. 3, the corresponding encoder features and BA map generated by the BA unit are fed into the AGGC module. The encoder features are first refined by the SA and BA to highlight the important spatial locations and boundary details. Then, the GCM unit is used to correlate the different locations to consistently detect the lesion regions and generate the enhanced encoder features. With the guidance of the enhanced encoder features, the initial decoder features of the current level are generated by combining with the previous decoder features generated by the previous BCSR block. Furthermore, considering the importance of high-level semantic information for suppressing irrelevant background noises, we design an SG unit to aggregate multiscale high-level features and formulate an SG map, which is further combined with the decoder features by means of residual connection to obtain the corresponding final decoder features

$$F^i = \delta(\text{conv}(\bar{F}^i + \bar{F}^i \circ S_d))$$

where $\bar{F}^i = \text{concat}(F^{i+1}, f_{aggc}^i)$, \text{concat}() is channel-wise concatenation operation, the values of $f_{aggc}^i$ are the output features generated by the AGGC module, $S_d$ denotes the SG map generated by the SG unit, $\circ$ represents element-wise multiplication, $\delta(\cdot)$ denotes the rectified linear unit (ReLU) activation, and \text{conv}() is a customized convolutional block. At each BCSR block, we use a convolutional layer with the kernel size of $1 \times 1$ to produce the corresponding segmentation map $S_i$

$$S_i = \sigma(\text{conv}_{1\times1}(F^i))$$

where \text{conv}_{1\times1} denotes a convolutional layer with the kernel size of $1 \times 1$, and $\sigma(\cdot)$ is the sigmoid activation. Our network produces five side-output maps ($S_b, S_3, S_4, S_5, S_2$), where the output $S_2$ of the last BCSR block is used as the final infection segmentation mask.

B. AGGC Module

The clinical practice and data analysis have demonstrated that the pneumonia lesion areas, including the COVID-19,
have no specific distribution characteristics and are generally scattered. Moreover, lung opacities in the CT image are inhomogeneous, and there is no clear center or boundary. These undoubtedly pose great challenges to the segmentation model, including the following: 1) the scattered lesion areas can easily lead to missed detection or incomplete detection and 2) the boundary of the lesion area is not easy to segment accurately and sharply. To address these issues, we introduce more effective and valuable short-connection encoder as guidance. The evolution of encoder features is specifically manifested in two aspects: first, we use SA and BA to filter the features in terms of highlighting the important spatial locations and boundary details. Second, we capture the context dependence of different regions, which constrains the generation of more complete and accurate segmentation result. The overall architecture is illustrated in Fig. 3. The AGGC module can not only ensure the integrity of the features of the local lesion area, but also perceive the global correlation between different areas. The specific process is as follows.

Although the encoder features can provide rich details and other useful information for the decoder, it also has much redundancy. Therefore, the SA [35] is introduced to modify the encoder features from the spatial dimension, focusing on highlighting the important spatial positions. Specifically, we first apply the average-pooling and max-pooling operations along the channel axis on the input features and then concatenate them to generate an efficient feature descriptor. Finally, a convolution layer with the kernel size of $3 \times 3$ followed by a sigmoid function is used to generate an SA map $A_i$:

$$A_i = \sigma(\text{conv}_{3 \times 3}(\text{concat}(\text{avepool}(f^i), \text{maxpool}(f^i))))$$  

where avepool and maxpool denote the average pooling and max pooling along the channel axis, respectively. With the SA map, the initial encoder features $f^i$ can be refined as spatial-enhanced features $f'^i$ via residual connection

$$f'^i = A_i \odot f^i + f^i$$  

where $\odot$ represents element-wise multiplication. The visualization of spatial-enhanced features is shown in Fig. 4(c).

A convolution layer with the kernel size of $3 \times 3$ followed by a sigmoid function is used to generate an SA map $A_i$:

$$A_i = \sigma(\text{conv}_{3 \times 3}(\text{concat}(\text{avepool}(f^i), \text{maxpool}(f^i))))$$  

where avepool and maxpool denote the average pooling and max pooling along the channel axis, respectively. With the SA map, the initial encoder features $f^i$ can be refined as spatial-enhanced features $f'^i$ via residual connection

$$f'^i = A_i \odot f^i + f^i$$  

where $\odot$ represents element-wise multiplication. The visualization of spatial-enhanced features is shown in Fig. 4(c).

It can be seen that the spatial-enhanced features have a high response in the infected area to be segmented and a low response in the irrelevant background area, thus achieving the effect of highlighting important spatial locations and suppressing interferences.

