Received: 19 July 2021 | Revised: 16 February 2022 | Accepted: 21 February 2022

DOI: 10.1002/mp.15600

RESEARCH ARTICLE

MEDICAL PHYSICS

HFCF-Net: A hybrid-feature cross fusion network for COVID-19 lesion segmentation from CT volumetric images

Yanting Wang[1] | **Qingyu Yang**[1] | **Lixia Tian**[1] | **Xuezhong Zhou**[1] | **Islem Rekik**[2,3] | **Huifang Huang**[1]

[1]School of Computer and Information Technology, Beijing Jiaotong University, Beijing, China

[2]BASIRA Laboratory, Faculty of Computer and Informatics, Istanbul Technical University, Istanbul, Turkey

[3]School of Science and Engineering, Computing, University of Dundee, Dundee, UK

Correspondence
Huifang Huang, School of Computer and Information Technology, Beijing Jiaotong University, Beijing 100044, China.
Email: hfhuang@bjtu.edu.cn; huifangbj@hotmail.com

Funding information
Natural Science Foundation of Beijing, Grant/Award Number: M21012

Abstract

Background: The coronavirus disease 2019 (COVID-19) spreads rapidly across the globe, seriously threatening the health of people all over the world. To reduce the diagnostic pressure of front-line doctors, an accurate and automatic lesion segmentation method is highly desirable in clinic practice.

Purpose: Many proposed two-dimensional (2D) methods for sliced-based lesion segmentation cannot take full advantage of spatial information in the three-dimensional (3D) volume data, resulting in limited segmentation performance. Three-dimensional methods can utilize the spatial information but suffer from long training time and slow convergence speed. To solve these problems, we propose an end-to-end hybrid-feature cross fusion network (HFCF-Net) to fuse the 2D and 3D features at three scales for the accurate segmentation of COVID-19 lesions.

Methods: The proposed HFCF-Net incorporates 2D and 3D subnets to extract features within and between slices effectively. Then the cross fusion module is designed to bridge 2D and 3D decoders at the same scale to fuse both types of features. The module consists of three cross fusion blocks, each of which contains a prior fusion path and a context fusion path to jointly learn better lesion representations. The former aims to explicitly provide the 3D subnet with lesion-related prior knowledge, and the latter utilizes the 3D context information as the attention guidance of the 2D subnet, which promotes the precise segmentation of the lesion regions. Furthermore, we explore an imbalance-robust adaptive learning loss function that includes image-level loss and pixel-level loss to tackle the problems caused by the apparent imbalance between the proportions of the lesion and non-lesion voxels, providing a learning strategy to dynamically adjust the learning focus between 2D and 3D branches during the training process for effective supervision.

Result: Extensive experiments conducted on a publicly available dataset demonstrate that the proposed segmentation network significantly outperforms some state-of-the-art methods for the COVID-19 lesion segmentation, yielding a Dice similarity coefficient of 74.85%. The visual comparison of segmentation performance also proves the superiority of the proposed network in segmenting different-sized lesions.

Conclusions: In this paper, we propose a novel HFCF-Net for rapid and accurate COVID-19 lesion segmentation from chest computed tomography volume data. It innovatively fuses hybrid features in a cross manner for lesion segmentation, aiming to utilize the advantages of 2D and 3D subnets to complement each other for enhancing the segmentation performance. Benefitting from the cross fusion mechanism, the proposed HFCF-Net can segment the lesions more accurately with the knowledge acquired from both subnets.

KEYWORDS
COVID-19, cross fusion, hybrid-feature fusion, lesion segmentation

1 | INTRODUCTION

Coronavirus disease 2019 (COVID-19) caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), was first discovered in December 2019 and then rapidly spread to many countries around the world in early 2020.[1–3] Since the coronavirus can be spread by droplets, aerosols, and other methods,[4] the COVID-19 is highly infectious. According to statistics from the Center for Systems Science and Engineering of Johns Hopkins University, as of December 21, 2020, 191 countries and regions all over the world have reported a total of more than 76 million confirmed cases and about 1.69 million patients have died from coronavirus.[5]

The typical clinical manifestations of COVID-19 patients mainly include fever, cough, shortness of breath, loss of taste, and smell, usually accompanied by pneumonia.[1,6] Presently, real-time reverse transcription polymerase chain reaction (RT-PCR) is the mainstream method for diagnosing coronavirus pneumonia, but it has the disadvantages of time-consuming detection and a high false-negative rate in the early stages of the disease.[7] In contrast, computed tomography (CT) imaging is relatively easy to perform rapid scanning. It can more directly observe the lesions of lungs with high-resolution three-dimensional (3D) images for disease diagnosis.[8] Some studies have also shown that CT imaging can serve as the supplement of RT-PCR detection to support early detection of COVID-19.[6,9,10] The Guidelines for the Diagnosis and Treatment of Pneumonia Caused by COVID-19 (fifth edition) issued by the Chinese government have included the clinical manifestations of CT images in the diagnostic criteria.

In the diagnosis process, it is essential for doctors to observe the visual representations of the anatomy provided by medical images.[11,12] However, manual observation is labor-intensive and time-consuming work. Research indicates that an experienced radiologist can only interpret about four chest CT scans per hour.[13] Meanwhile, the infected regions exhibit various manifestations, such as ground glass shadow and lung consolidation, also accompanied by irregular shapes and fuzzy boundaries caused by low contrast, which may further pose challenges for the precise lesion detection and aggravate the burden of doctors.[8] Therefore, a fast auto-segmentation computer-aided diagnosis tool of COVID-19 lesions is urgently needed in the clinic applications since accurate segmentation of lesion regions is of value not only in facilitating the diagnosis but also in assessing the severity and prognosis of the disease.[14]

In recent years, as a continuously developing emerging technology, deep learning has achieved remarkable results in many aspects of the medical field.[15] Some studies have applied deep learning networks to the segmentation of pneumonia lesions, resulting in better performance consequently. The current methods for COVID-19 lesion segmentation mainly include two-dimensional (2D)-based and 3D-based segmentation methods. The 2D-based segmentation methods explore 2D convolutional neural networks (CNNs) to predict the lesion region of each slice in CT volume data.[16–22] For example, Wang et al.[11] proposed a novel noise-robust learning framework based on self-ensembling of 2D CNN for slice-by-slice segmentation. Fan et al.[17] designed a semi-supervised lung infection segmentation deep network (Semi Inf-Net) for CT slices. Laradji et al.[18] trained a 2D weakly supervised CNN with the transformation-consistency constraints for increasing robustness. Yao et al.[19] presented an unsupervised pixel-level anomaly modeling framework with the 2D U-Net backbone. Although these deep networks for lesion segmentation in slices have achieved better results, 2D networks cannot leverage the inter-slice spatial information, which leads to limited performance improvement.

The 3D-based segmentation approaches can effectively exploit the 3D spatial information of segmented tissues to produce more accurate label maps. However, most of the current studies focused on lesion segmentation in slices, while very little attention has been paid to the segmentation of 3D infection regions in CT volume data.[23–25] There are two main reasons for this: (1) long training time and convergence difficulties; and (2) high computational cost and memory consumption. Under limited hardware resource, the image size of input data is often reduced to ensure the successful running of 3D networks, which unavoidably causes the loss of global information and affects segmentation performance. Therefore, it is still a challenging task to employ 3D networks to segment lesions of COVID-19.

To address the above issues, we propose a novel end-to-end hybrid-feature cross fusion network (HFCF-Net) for the COVID-19 lesion segmentation of CT volume data. The proposed network integrates 2D and 3D subnets, which can effectively extract features within and between slices, and then uses the cross feature fusion to enhance the interaction of both subnets. Specifically, we first split the 3D data into 2D slice sequences and input them to 2D and 3D branches to be processed, respectively. Next, considering the complementarity between features of two subnets that can be exploited to enhance the useful information of lesion segmentation, the two output maps from each layer of 2D and 3D decoders are regarded as complementary guidance information and transferred to their own complementary decoder in

turn, aiming to boost the ability of the entire network to perceive lesion information. Fusing 2D features can utilize the fast convergence characteristics of the 2D subnet to help the 3D subnet reduce optimization burden. Meanwhile, with the guidance of the contextual information extracted by the 3D subnet, the segmentation performance of the 2D subnet can be well improved. Moreover, within an end-to-end system, both branches are jointly optimized during the training process to achieve the precise location and segmentation of infection regions. The proposed HFCF-Net has been evaluated on a publicly available dataset and achieved superior performance compared with the state-of-the-art segmentation networks.