As mentioned earlier, segmenting the lesion areas with clear and definite boundary is very important for diagnosis. To maintain the sharpness and clarity of the boundary in each decoding stage, we directly use the boundary map $S_b$ generated by the BA unit to emphasize the important boundary details. To highlight the boundary locations without losing other important information, we still use the residual connection for integration

$$f'^i = S_b \odot f'^i + f^i$$  

where $S_b$ represents the boundary-enhanced features, and $S_b$ is the boundary map generated by the BA unit.

In addition to highlighting the important location and boundary of the initial encoder features, we also need to solve the problem of incomplete detection caused by scattered lesions. Therefore, modeling the dependencies between different pixels is an effective solution. In this way, if an infected area is segmented, other scattered interference areas highly related to it may also be located. To this end, we introduce the GCM unit to achieve feature alignment and mutual enhancement between features by modeling the dependencies between different locations in the feature map. Following [36], we first learn the global context-aware feature map $G^i$ by measuring the influence of the embedding features of all other spatial locations on the current location:

$$G^i = \nabla(\Delta(\tilde{f}_b^i \otimes o^j))$$

$$= \nabla(\Delta(\tilde{f}_b^i) \otimes \text{Norm}((\Delta(\tilde{f}_b^i))^T \otimes \Delta(\tilde{f}_b^i))^T))$$  

where the values of $\tilde{f}_b^i$ are the normalized boundary-enhanced features, $\Delta(\cdot)$ reshapes a matrix of $\mathbb{R}^{D_1 \times D_2 \times D_1}$ into $\mathbb{R}^{D_1 \times D_2}$, $\nabla(\cdot)$ is the inverse operator of $\Delta(\cdot)$, $\otimes$ is the matrix multiplication, $o^j$ is the global context relationship map, and $\text{Norm}$ is the normalization operation. Then, $G^i$ that encodes the global context relationship is used to further refine the boundary-enhanced features in a residual connection manner

$$f'^i_{agg} = \zeta \cdot (G^i \odot f'^i) + f'^i$$

where $\zeta$ is a learnable weight parameter that controls the contribution of the global contextual information.

The entire AGGC model adopts a progressive enhancement structure, which not only highlights important spatial locations and boundary details, but also encodes the context dependencies between different spatial positions. The operation sequence of these three enhancement strategies is SA, BA, and global context. The reasons for this design are as follows: for encoder features with rich information, reducing redundancy is a primary task; otherwise, the continued existence of redundancy in the later learning process will weaken the expression of features. Then, to ensure that the boundary details are well preserved and highlighted, we apply the BA to spatial-enhanced features. Finally, calculating the global context dependence on relatively clear and distinct features after spatial and boundary enhancements can further strengthen the discrimination of features. This process can obtain encoder features with prominent spatial positions, highlighted boundary details, and comprehensive global context, which can then be used as the guidance for decoder features.

### C. BA Unit

For the medical image segmentation task, especially the COVID-19 segmentation where the lesion areas are scattered...
throughout the image, it is particularly important to clearly and accurately outline the boundaries of the lesion areas for later diagnosis. If only the semantic information of each pixel is considered and the boundary information is ignored in the segmentation process, this may cause the boundary of the segmentation result to be blurred and not sharp enough.

Inspired by [8], we introduce the boundary constraint into our approach by utilizing the BA unit, as shown in Fig. 5. On the one hand, we also introduce the boundary supervised learning to the network. In this way, the network can gradually learn the features of generating a clear boundary map. Specifically, the low-level features \( f^1 \) are fed into three convolution layers with different kernel sizes, and then, a sigmoid function is employed to generate the final BA map

\[
S_b = \sigma(\text{conv}_{1 \times 1}(\text{conv}_{3 \times 3}(\text{conv}_{1 \times 1}(f^1))))
\]

(8)

where the values of \( f^1 \) are the encoder features of the first layer, and \( \text{conv}_{n \times n} \) denotes the convolution layer with the kernel size of \( n \times n \). The learning process of BA map is a supervised learning manner, which can guarantee the effectiveness and pertinence of learning. As shown in Fig. 4(d), the BA map \( S_b \) effectively highlights the boundary positions of the infected regions to be segmented, which can be further used to refine the high-level features in the AGGC module.

In addition to generating boundary map through label supervised learning, we also treat the generated boundary map as an attention weight to highlight the encoder features as introduced in (5). Different from the work [8], we did not use the boundary features to perform feature enhancement. Instead, we adopt a method similar to SA, treat the generated boundary map as an attention map, and use the residual connection for feature enhancement. In this way, boundary constraints can be introduced more intuitively and efficiently, and at the same time, unnecessary information redundancy in the feature space can be reduced. This BA map will be used to enhance all the high-level features (i.e., \( f^2 \), \( f^3 \), and \( f^4 \)).

### D. SG Unit

The features learned by different encoder layers include different information. For example, the low-level features highlight the details of the scene, such as the texture and boundary, while the high-level features tend to learn the semantic and category attributes of the scene. For the segmentation task of COVID-19 lesion areas, there are relatively more background interferences, and how to effectively suppress them is particularly important. Considering the importance of high-level semantic information for suppressing irrelevant background interference, we design an SG unit to aggregate multiscale high-level features and formulate an SG map to refine the decoder features.

For the feature fusion of different scales, the correlation between feature maps of adjacent scales on the feature pyramid is the strongest. However, in the previous multiscale feature fusion methods, some features are usually upsampled or downsampled multiple times, which will lose the semantic information of the original features and even cause semantic ambiguity. Based on this, we try our best to avoid the worthless and unadvisable multiple cross-scale sampling of single layer. As illustrated in Fig. 2, the spatial resolution of high-level features \( f^2 \), \( f^3 \), and \( f^4 \) is unified on the intermediate resolution of the third encoder layer, which not only avoids the information loss due to multiple sampling, but also preserves the correlation between adjacent-scale features to the utmost extent. To be more specific, the encoder features \( f^2 \) and \( f^4 \) are \( \times 2 \) downsampled and \( \times 2 \) upsampled to the resolution of features \( f^3 \), respectively, and then, they are fused together to obtain an SG map through the addition and convolution operations. The process can be formulated as follows:

\[
S_s = \text{conv}(\text{down}_2(f^2)) + f^3 + \text{up}_2(f^4)
\]

(9)

where \( \text{up}_2 \) and \( \text{down}_2 \) are the \( \times 2 \) spatial upsampling and downsampling, respectively. The learned semantic map \( S_s \) encodes the multilevel semantic information, which is served as a global guidance in BCSR block. Since we choose the middle layer as the scale standard of feature output, the semantic map \( S_s \) only needs to be downsampled once in the first BCSR block and upsampled once in the last BCSR block, respectively. It effectively avoids the multiple cross-level sampling of \( S_s \) and retains the feature information of \( S_s \) to the maximum extent.

### E. Loss Function

For the supervision of network, the overall loss function involves two parts: first, the segmentation loss on three decoder outputs (i.e., \( S_2 \), \( S_3 \), and \( S_4 \)) and the semantic loss on the semantic map \( S_s \) via weighted intersection of union (IOU) loss and weighted binary cross entropy (BCE) loss. Inspired by [37], the weighted IOU loss provides effective global (image level) supervision, and the weighted BCE loss provides effective local (pixel level) supervision. Combining them, accurate segmentation result can be obtained from the image level and pixel level, which is defined as follows:

\[
l_s = \sum_{k=S_2,S_3,S_4} (l_{\text{wio}}^k + l_{\text{wbce}}^k)
\]

(10)

\[
l_{\text{wbce}}^k = \sum_{x=1}^{u} \sum_{y=1}^{h} Gxy \log \left( \frac{P_{xy}^k}{S_{xy}} \right) * (1 + \epsilon_{xy})
- \sum_{x=1}^{u} \sum_{y=1}^{h} (1 - Gxy) \log \left( 1 - \frac{P_{xy}^k}{1 + \epsilon_{xy}} \right) * (1 + \epsilon_{xy})
\]

\[
\sum_{x=1}^{u} \sum_{y=1}^{h} \epsilon_{xy}
\]

(11)

\[
l_{\text{wio}}^k = 1 - \frac{\sum_{x=1}^{u} \sum_{y=1}^{h} (P_{xy}^k \cap G) * (1 + \epsilon_{xy})}{\sum_{x=1}^{u} \sum_{y=1}^{h} (P_{xy}^k \cup G) * (1 + \epsilon_{xy})}
\]

(12)

where \( l_{\text{wio}}^k \) and \( l_{\text{wbce}}^k \) denote the weighted IOU loss and weighted BCE loss on different supervision types, respectively, \( P_{xy}^k \) is the prediction map (i.e., \( S_2 \), \( S_3 \), \( S_4 \), and \( S_s \)), \( G \) is the
segmentation ground truth, \(w\) and \(h\) are the width and height of the input image, \((x, y)\) denotes the coordinate of each pixel in image, and \(\epsilon\) is a weight hyperparameter. Second, the boundary loss \(l_{\text{bce}}^b\) on the BA map \(S_b\) via standard BCE loss, which is defined as follows:

\[
l_{\text{bce}}^b = \sum_{x=1}^{w} \sum_{y=1}^{h} -G_x^b \log (P_x^b) - (1 - G_x^b) \log (1 - P_x^b)
\]

(13)

where \(G_x^b\) is the boundary ground truth, and \(P_x^b\) is the predicted boundary map.