In summary, the main contributions of this paper are as follows:

1. We propose a novel end-to-end HFCF-Net to achieve the lesion segmentation of COVID-19. The proposed network designs a cross fusion module to take advantage of 2D and 3D subnets to complement each other, avoiding the problems of missing inter-slice information of the 2D network and slow convergence of the 3D network. To our best knowledge, there has been no work investigating the hybrid-feature (intra- and inter-slice features) fusion in COVID-19 segmentation so far.
2. We design a novel network consisting of a 2D multi-scale subnet and a 3D lightweight subnet to effectively probe the intra- and inter-slice features for lesion segmentation. Specifically, the 2D subnet incorporates an aggregate interaction module and a Res2Net global module to improve the segmentation quality, and the 3D subnet adopts a lightweight design to reduce computation without information loss.
3. We propose a novel imbalance-robust adaptive learning loss function with an adaptive learning strategy. The proposed loss function not only supervises the whole network from both image- and pixel-level aspects to alleviate the imbalance of positive and negative examples but also transfers the learning emphasis of the network adaptively to avoid interference from subnets, thus obtaining a better optimization result.

The rest of this paper is organized as follows. Section 2 introduces details of the proposed HFCF-Net. The experiments and results are given in Section 3. Finally, we further present our discussion and conclusion in Sections 4 and 5, respectively.

2 | MATERIALS AND METHODS

The proposed HFCF-Net for lesion segmentation is shown in Figure 1 and summarized as follows: (1) decompose the 3D volume data to 2D slice sequences and then input them into the 2D multi-scale subnet to extract lesion features within slices; (2) input the 3D volume data into the 3D lightweight subnet to extract spatial lesion features; (3) feed features of both 2D and 3D decoders to the cross fusion module in turn for integrating 2D and 3D information to improve the lesion segmentation.

In this section, we firstly describe three modules in our network, including the 2D multi-scale subnet, the 3D lightweight subnet, and the cross fusion module. Next, we present a novel loss function with an adaptive learning strategy.

2.1 | The design of 2D multi-scale subnet

The layout of the proposed 2D subnet is given in Figure 1. We employ the "encoder–decoder" structure similar to the U-Net[26] but strengthen it in two aspects. Firstly, different from the original network that directly concatenates features from the encoder and decoder, we utilize the aggregate interaction module (AIM)[27] to integrate the multi-scale semantic features from neighboring encoder stages to replace skip connection to lessen the gap between the current encoder stage and symmetrical decoder stage, contributing to a better fusion process. Secondly, we introduce the Res2Net module[28] in the high-level stages of the decoder, where the module evenly splits the input feature maps into four subsets along the channel dimension and sends all the subsets except for the first one to the hierarchical residual-like connected convolutional operators. Then, the first feature subset and other convolutional outputs are concatenated to fuse information altogether. This aims to implicitly explore intra-stage multi-scale feature representations to generate abundant scale-specific information for further refining the segmentation results.

In addition, we replace the typical deconvolution layers with a 1×1 convolution layer followed by a bilinear interpolation to recover the feature resolution, which can avoid the grid effect and decrease the parameter number. As a result, the 2D subnet can efficiently capture multi-scale information embedded in the slices to enhance the representation ability of features for achieving better segmentation performance.

2.2 | The design of 3D lightweight subnet

Generally, 3D networks often suffer from high computational cost and Graphics Processing Unit (GPU) consumption caused by a large number of network parameters, which limits the depth of networks that is crucial for performance gains.[29] To make full use of the spatial context information under the limited hardware conditions for boosting the segmentation performance, we design a 3D lightweight

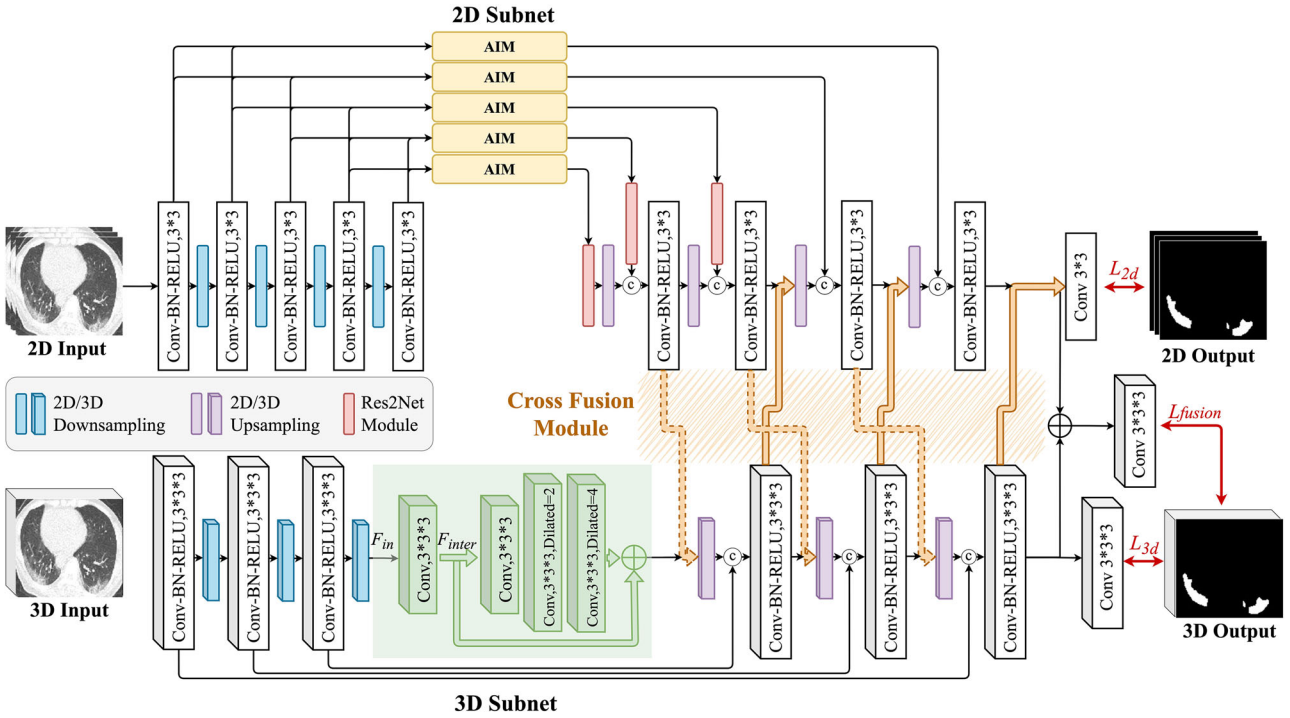

FIGURE 1 The overall architecture of the proposed hybrid-feature cross fusion network (HFCF-Net). It contains a 2D multi-scale subnet, a 3D lightweight subnet, and a cross fusion module

subnet with low computational cost to further explore the segmentation of the original 3D data. The proposed 3D lightweight subnet is shown in Figure 1.

Inspired by the great performance of 3D U-Net, we also use a similar backbone to design a lightweight backbone with two important aspects. *First*, considering the limitation of memory and model complexity, we explore a lightweight 3D U-Net by reducing the number of convolution kernels. Compared with the original 3D U-Net where the number of filters is doubled layer by layer, we adjust the filter setting, keeping the number of filters invariable in the shallow stages and halving the number of filters in the deep stages to avoid the degradation of high-resolution information. Consequently, the number of filters for each stage changes from the original setting, that is, {32, 64, 128, 256, 512} to the new setting, that is, {32, 64, 64, 128, 256}.

Second, since reducing the number of filters may affect the extraction of global contextual information, we design a dilated residual (DilRes) block at the bottleneck layer of the encoder–decoder structure to attenuate the global information loss. Most studies used a simple downsampling operation on the feature maps to get a wider receptive field for enhancing the global information representation. However, many stacked pooling layers may make the high-level feature maps too small to reserve the segmentation target for generating the global guidance information. Compared with the downsampling operation, dilated convolutions can expand receptive field without losing resolution through inserting holes into the familiar convolution operation.[30] It not only obtains a wider receptive field by setting the dilated rate to capture the global context but also maintains the relative spatial position of the feature maps required for precise segmentation. Specially, the proposed DilRes block mainly consists of two components. One component is a $3 \times 3 \times 3$ plain convolution that transforms input feature maps F_{in} to intermediate feature maps F_{inter} to capture local features. The other component is a successive dilated convolution operation with a residual connection that takes F_{inter} as input and extracts the long-range contextual information. Three dilated convolutions with different dilation rates (1, 2, and 4) are stacked to expand the receptive field further, and a residual connection is used to fuse F_{inter} with the output feature maps obtained through three successive dilated convolutions to facilitate the training.