Therefore, the overall loss function can be defined as follows:

\[
l_{\text{total}} = l_s + l_{\text{bce}}^b.
\]

(14)

IV. EXPERIMENTS

A. Datasets and Evaluation Metrics

At present, there are few public COVID-19 lung CT datasets for infection segmentation. To have relatively sufficient samples for training, we merged the two public datasets [38], [39] to obtain 1018 high-quality CT images and further divided them into 719 training images and 300 testing images. For each CT slice, it contains the corresponding infection mask, and the infection edge obtained from infection mask using the Canny operator. Note that, the COVID-19 classification and the infection edge obtained from infection mask using each CT slice, it contains the corresponding infection mask, them into 719 training images and 300 testing images. For to obtain 1018 high-quality CT images and further divided ples for training, we merged the two public datasets [38], [39].

For these six measurements, in addition to the MAE score, all other indicators are that the larger the value, the better the performance.

B. Training Strategies and Implementation Details

We implement the proposed network via the PyTorch toolbox and are accelerated by an NVIDIA GeForce RTX 2080Ti GPU. We also implement our network using the MindSpore Lite tool. The Res2Net-50 [34] pretrained on ImageNet [47] is employed as the backbone feature extractor in the experiment. We have two points in particular to note: first, the fundamental reason we choose Res2Net-50 as the backbone network is for fair comparison, so we follow the MindSpore Lite tool. The Res2Net-50 [34] pretrained on ImageNet [47] is employed as the backbone feature extractor in the experiment. We have two points in particular to note: first, the fundamental reason we choose Res2Net-50 as the backbone network is for fair comparison, so we follow the MindSpore Lite tool. The Res2Net-50 [34] pretrained on ImageNet [47] is employed as the backbone feature extractor in the experiment. We have two points in particular to note: first, the fundamental reason we choose Res2Net-50 as the backbone network is for fair comparison, so we follow the MindSpore Lite tool. The Res2Net-50 [34] pretrained on ImageNet [47] is employed as the backbone feature extractor in the experiment. We have two points in particular to note: first, the fundamental reason we choose Res2Net-50 as the backbone network is for fair comparison, so we follow the MindSpore Lite tool. The Res2Net-50 [34] pretrained on ImageNet [47] is employed as the backbone feature extractor in the experiment. We have two points in particular to note: first, the fundamental reason we choose Res2Net-50 as the backbone network is for fair comparison, so we follow the MindSpore Lite tool. The Res2Net-50 [34] pretrained on ImageNet [47] is employed as the backbone feature extractor in the experiment. We have two points in particular to note: first, the fundamental reason we choose Res2Net-50 as the backbone network is for fair comparison, so we follow the MindSpore Lite tool. The Res2Net-50 [34] pretrained on ImageNet [47] is employed as the backbone feature extractor in the experiment. We have two points in particular to note: first, the fundamental reason we choose Res2Net-50 as the backbone network is for fair comparison, so we follow the MindSpore Lite tool. The Res2Net-50 [34] pretrained on ImageNet [47] is employed as the backbone feature extractor in the experiment.

References

[38] Cong, M., Cong, Y., Cong, H., Cong, G., & Cong, J. (2021). BCS-Net: BOUNDARY, CONTEXT, AND SEMANTIC FOR AUTOMATIC COVID-19 LUNG INFECTION SEGMENTATION. IEEE Access, 9, 50190-50199.
are gray scale, and when we load the CT image, we still use the RGB three-channel method, but the corresponding pixel values in the three channels are the same. To computational efficiency, we resize all the training and testing images to $352 \times 352$. The batch size is 8, and the training process converges until 200 iterations. We use Adam [49] as a learning rate optimizer with betas of $(0.9, 0.999)$, and the learning rate is set to $3e^{-4}$. The training loss curve of our network is shown in Fig. 6. The average inference time of our method is 0.036 s (27 frames/s) for processing an image with the size of $352 \times 352$, the Giga floating point operations (GFLOPs) is 28.02, and the parameters are 44, 822, and 580. The code and results of our BCS-Net can be found from the link of https://github.com/rmcong/BCS-Net-TIM22.

C. Comparison With State-of-the-Art Method

We compare our method with some state-of-the-art methods, including five classical segmentation models (UNet [33], UNet++ [44], Attention-UNet [45], Resnet34-UNet [33], context encoder network (CE-Net) [46], and FCN [32]) and three state-of-the-art segmentation methods for COVID-19 (Inf-Net [8], AnamNet [9], and CopleNet [10]). To ensure the fairness of the experiment, all the comparison methods are retrained on the same dataset under the default parameter settings.

1) Qualitative Comparison: Fig. 7 shows the qualitative comparisons of our model with the other state-of-the-art methods. It can be seen that the results of our method have greater advantages in terms of detection accuracy, completeness, and sharpness. For example, in the first image, the general medical image segmentation networks (e.g., UNet [33], UNet++ [44], Attention-UNet [45], and Resnet34-UNet [33]) often cannot effectively suppress the interference of background regions, such as the areas between the two lungs. In contrast, the detection results of the COVID-19 segmentation network are better, but the existing methods (e.g., Inf-Net [8], AnamNet [9], and CopleNet [10]) cannot completely suppress these interferences. For example, the area under the left lung is not effectively suppressed. Our proposed method has a better performance in these aspects and has a stronger ability to detect details. The last image also confirms that our proposed method can accurately detect infected areas and suppress irrelevant background areas. In addition, our method has more complete structure and sharper boundaries. In the third image, the existing methods (e.g., Inf-Net [8], AnamNet [9], and CopleNet [10]) cannot completely and continuously detect the infected regions at
the bottom of the left lung, while our method can detect them clearly, accurately, and completely. Moreover, compared with the existing methods (e.g., UNet++ [44], CopleNet [10], and Attention-UNet [45]), our results have clearer boundaries, which are very important for doctors to diagnose and treat. In general, our method exhibits competitive visual effects in both the segmentation ability of the infected regions and the suppression ability of the background regions.

2) Quantitative Comparison: The quantitative comparisons are reported in Table I. In addition to the precision score, our proposed BCS-Net achieves the best performance in other five measurements on the testing dataset. Compared with the classical segmentation models (e.g., UNet [33], UNet++ [44], Attention-UNet [45], and Resnet34-UNet [33]), the well-designed methods for COVID-19 are more competitive. For example, compared with the UNet [33], the percentage gain of the Inf-Net [8] reaches 6.6% for DSC score, and the percentage gain of our method is 8.6%. For $E_{\phi}$, the performance improvement is more obvious. Specifically, the performance of Inf-Net [8] is increased by 5.7% compared with the UNet++ [44], and the performance is increased by 9.0% compared with the Resnet34-UNet [33]. For the infection segmentation method for COVID-19, our proposed BCS-Net achieves better performance. For example, compared with the second best method, the percentage gain of DSC is 1.9%, the Recall score is 1.8%, and the $E_{\phi}$ score achieves 1.0%. In general, our detection effect is superior in quantitative measurements.

D. Ablation Study

We conduct several ablation experiments to validate the performance and effectiveness of two contributed components of our proposed model, including AGGC module and SG unit. The results are given in Table II and Fig. 8.

First, we verify the effect of AGGC module on the whole network. The main function of the AGGC module is to strengthen the important spatial position information and correlate the dependencies of different pixels, thereby ensuring a more accurate and complete detection. We remove the AGGC module, including the guidance of SA and BA, and the global context awareness. In other words, in the BCSR block, the decoding features of the previous layer will not enter the AGGC module but will be passed directly backwards. Compared with No. 1 and No. 2 in Table II, without the AGGC module, the five indicators fell, and the MAE score remained flat. As shown in Fig. 8(d), after removing the AGGC module, there are unsuppressed interference noises in the top-left corner, top-right corner, and bottom-right corner of the first image, and the boundaries are blurred. In the second image, the small lesions on the top-right after removing the AGGC module are not detected, while in the third image, there are obvious false detections in the left region. These all demonstrate the effectiveness of our AGGC module.

Second, to discuss the role of the SG unit, we design two ablation experiments. One is that we directly delete the SG unit, that is, without introducing global semantic information guidance. The other is that we replace the proposed SG unit with the PPD module in Inf-Net [8]. The related results are reported in No. 3 and No. 4 of Table II. Compared with the full model, all indicators will be reduced after removing or replacing the SG unit. In addition, we also found that the complex PPD model has three indicators (i.e., DSC, Prec, and Recall) that it is not as good as removing the SG unit directly. Comparing the SG and PPD modules of introducing global SG (i.e., No. 1 and No. 4), we can see that the designed SG model achieves better performance. After removing the SG module, the top-left and top-right regions in the first image of Fig. 8(e), the top-left region of the second image, and the top-right region of the third image have some irrelevant interferences that are not effectively suppressed. Replacing the SG module with the PPD module also does not improve the performance, and both the left region and the top-right background region in the third image are incorrectly detected as infected regions.