Since the DilRes block is built upon the top-level feature maps with relatively low resolution, it cannot bring additional computational overheads. Furthermore, all intermediate feature maps inside the dilated convolution block have the same resolution as input feature maps F_{in} due to without using pooling operations, not causing too much information loss.

2.3 | The design of cross fusion module

Although 3D networks can exploit rich spatial information along the z dimension, there exists an apparent

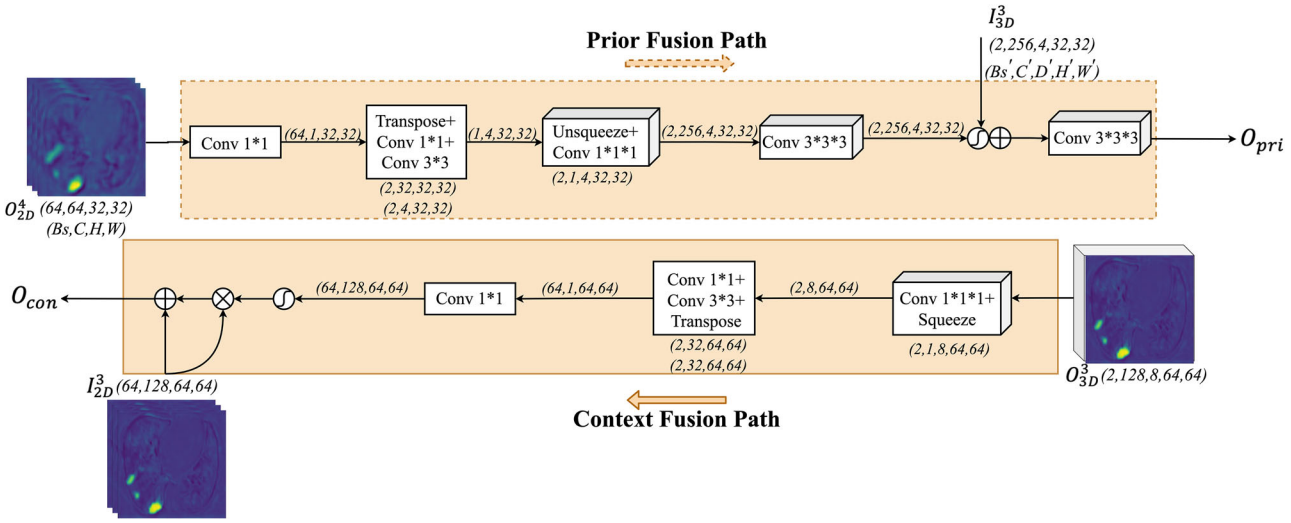

FIGURE 2 Illustration of the cross fusion block (CFB). Take the CFB[1] as an example. In the prior fusion path, the 2D feature maps $O_{D^4_{2D}}$ is transmitted to the 3D decoder layer and fused with the 3D feature maps $I_{D^3_{3D}}$. In the context fusion path, the 3D feature maps $O_{D^3_{3D}}$ is transmitted to the 2D decoder layer and fused with the corresponding 2D feature maps $I_{D^3_{2D}}$. Bs, C, H, and W: the batch size, channel number, height, and width of 2D feature maps; Bs′, C', D', H', and W': the batch size, channel number, depth, height, and width of 3D feature maps

imbalance between the voxel numbers of the lesion and non-lesion regions of a 3D image. This can cause the loss of negative samples (non-target) to dominate the total loss, which provides misleading information and overwhelms the positive feedback of the target, resulting in the inefficient training process.[31] In contrast, the imbalance problem inside the slices is not as severe as that of 3D images so that 2D networks spend less training time and converge faster than 3D networks, while 2D networks neglect useful spatial information to prevent the further improvement of its performance. To address these dilemmas, we propose fusing 3D and 2D subnets to jointly utilize their own characteristics to complement each other for obtaining better lesion segmentation maps.

We design a cross fusion module to fuse both types of features at the same scale for producing more discriminative fusion feature maps. Figure 1 shows the structure of the cross fusion module that consists of three fusion blocks to bridge each layer of decoders from 2D and 3D subnets for obtaining better ability of perceiving lesions and segmentation performance. Figure 2 shows the details of the fusion block, which contains two cross paths, that is, a prior fusion path and a context fusion path.

In the prior fusion path, we regard 2D feature maps as auxiliary features and fuse them with 3D feature maps at the same scale, aiming to utilize the intra-slice information to provide the 3D subnet with prior lesion knowledge and improve the efficiency of 3D feature extraction. In the meantime, the 3D feature maps can be used as a kind of context attention to boost the ability of the 2D subnet in perceiving the context information. To this end, the context fusion path is proposed to combine the context information captured by the 3D feature maps as auxiliary knowledge with the feature maps of 2D decoder, providing the context attention guidance and enhancing the segmentation precision of the 2D subnet.

We define the ith cross fusion block as CFB^i. I^i_{2D} and O^i_{2D} denote the input and output feature maps of the ith decoder layer ($i = 1, 2, 3, 4$ from shallow layer to deep layer) in the 2D subnet, respectively. Similarly, I^i_{3D} and O^i_{3D} ($i = 1, 2, 3$) are the input and output feature maps of the ith decoder layer in the 3D subnet, respectively.

As shown in Figure 2, for the first fusion block CFB^1, the output feature maps O^4_{2D} from the 2D subnet and the input feature maps I^3_{3D} from the 3D subnet are the inputs of prior fusion path. We expect that incorporating the intra-slice feature maps with prior lesion information would promote the exploration efficiency of the 3D subnet. However, the 2D feature maps O^4_{2D} and the 3D feature maps I^3_{3D} have different sizes. Hence, the dimension transformation is performed to make the 2D feature maps O^4_{2D} have the same size as I^3_{3D}, and then the feature maps would pass through a convolution layer to refine weights, shown as follows

$$F = \mathrm{Conv}\left(\mathrm{Tran}\left(O^4_{2D}\right)\right) \tag{1}$$

where "Tran" represents dimension transformation operations, including channel adjustment and matrix transposition, and "Conv" indicates a convolution operation. The obtained feature maps F has the same size as the feature maps I^3_{3D}. Then we employ sigmoid activation functions to get the intra-slice attention maps and fuse them with I^3_{3D} to provide the coarse lesion location by

highlighting potential infection regions, shown as

$$O_{\text{pri}} = \text{Conv}\left(\text{sigmoid}\,(F) + I^3_{\text{3D}}\right). \tag{2}$$

In the context fusion path, the output feature maps O^3_{3D} and the input feature maps I^3_{2D} are the inputs of this path. The dimension of O^3_{3D} is also adjusted to the same size as the feature maps I^3_{2D} by the dimension transformation. Then we apply a sigmoid activation function to each slice of the transformed feature maps $\text{Tran}(O^3_{\text{3D}})$ to obtain attention maps, which accentuate lesion regions and reduce the response of interference regions for the feature maps I^3_{2D}. Subsequently, the resulted feature maps are fused with the input feature maps I^3_{2D} to retain intra-slice context information but highlight the lesion regions using the inter-slice lesion context information. These operations are formally shown as follows:

$$O_{\text{con}} = I^3_{\text{2D}} + I^3_{\text{2D}} \otimes \text{sigmoid}\left(\text{Tran}\left(O^3_{\text{3D}}\right)\right), \tag{3}$$

where $\otimes$ represents the element-wise multiplication.

In this way, we obtained the 3D features O_{pri} combined with prior information and the 2D features O_{con} fused with 3D context information. Then they would be fed to the follow-up layers to perform feature extraction.

The whole network was trained based on contextual information from the original 3D data and sufficient feature representations within slices from the 2D branch. At the same time, cross fusion can also enable the network to fuse 2D and 3D information in turn, avoiding the attenuation of 2D or 3D features caused by information fusion in single direction. With the guidance of the fusion information generated by efficient interaction, the problem of spatial information missing in the 2D subnet and the imbalance problem existing in the 3D subnet have also been well mitigated, which can gradually boost the segmentation performance with the increase of iteration times.

2.4 | Imbalance-robust adaptive learning loss

The total loss function proposed for training our HFCF-Net includes the loss $L_{\text{2D}-\text{GT}}$ between the 2D prediction maps and the 2D ground truth ones, the loss $L_{\text{3D}-\text{GT}}$ between the 3D prediction maps and the 3D ground truth ones, and the consistency loss $L_{\text{Fusion}-\text{GT}}$ between the fusion maps and the 3D ground truth ones. The three losses adopt the same form of loss function. For more efficient training, we design the loss function from two innovative perspectives. First, the proposed loss function combines the binary cross-entropy loss (BCE) L_{bce} and reweighted Dice loss $L_{\text{reweighted}-\text{Dice}}$ to make use of their advantages to alleviate the imbalance phenomenon existing in the data. Second, we utilize an adaptive learning strategy to adjust the learning attention between 2D and 3D subnets in the training process. Our motivation lies in the fact that the well-trained 2D subnet can provide prior information of lesion location for the 3D subnet, avoiding network degeneration caused by the wrong optimization direction.

Since the lesions often only occupy a small region of the lung, segmentation results tend to be strongly biased toward the background when the network is trained with the cross-entropy loss function. The Dice loss function can effectively tackle this problem by implicitly establishing a balance between foreground and background classes.[11] However, we also encounter an imbalance problem on the size of lesions. In some cases, the large lesions could be about 20–50 times bigger than the small ones, but the Dice loss treats equally lesions with different sizes, and the networks tend to miss small-sized targets. Shirokikh et al.[32] have proposed a loss function reweighting strategy to promote the detection quality by increasing the weight of small lesions. Inspired by the work of Shirokikh et al., we propose a novel reweighted Dice loss function for assigning the larger weights for small lesions. To address the twofold imbalance, the proposed loss function is composed of the global (image-level) loss and the local (pixel-level) loss, that is, reweighted Dice loss $L_{\text{reweighted}-\text{Dice}}$ and BCE loss L_{bce}

$$L = L_{\text{reweighted}-\text{Dice}} + L_{\text{bce}} = 1 - \frac{2\sum_i w_i p_i y_i}{\sum_i w_i \left(p_i^2 + y_i^2\right)} - \left(y_i \log p_i + (1 - y_i)\log\left(1 - p_i\right)\right), \tag{4}$$

where p_i denotes the ith pixel of the prediction probability map, y_i is the ith element of the corresponding ground truth binary mask, and w_i is the assigned weight.

The overall loss function of the network consists of three losses, that is, $L_{\text{2D}-\text{GT}}$, $L_{\text{3D}-\text{GT}}$, and $L_{\text{Fusion}-\text{GT}}$. The three losses are calculated according to the loss function in Equation (4). If 2D and 3D representation learning deserves equal attention, the two branches may interfere with each other to affect their own performance when they are trained together. Therefore, we adopt a learning strategy to transfer the learning focus between subnets adaptively by controlling the loss weights of the two subnets with parameter α, and the overall loss function is defined as

$$L_{\text{total}} = \alpha L_{\text{2D}-\text{GT}} + (2 - \alpha) L_{\text{3D}-\text{GT}} + L_{\text{Fusion}-\text{GT}}, \tag{5}$$

where $\alpha = 1 - (T/T_{\text{max}})^2$ is the adaptation factor, T_{max} is the total number of iterations, and T is the current iteration.

The total loss function is designed to first focus on training the 2D subnet and then gradually pay attention to training the 3D subnet, which is achieved by

the adaptation factor α. α automatically decreases with the increase of the training epochs, which means that the optimization focus of the total loss function transfers from 2D representation learning to 3D representation learning. In the initial stage of training, both subnets may have poor performance due to without sufficient feature fusion. However, due to the fast convergence of the 2D subnet, it first produces relatively better performance during the training process. With the proceeding of cross feature fusion, the 3D subnet can obtain reliable prior information about lesion location from 2D features and thus learn more discriminative representations about proper infection regions directly, which can significantly improve the training efficiency. During the process of training, the emphasis of the network transforms from intra-slice feature representation learning to exhaustive volumetric information exploration, thus improving the segmentation accuracy. Different from the conventional learning network, this strategy allows both subnets with different goals to be trained in a synergetic manner instead of performing a blind optimization process at the beginning to result in worse performance.

2.5 | Dataset and data preprocessing

We evaluated our network on a publicly accessible COVID-19-20 dataset,[33] which was provided by the Multi-National NIH Consortium for CT AI in COVID-19 via the NCI The Cancer Imaging Archive (TCIA) public website (https://www.cancerimagingarchive.net/)[34] for the COVID-19 Lung CT Lesion Segmentation Challenge-2020.[35] The dataset contained 199 cases for training and 50 cases for online validation. Since the challenge was over, we only got the ground truth annotations of the training set, but without true labels of online validation cases. Therefore, we conducted our experiments based on the 199 cases of training set. The size of CT scans was $512 \times 512 \times (39–361)$, and the voxel size was all $1 \times 1 \times 1$ mm^3. For efficient training, the 3D CT scans were preprocessed and then input to the network. For highlighting the anatomical structures and removing the irrelevant issues, we truncated the original intensity values of CT scans into [-1200 HU, 300 HU], which means to set the value above 300 to 300 and below -1200 to -1200. Then, the CT scans were further normalized to the standard normal distribution with z-score to avoid the influence of outliers.

In the training stage, we randomly cropped each 3D volume image into many patches with a size of $32 \times 256 \times 256$ as 3D input data, instead of using rough linear interpolation to resize the image, which can retain more information. In our experiments, the overall 199 cases were divided into the training and testing sets. The training set contained 150 CT scans, and the testing set had 49 scans. To evaluate the efficacy of our network compared to other networks, we adopted the fivefold cross-validation on the training set for adjusting hyper-parameters, and the results obtained on the testing set with the optimum value of hyper-parameters were compared.

3 | EXPERIMENTS AND RESULTS

3.1 | Experimental settings

3.1.1 | Implementation details

We implemented our HFCF-Net using Pytorch with an Nvidia Tesla T4 GPU. All experiments were performed on the same environment. During the training process, we first separately pre-trained the 2D and 3D subnets, and then the weights of the pre-trained networks were used to initialize 2D and 3D encoders of the proposed HFCF-Net. The other parts were initialized with the Kaiming initialization[36] that considers the nonlinearity of Rectified Linear Units (ReLU) to help with convergence of deep networks. All network weights of HFCF-Net were learned via the Adam optimizer with a weight decay of 0.0005. The batch size is set to 2. The initial learning rate was set to 0.0001 and decayed according to the polynomial schedule $\mathrm{lr} = \mathrm{lr} \times (1 - T/T_{\max})^{0.9}$. Moreover, we employed online data augmentation techniques, including random flipping and random rotating, to further alleviate the risk of overfitting. For a fair comparison, all compared networks were implemented on the same computer, and conducted the hyper-parameter optimization. The hyper-parameter values are listed in Table S2 and the optimization details can be found in Figures S3–S12.

3.1.2 | Evaluation metrics

For evaluating the validity of the proposed HFCF-Net, we chose the fusion output as the final prediction result S_{pre}. The similarity between the final prediction result S_{pre} and the ground truth G was quantified by five widely adopted metrics, that is, Dice similarity coefficient (Dice), intersection over union (IoU), the 95th percentile of Hausdorff distance (HD_{95}), sensitivity (Sen), and specificity (Spe).

The Dice and IoU are statistics used to gauge the similarity of two samples by calculating the ratio of the intersection area to the total area, computed as:

$$\mathrm{Dice} = \frac{2 \times \left| S_{\mathrm{pre}} \cap G \right|}{\left| S_{\mathrm{pre}} + |G| + \in \right|} \tag{6}$$

$$\mathrm{IoU} = \frac{\left| S_{\mathrm{pre}} \cap G \right|}{\left| S_{\mathrm{pre}} \cup G \right|} \tag{7}$$

where $\in$ is a smoothing factor to avoid zero denominator.