### Table I

| Methods       | DSC ↑ | Prec ↑ | Recall ↑ | $S_m$ ↑ | $E_{\phi}$ ↑ | MAE ↓ |
|---------------|-------|--------|----------|---------|--------------|-------|
| UNet          | 0.777 | 0.804  | 0.814    | 0.862   | 0.917        | 0.020 |
| UNet++        | 0.771 | 0.836  | 0.780    | 0.867   | 0.906        | 0.021 |
| Attention-UNet| 0.746 | 0.818  | 0.768    | 0.853   | 0.913        | 0.021 |
| Resnet34 UNet | 0.720 | 0.702  | 0.836    | 0.812   | 0.873        | 0.030 |
| FCN           | 0.800 | 0.839  | 0.791    | 0.855   | 0.949        | 0.020 |
| CE-Net        | 0.818 | 0.834  | 0.824    | 0.854   | 0.960        | 0.017 |
| AmanNet       | 0.775 | 0.831  | 0.776    | 0.856   | 0.920        | 0.021 |
| CopleNet      | 0.816 | 0.850  | 0.821    | 0.874   | 0.944        | 0.016 |
| Inf-Net       | 0.828 | 0.831  | 0.846    | 0.877   | 0.963        | 0.016 |
| BCS-Net [ours]| 0.844 | 0.841  | 0.861    | 0.880   | 0.972        | 0.015 |

### Table II

| No. | Variations    | DSC ↑ | Prec ↑ | Recall ↑ | $S_m$ ↑ | $E_{\phi}$ ↑ | MAE ↓ |
|-----|---------------|-------|--------|----------|---------|--------------|-------|
| 1   | BCS-Net       | 0.844 | 0.841  | 0.861    | 0.880   | 0.972        | 0.015 |
| 2   | w/o AGGC      | 0.838 | 0.835  | 0.856    | 0.874   | 0.970        | 0.015 |
| 3   | w/o SG        | 0.837 | 0.840  | 0.850    | 0.872   | 0.969        | 0.016 |
| 4   | w/o SG w/ PPD | 0.836 | 0.838  | 0.849    | 0.872   | 0.969        | 0.016 |

![Fig. 8. Visual comparison of BCS-Net variants equipped with different modules. (a) Images. (b) Ground truth. (c) BCS-Net. (d) w/o AGGC. (e) w/o SG. (f) w/o SG w/ PPD.](image-url)
TABLE III

| No. | Output | DSC ↑ | P-rec ↑ | Recall ↑ | S_M ↑ | E_0 ↑ | MAE ↓ |
|-----|--------|-------|---------|----------|-------|-------|-------|
| 1   | S_1    | 0.839 | 0.851   | 0.846    | 0.867 | 0.965 | 0.017 |
| 2   | S_3    | 0.802 | 0.776   | 0.855    | 0.853 | 0.954 | 0.017 |
| 3   | S_2[ours] | 0.844 | 0.841   | 0.861    | 0.880 | 0.972 | 0.015 |

Both ablation experiments illustrate the role of the SG module designed in this article.

The final output of the network is derived from features at the second decoding stage mainly based on the following two points. First, we design the AGGC module to make full use of the features of the first encoder layer (f^1) to supplement boundary information for high-level encoder features (f^2, f^3, and f^4). In implementation, the features f^1 are combined with f^2, f^3, and f^4 through the AGGC module, respectively. Under such a model framework, we do not perform the AGGC module in the last decoding layer, so the network does not set the output of S_1. Second, as we all know, generating the final map at a lower resolution from a higher decoding stage will consume less computational resources, which is beneficial to computational efficiency. Moreover, it has been experimentally verified that its performance will not be obviously degraded. Table III presents the final segmentation results obtained by different decoding layers. It can be found that, in general, S_2 achieves better performance than S_1 and S_3, which also illustrates the effectiveness of our setup. Based on these observations and consideration, the output from the second decoding layer is used as the final segmentation result.

V. Conclusion

This article focuses on COVID-19 lung infection segmentation task and proposes an end-to-end framework dubbed as BCS-Net to achieve this. The proposed network follows an encoder–decoder architecture equipped with BCSR blocks, which considers the boundary refinement, context modeling, and semantic constraint jointly. The AGGC module is designed to select the most valuable encoder features from the perspective of important spatial/boundary locations and context dependence. Moreover, we design a new SG unit to provide the SG and suppress the background noises by aggregating multiscale high-level features at the intermediate resolution. Extensive experiments and ablation studies demonstrate the effectiveness of the proposed BCS-Net architecture.