TABLE 1 Quantitative evaluation results of different segmentation networks

Network	Dice↑	IoU↑	HD_{95}↓	Sen↑	Spe↑	FLOPs (GFLOPs)	Param (M)	Training time (h)
V-Net (3D)[40]	0.6650	0.5112	8.4323	0.7054	0.9975	750.982	45.596	20
U-Net (3D)[39]	0.6858	0.5407	3.7606	0.8109	0.9975	1895.000	16.320	25
ConResNet (3D)[14]	0.7065	0.5697	3.6469	0.7860	0.9985	583.593	19.300	17.2
2.5D-Net[37]	0.6502	0.5012	3.5543	0.8254	0.9971	/	/	/
U-Net[26]	0.6124	0.4512	3.3811	0.7653	0.9967	40.081	17.266	3.5
U^2-Net[38]	0.5794	0.3949	4.3187	0.6299	0.9969	37.540	44.009	3
COPLE-Net[11]	0.6231	0.4590	3.2408	0.7359	0.9969	11.148	10.521	2.5
Inf-Net[17]	0.6304	0.4622	3.4295	0.6734	0.9975	6.381	30.337	2.5
Eff-Net[20]	0.6113	0.4451	4.6188	0.6328	0.9976	7.729	24.430	2
Weakly-Net[18]	0.5940	0.4286	5.6558	0.6227	0.9975	73.402	134.266	3
HFCF-Net	**0.7485**	**0.6068**	**3.1764**	**0.8358**	**0.9990**	798.374	33.670	18

Abbreviations: Dice, Dice similarity coefficient; FLOPs, floating point operations; HD_{95}, 95th percentile of Hausdorff distance; GFLOPs, giga floating point of operations; HFCF-Net, hybrid-feature cross fusion network; IoU, intersection over union; Sen, sensitivity; Spe, specificity; Param, parameter number.
The bold values means the best performance.

The HD evaluates the segmentation quality and a smaller value of the HD indicates better segmentation results. The HD is computed by the following expression:

$$\mathrm{HD} = \max\left\{\max_{s\in\bar{S}_{\mathrm{pre}}}\min_{g\in\bar{G}}\|s-g\|^2, \times \max_{g\in\bar{G}}\min_{s\in\bar{S}_{\mathrm{pre}}}\|g-s\|^2\right\} \quad (8)$$

where $\bar{S}_{\mathrm{pre}}$ and $\bar{G}$ denotes the set of lesion boundary points of the prediction result and the ground truth, respectively.

HD_{95} is similar to HD, and utilizes the 95th percentile of the distances instead of the maximal value in Equation (8). The purpose of using this metric is to eliminate the impact of outliers.

Besides, we also introduced other two metrics to measure the complexity of the network, that is, floating point operations (FLOPs) and parameter number (Param).

3.2 | Comparison of segmentation performance

To validate the efficacy of the proposed network in COVID-19 lesion segmentation, we considered nine state-of-the-art and classical networks for comparison, including: (1) COPLE-Net[11] that employs the squeeze-and-excitation block and Atrous Spatial Pyramid Pooling (ASPP) module to extract features and integrates a self-ensembling training framework to promote the robustness against noise; (2) Inf-Net[17] that uses reverse attention module to explore discriminative infection regions and adopts a parallel partial decoder to generate the global map; (3) a weakly supervised segmentation network (Weakly-Net)[18] based on spatial transformation consistency; (4) a modified lightweight U-Net with EfficientNetB7 backbone (Eff-Net)[20]; (5) 2.5D segmentation network (2.5D-Net)[37] that decomposes the 3D segmentation problem into three independent 2D segmentation problems; (6) two-level nested U-structure network (U^2-Net)[38]; (7) ConResNet[14] that designs the context residual module to explicitly perceive 3D context to boost the network's ability; (8) 2D U-Net[26]; (9) 3D U-Net[39]; (10) V-Net.[40] Note that 2D networks adopted the same training manner as the original papers did, that is, using only slices with lesions for training. In the testing phase, the compared networks were evaluated on all slices of the testing set.

Table 1 shows the quantitative performance comparison of these networks on the testing set. It can be observed that the proposed HFCF-Net consistently achieved the best performance among the compared networks in five performance metrics in the COVID-19 lesion segmentation task. Compared to the ConResNet with the greatest Dice score across other networks, our network improved the Dice score by 4.20%. The performance improvement is mainly attributed to the hybrid-feature fusion between two branches, which provides reliable feature representation and effective information exchange process. It should be noted that 3D segmentation networks (V-Net, 3D U-Net, ConResNet) achieved better Dice and IoU scores, and improved the average Dice score by 7.73% compared to 2D networks.

Figure 3 shows the training loss and validation Dice curves of all compared networks. It can be observed that our proposed HCHF-Net achieved the largest validation Dice score. Meanwhile, the training iterations of 3D networks (proposed HCHF-Net, V-Net, 3D U-Net, and ConResNet) are much more than 2D networks, which indicates the slow convergence of 3D networks. Figure 4

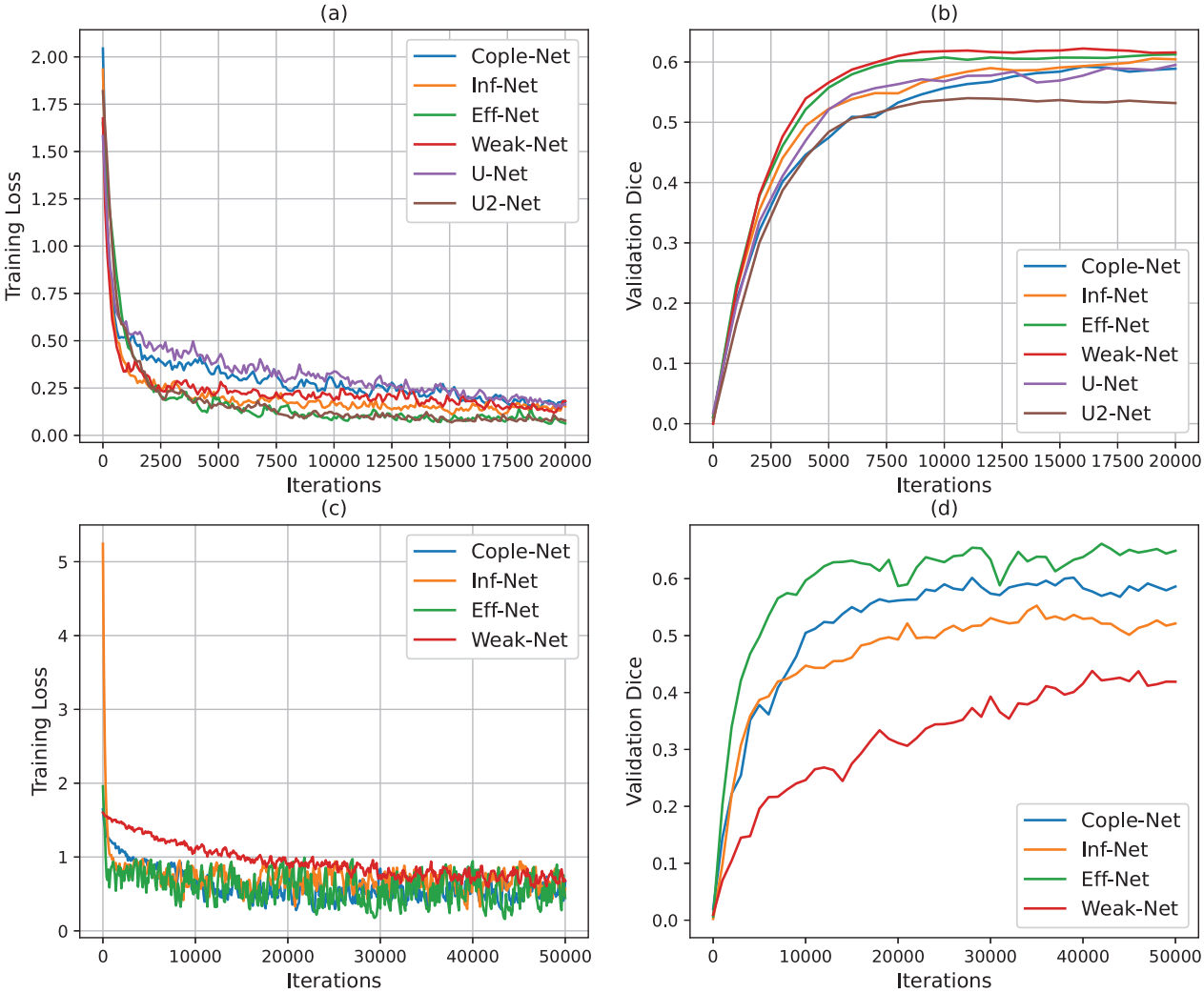

FIGURE 3 Training loss and validation Dice curves of the compared networks. (a) Training loss curves of 2D networks; (b) validation Dice curves of 2D networks; (c) training loss curves of 3D networks; (d) validation Dice curves of 3D networks

shows visual comparison results of some representative segmentation networks. Obviously, the results obtained by our network are most close to the ground truth. In contrast, other networks provided some unsatisfactory results, including fuzzy boundaries and incomplete shapes. It is worth noting that the 3D networks performed better in locating small lesions and segmenting complete lesions than 2D networks. Besides, although the 2.5D-Net achieved a competitive performance in the quantitative comparison, it gave uncompetitive visual results, which exhibit some fragmented regions and unsmooth boundaries. This is mainly because that the ability of 2D convolutional kernels to handle the context information is weak and the limited context information does not bring significant improvement. We conducted an additional experiment to compare the performance of different 2.5D networks and 3D network. The results can be found in Table S1 and Figure S1.

To further better compare the performance of different networks in dealing with lesions at different scales, we split testing images into three groups based on the proportion of lesions: large lesion group containing the cases with lesion proportion greater than 0.02, medium lesion group between 0.02 and 0.005, and small lesion group smaller than 0.005. We listed the quantitative evaluation results in Table 2 and visualized the 3D structure of segmentation results in Figure 5. The quantitative results show that our network had a better performance than others, especially in the segmentation of small and medium lesions. It shows that HFCF-Net achieved the highest Dice score of 69.83%, 71.89%, and 84.73% in the segmentation of small, medium, and large lesions, respectively. Compared to other networks with the best Dice score, our HFCF-Net improved the Dice score by 4.20%, 6.68%, and 3.58% in different-sized lesion segmentation tasks, respectively. Besides, HFCF-Net also outperformed other networks in the 3D visual comparison, demonstrating our network's superiority in segmenting different-sized lesions. It should be noted that our network obtained such a significant

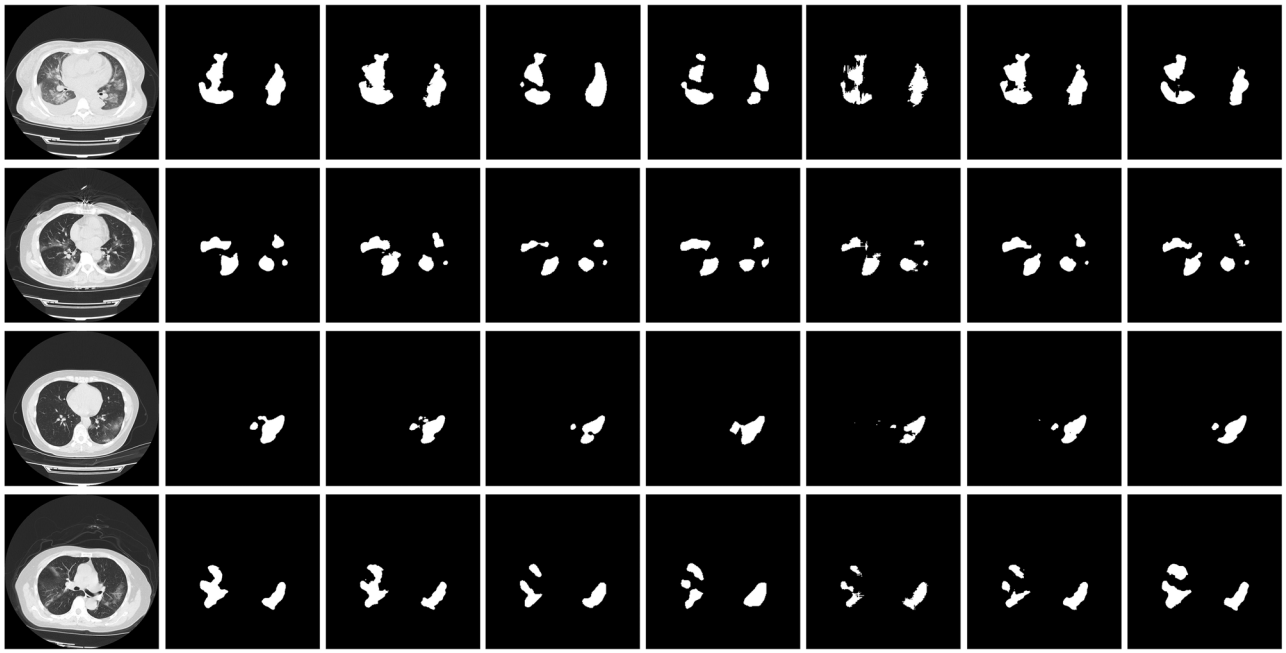

FIGURE 4 Visual comparison of lung lesion segmentation results from different networks. From left to right: (a) computed tomography (CT) image, (b) ground truth, (c) the proposed hybrid-feature cross fusion network (HFCF-Net), (d) COPLE-Net, (e) Inf-Net, (f) 2.5D-Net, (g) ConResNet, and (f) 3D U-Net

TABLE 2 Quantitative comparison results of segmentation networks dealing with different-sized lesions

	Small lesion				Medium lesion				Large lesion			
Network	**Dice**	**IoU**	**HD_{95}**	**Sen**	**Dice**	**IoU**	**HD_{95}**	**Sen**	**Dice**	**IoU**	**HD_{95}**	**Sen**
V-Net (3D)[40]	0.6164	0.4667	13.0539	0.6094	0.6372	0.4803	7.1947	0.7582	0.7451	0.6051	1.0469	0.9143
U-Net (3D)[39]	0.6467	0.5066	7.1682	0.7476	0.6044	0.4567	4.0483	0.8077	0.7843	0.6487	0.6914	0.9051
ConResNet (3D)[14]	0.6270	0.4730	6.0972	0.7422	0.6452	0.5038	3.1777	0.7917	0.8115	0.7135	0.5834	0.8285
2.5D-Net[37]	0.6138	0.4693	4.2922	0.7759	0.6166	0.4593	2.4144	0.8584	0.7135	0.5710	0.6413	0.8975
U-Net[26]	0.5701	0.4243	5.0424	0.6766	0.5734	0.4201	2.2579	0.8162	0.6263	0.4732	0.9951	0.8769
U^2-Net[38]	0.4884	0.3424	6.6258	0.5106	0.5584	0.4014	2.1475	0.7282	0.6756	0.5174	0.7500	0.8603
COPLE-Net[11]	0.5717	0.4238	5.1161	0.6502	0.5951	0.4377	2.4657	0.7791	0.6669	0.5118	0.9607	0.8679
Inf-Net[17]	0.5482	0.3948	5.9282	0.5743	0.6521	0.4929	**1.6291**	0.7474	0.7382	0.5897	0.6290	0.8390
Eff-Net[20]	0.5245	0.3733	6.8818	0.5214	0.6496	0.4898	1.8255	0.7462	0.7377	0.5903	0.7408	0.8113
Weakly-Net[18]	0.5065	0.3561	8.4835	0.6284	0.4689	0.4689	2.5002	0.7567	0.7383	0.5937	0.7528	0.8352
HFCF-Net	**0.6983**	**0.5547**	**3.9752**	**0.7951**	**0.7189**	**0.5733**	2.1817	**0.8626**	**0.8473**	**0.7373**	**0.4171**	**0.9074**

Abbreviations: Dice, Dice similarity coefficient; HD_{95}, 95th percentile of Hausdorff distance; HFCF-Net, hybrid-feature cross fusion network; IoU, intersection over union; Sen, sensitivity.
The bold values means the best performance.

performance gain without increasing too much computational burden that was much less than that of 3D U-Net.