REFERENCES

[1] D. Mishra, S. Chaudhury, M. Sarkar, and A. S. Soin, “Ultrasound image segmentation: A deeply supervised network with attention to boundaries,” IEEE Trans. Biomed. Eng., vol. 66, no. 6, pp. 1637–1648, Jun. 2019.
[2] G. Soldati et al., “On lung ultrasound patterns specificity in the management of COVID-19 patients,” J. Ultrasound Med., vol. 39, no. 11, pp. 2283–2284, Nov. 2020.
[3] I. D. Apostolopoulos and T. Bessiana, “COVID-19: Automatic detection from X-ray images utilizing transfer learning with convolutional neural networks,” Phys. Eng. Sci. Med., vol. 43, no. 2, pp. 635–640, 2020.
[4] N. E. M. Khalifa, F. Smarandache, G. Manogaran, and M. Loey, “A study of the neurovasop psychic signification on deep transfer learning models: An experimental case on a limited COVID-19 chest X-ray dataset,” Cognit. Comput., pp. 1–10, Jan. 2021.
[5] A. Abbas, M. M. Abdelsamea, and M. M. Gaber, “Classification of COVID-19 in chest X-ray images using DeTraC deep convolutional neural network,” Int. J. Speech Technol., vol. 51, no. 2, pp. 854–864, Feb. 2021.
[6] T. Zhou, S. Canu, and S. Ruan, “An automatic COVID-19 CT segmentation network using spatial and channel attention mechanism,” 2020, arXiv:2004.06673.
[7] X. Wang, X. Deng, Q. Fu, Q. Zhou, and J. Feng, “A weakly-supervised framework for COVID-19 classification and lesion localization from chest CT,” IEEE Trans. Med. Imag., vol. 39, no. 8, pp. 2615–2625, May 2020.
[8] D. 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.
[9] N. Paluri et al., “Anam-Net: Anamorphic depth embedding-based lightweight CNN for segmentation of anomalies in COVID-19 chest CT images,” IEEE Trans. Neural Netw. Learn. Syst., vol. 32, no. 3, pp. 932–946, Mar. 2021.
[10] G. Wang et al., “A noise-robust framework for automatic segmentation of COVID-19 pneumonia lesions from CT images,” IEEE Trans. Med. Imag., vol. 39, no. 8, pp. 2633–2663, Aug. 2020.
[11] Z. Han et al., “Accurate screening of COVID-19 using attention-based deep 3D multiple instance learning,” IEEE Trans. Med. Imag., vol. 39, no. 8, pp. 2584–2594, Aug. 2020.
[12] K. Roy, D. Banik, D. Bhattacharjee, O. Krejcar, and C. Kollmann, “LwMLA-NET: A lightweight multi-level attention-based NETwork for segmentation of COVID-19 lungs abnormalities from CT images,” IEEE Trans. Instrum. Meas., vol. 71, pp. 1–13, 2022.
[13] X. Wang et al., “SSA-Net: Spatial self-attention network for COVID-19 pneumonia infection segmentation with semi-supervised few-shot learning,” Med. Image Anal., vol. 79, Jul. 2022, Art. no. 102459.

[14] C. Li et al., “ASIF-Net: Attention steered interweave fusion network for RGB-D salient object detection,” IEEE Trans. Cybern., vol. 51, no. 1, pp. 88–100, Jan. 2021.

[15] Q. Tang et al., “BridgeNet: A joint learning network of depth map super-resolution and monocular depth estimation,” in Proc. 29th ACM Int. Conf. Multimedia, Oct. 2021, pp. 2148–2157.

[16] C. Li, S. Anwar, J. Hou, R. Cong, C. Guo, and W. Ren, “Underwater image enhancement via medium transmission-guided multi-color space embedding,” IEEE Trans. Image Process., vol. 30, pp. 4985–5000, 2021.

[17] Z. Chen, R. Cong, Q. Xu, and Q. Huang, “DPANet: Depth potentiality-aware gated attention network for RGB-D salient object detection,” IEEE Trans. Image Process., vol. 30, pp. 7012–7024, 2021.

[18] R. Cong et al., “Global-and-local collaborative learning for co-salient object detection,” IEEE Trans. Cybern., early access, Jul. 22, 2022, doi: 10.1109/TCYB.2022.3169431.

[19] R. Cong, Y. Zhang, L. Fang, J. Li, Y. Zhao, and S. Kwong, “RRNet: Relational reasoning network with parallel multi-scale attention for salient object detection in optical remote sensing images,” IEEE Trans. Geosci. Remote Sens., vol. 60, pp. 1644–1658, 2022.

[20] X. Zhou et al., “Edge-guided recurrent positioning network for salient object detection in optical remote sensing images,” IEEE Trans. Cybern., early access, Apr. 13, 2022, doi: 10.1109/TCYB.2022.3163152.

[21] H. Wen et al., “Dynamic selective network for RGB-D salient object detection,” IEEE Trans. Image Process., vol. 30, pp. 9179–9192, 2021.

[22] Y. Mao, Q. Jiang, R. Cong, W. Gao, F. Shao, and S. Kwong, “Cross-modality fusion and progressive integration network for saliency prediction on stereoscopic 3D images,” IEEE Trans. Multimedia, vol. 24, pp. 2435–2448, 2022.

[23] C. Li, R. Cong, J. Hou, S. Zhang, Y. Qian, and S. Kwong, “Nested network with two-stream pyramid for salient object detection in optical remote sensing images,” IEEE Trans. Geosci. Remote Sens., vol. 57, no. 11, pp. 9156–9166, Nov. 2019.

[24] C. Guo, C. Li, J. Guo, R. Cong, H. Fu, and P. Han, “Hierarchical features driven residual learning for depth map super-resolution,” IEEE Trans. Image Process., vol. 28, no. 5, pp. 2545–2557, May 2019.

[25] Y. Huang, F. Zheng, R. Cong, W. Huang, M. R. Scott, and L. Shao, “MCMT-GAN: Multi-task coherent modality transferable GAN for 3D brain image synthesis,” IEEE Trans. Image Process., vol. 29, pp. 8187–8198, 2020.

[26] G. Yue, W. Han, B. Jiang, T. Zhou, R. Cong, and T. Wang, “Boundary constraint network with cross layer feature integration for polyp segmentation,” IEEE J. Biomed. Health Informat., early access, May 10, 2022, doi: 10.1109/JBHI.2022.3173948.

[27] A. Dairi, F. Harrou, and Y. Sun, “Deep generative learning-based 1-SVM detectors for unsupervised COVID-19 infection detection using blood tests,” IEEE Trans. Instrum. Meas., vol. 71, pp. 1–11, 2022.

[28] R. R. Sharma, M. Kumar, S. Maheshwari, and K. P. Ray, “EVDHM-ARIMA-based time series forecasting model and its application for COVID-19 cases,” IEEE Trans. Instrum. Meas., vol. 70, pp. 1–10, 2021.

[29] K. He, G. Gkioxari, P. Dollár, and R. B. Girshick, “Mask R-CNN,” in Proc. IEEE Int. Conf. Comput. Vis., Oct. 2017, pp. 2980–2988.

[30] D. Bolya, C. Zhou, F. Xiao, and Y. J. Lee, “YOLACT: Real-time instance segmentation,” in Proc. IEEE/CVF Int. Conf. Comput. Vis., Oct. 2019, pp. 9156–9165.

[31] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, “Encoder-decoder with atrous separable convolution for semantic image segmentation,” in Proc. Eur. Conf. Comput. Vis., 2018, pp. 833–851.

[32] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2015, pp. 3431–3440.

[33] O. Ronneberger, P. Fischer, and T. Brox, “U-Net: Convolutional networks for biomedical image segmentation,” in Proc. Int. Conf. Med. Image Comput. Assist. Interv., 2015, pp. 234–241.

[34] 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.

[35] S. Woo, J. Park, J. Y. Lee, and I. S. Kweon, “CBAM: Convolutional block attention module,” in Proc. Eur. Conf. Comput. Vis., 2018, pp. 3–19.

[36] Q. Zhang et al., “Dense attention fluid network for salient object detection in optical remote sensing images,” IEEE Trans. Image Process., vol. 30, pp. 1305–1317, 2021.

[37] J. Wei, S. Wang, and Q. Huang, “FE-Net: Fusion, feedback and focus for salient object detection,” in Proc. AAMT Conf. Artif. Intell., 2020, vol. 34, no. 7, pp. 12321–12328.

[38] COVID-19 CT Segmentation Dataset. Accessed: Apr. 2020. [Online]. Available: https://medisegmentation.com/COVID19/

[39] COVID-19 CT Lung and Infection Segmentation Dataset. Accessed: Apr. 20, 2020. [Online]. Available: https://zenodo.org/record/3757476.

[40] Y. Huang, F. Zheng, R. Cong, W. Gao, F. Shao, and S. Kwong, “Cross-modality fusion and progressive integration network for saliency prediction on stereoscopic 3D images,” IEEE Trans. Multimedia, vol. 24, pp. 2435–2448, 2022.

[41] R. Cong et al., “Edge-guided recurrent positioning network for salient object detection in optical remote sensing images,” IEEE Trans. Geosci. Remote Sens., vol. 57, no. 11, pp. 9156–9166, Nov. 2019.

[42] R. Cong et al., “SSA-Net: Spatial self-attention network for COVID-19 pneumonia infection segmentation with semi-supervised few-shot learning,” Med. Image Anal., vol. 79, Jul. 2022, Art. no. 102459.