3.3 | Effectiveness of cross fusion module

The major contribution of our study is to design a cross fusion module to achieve the hybrid-feature fusion and information exchange between both subnets for improving segmentation performance. To verify the effectiveness of this module, we performed ablation experiments on three variants: baseline, HFCF-Net-A, and HFCF-Net-B. They refer to HFCF-Net without cross feature fusion, HFCF-Net without prior fusion, and HFCF-Net without context fusion, respectively. The comparison results of two single subnets (2D and 3D subnets) and four fusion networks on the testing set were listed in Table 3. It reveals that only summing two types

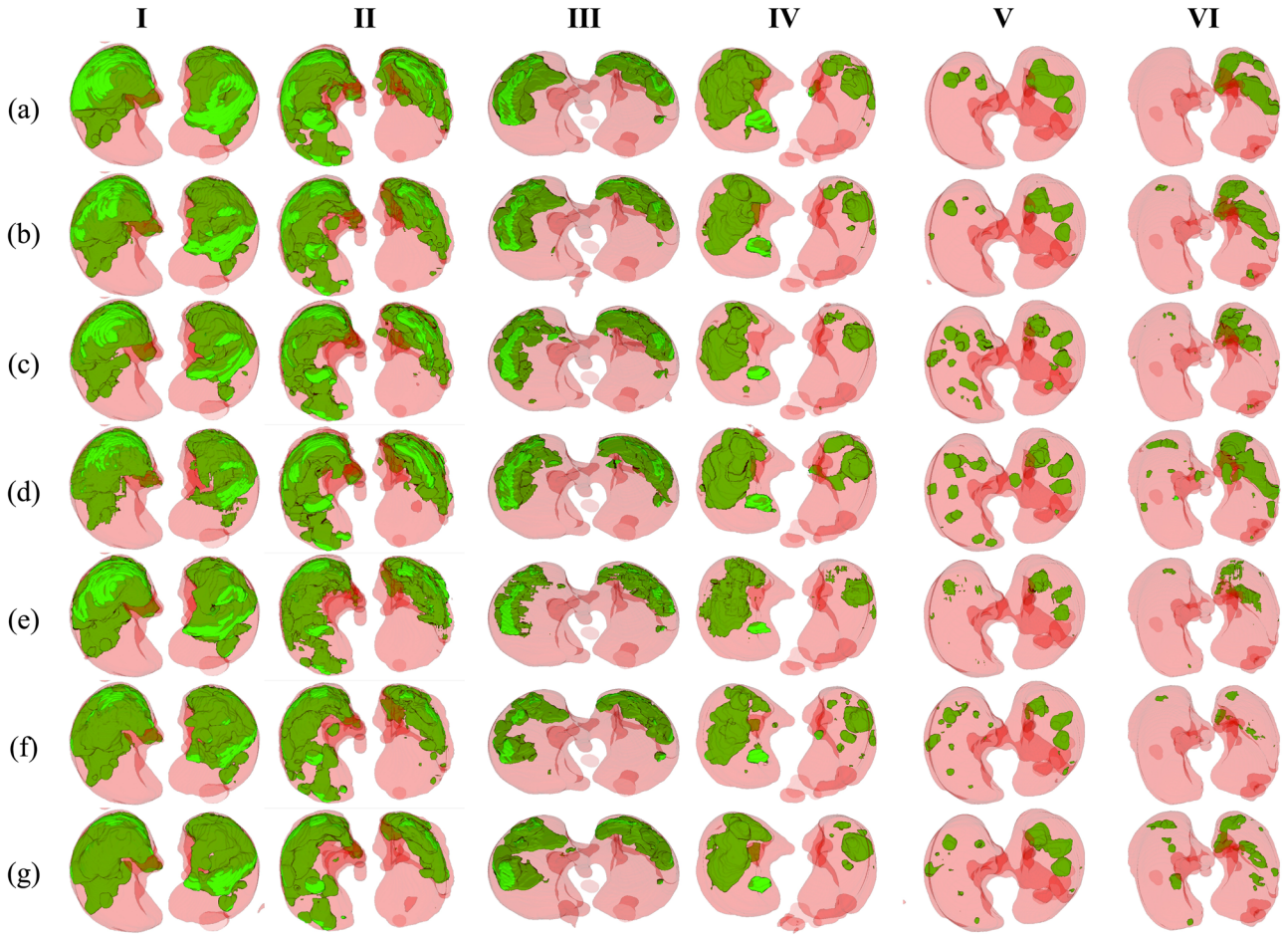

FIGURE 5 Visual comparison of segmentation networks dealing with different-sized lesions. Large lesions: case I; medium lesions: case II; small lesions: case III. For each case, the seven rows show the 3D segmentation visualization maps of (a) ground truth, (b) hybrid-feature cross fusion network (HFCF-Net), (c) COPLE-Net, (d) Inf-Net, (e) 2.5D-Net, (f) ConResNet, and (g) 3D U-Net, respectively

TABLE 3 Performance comparison of 2D subnet, 3D subnet, three hybrid-feature cross fusion network (HFCF-Net) variants, and HFCF-Net

	Final feature fusion	Prior fusion	Context fusion	Dice	IoU	HD_{95}	Sen	Spe
2D subnet	×	×	×	0 .6379	0.4736	3.1780	0.7802	0.9965
3D subnet	×	×	×	0 .7087	0.5567	3.5191	0.8278	0.9978
Baseline	√	×	×	0.7264	0.5896	3.7179	0.7636	**0.9992**
HFCF-Net-A	√	×	√	0.7318	0.5997	3.4297	0.7997	0.9989
HFCF-Net-B	√	√	×	0 .7344	0.5968	3.2743	0.8073	0.9988
HFCF-Net	√	√	√	**0.7485**	**0.6068**	**3.1764**	**0.8358**	0.9990

Abbreviations: Baseline, HFCF-Net without cross feature fusion; Dice, Dice similarity coefficient; HD_{95}, 95th percentile of Hausdorff distance; HFCF-Net-A, HFCF-Net without prior fusion; HFCF-Net-B, HFCF-Net without context fusion; IoU, intersection over union; Sen, sensitivity; Spe, specificity.
The bold values means the best performance.

of feature maps to complete feature fusion without information exchange can help the baseline network achieve a larger Dice score but a worse HD_{95} score. The reason may be that this simple feature fusion was unable to handle noises generated from both subnet features, resulting in disturbing the fusion effect. Afterward, gradually incorporating the cross fusion module, the HFCF-Net achieved a substantial performance improvement in all metrics. Specifically, compared with the network using prior fusion alone (HFCF-Net-A) and the one using context fusion alone (HFCF-Net-B), our network achieved better results (improving Dice by 1.67%, HD_{95} by 0.2533, sensitivity by 3.61% compared to HFCF-Net-A; improving Dice by 1.41%, HD_{95} by 0.0979, sensitivity by 2.85% compared to HFCF-Net-B). The performance gains indicate that the cross feature fusion indeed

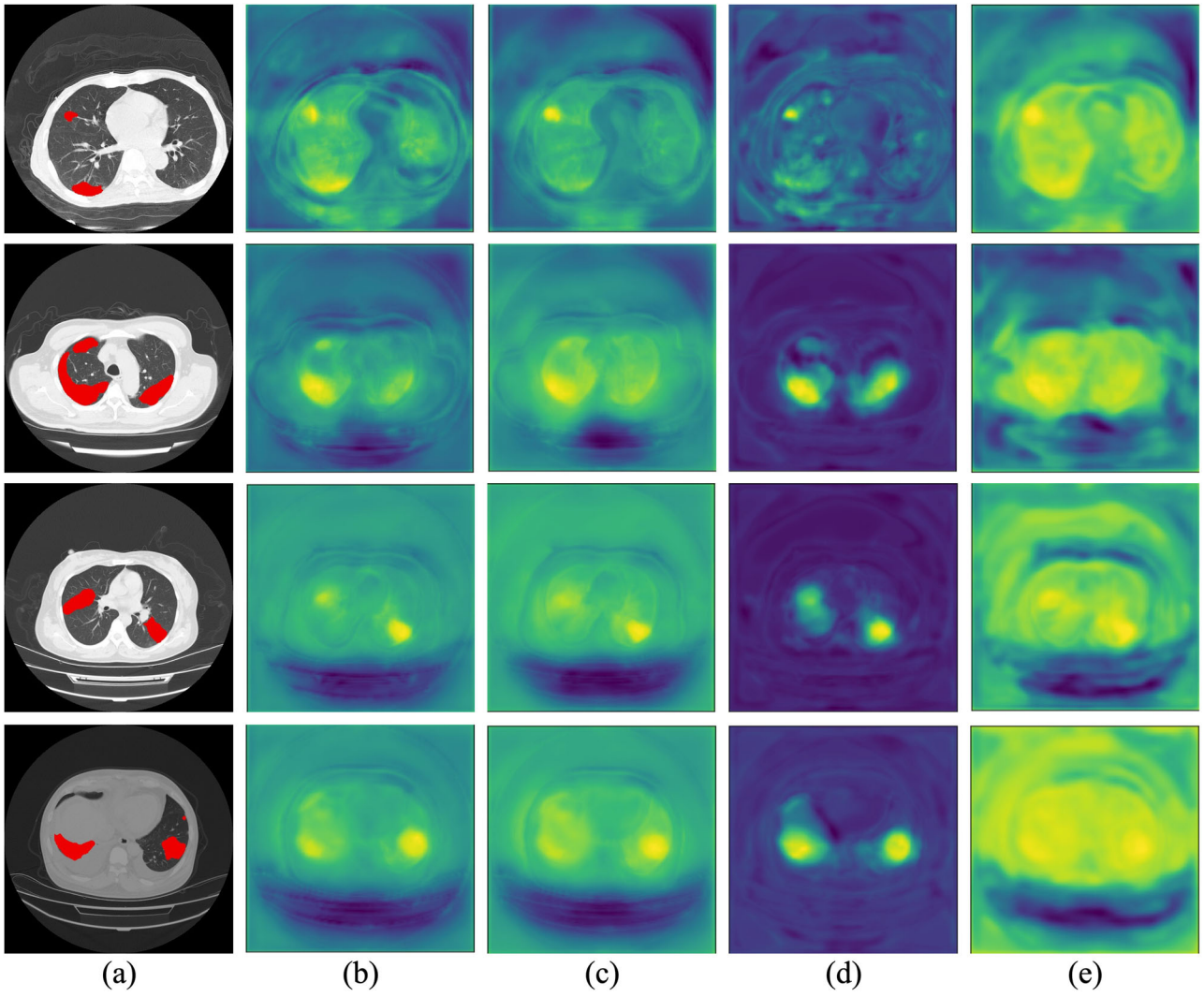

FIGURE 6 Visualization of the final fusion feature maps and transferred feature maps. From left to right: (a) ground truth, (b) final fusion feature maps with cross feature fusion, (c) final fusion feature maps without cross feature fusion, (d) 2D transferred feature maps, and (e) 3D transferred feature maps

effectively employed bidirectional information flow at different scales to simultaneously enhance the feature processing capability of the two subnets to boost segmentation accuracy. Moreover, the cross fusion manner and end-to-end training can accumulate the learned feature context and jointly optimize the 2D and 3D subnets, which can fully explore the hybrid features for better segmentation.

In Figure 6, to clearly show the efficacy of cross feature fusion, we visualized the final fusion feature maps with or without cross feature fusion, and 2D and 3D transferred feature maps (obtained by computing the summation of multiple channels of feature maps from the last prior fusion path and the last context fusion path, respectively). It can be observed that both transferred feature maps highlighted the infected regions. Meanwhile, the 3D transferred feature maps contained rich context information, and the 2D feature maps mainly presented the intra-slice location information of lesions. The fusion feature maps obtained through cross feature fusion of both 2D and 3D transferred features explored more discriminative features than the feature maps of the network without fusion, which is beneficial for accurate recognition of the infected regions.

We also displayed the segmentation results obtained by these variants in Figure 7. It shows that the results produced by our network are the closest to the ground truth. The network without cross feature fusion tended to ignore the small lesions and generated unsatisfied segmentation boundaries. As the repeated cross fusion of learned intra- and inter-slice information, the segmentation results were gradually improved. Adding the context fusion path can solve subtle segmentation faults, and involving the prior fusion path can further promote segmentation precision.

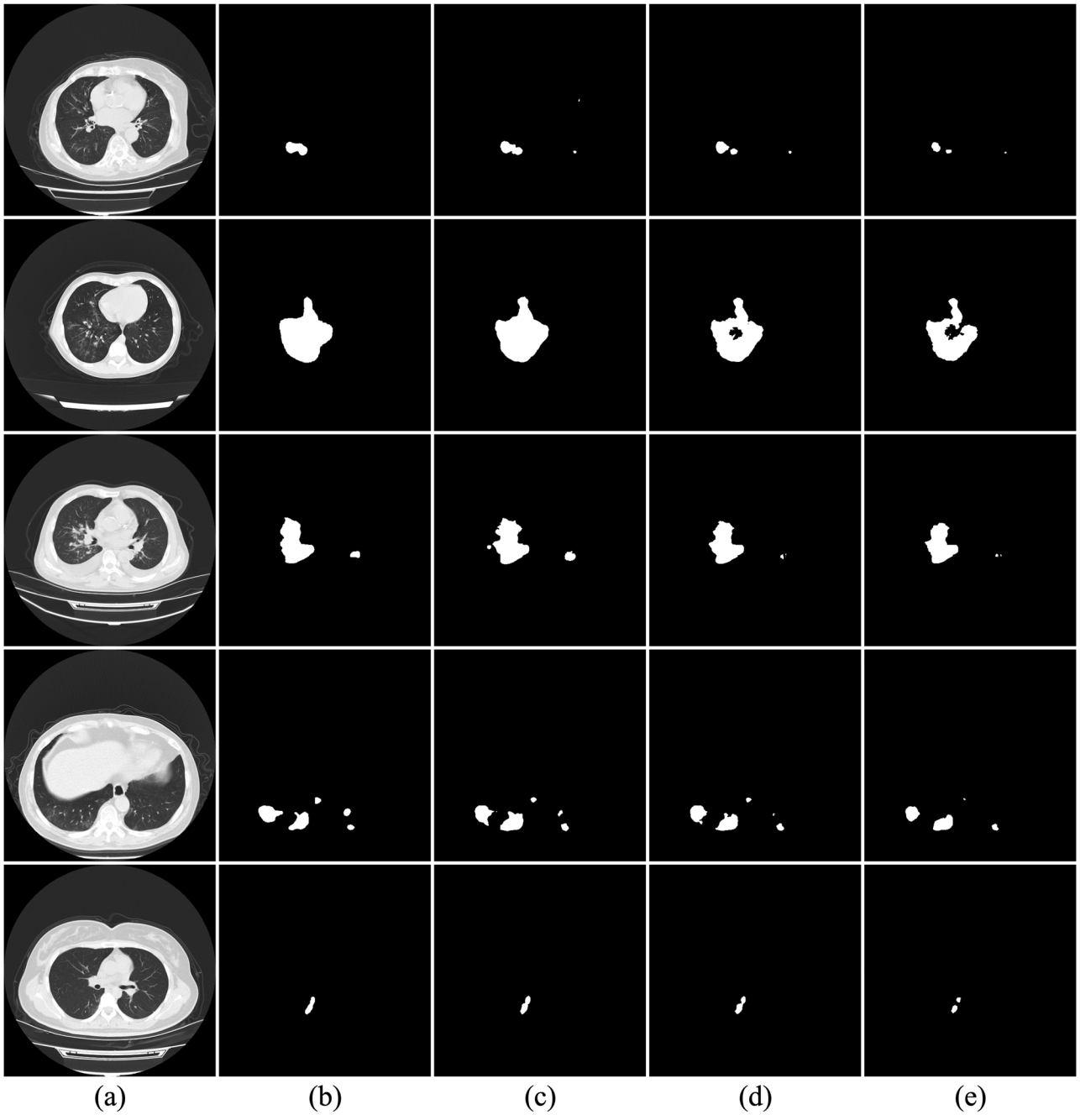

FIGURE 7 Visualization of segmentation results produced by hybrid-feature cross fusion network (HFCF-Net) and its variants. From left to right: (a) computed tomography (CT) images, (b) ground truth and segmentation results produced by (c) HFCF-Net, (d) HFCF-Net without context fusion, and (e) HFCF-Net without both fusion

3.4 | Effectiveness of 2D and 3D subnets

In the 2D subnet, we introduced aggregate interaction modules and Res2net modules to capture rich multi-scale information. In the 3D subnet, we improved the original 3D U-net to be a lightweight backbone with two important aspects, that is, adjusting the convolutional kernel number and inserting a DilRes block. To confirm the effectiveness of these designs, we conducted ablation experiments with different configurations for 2D and 3D subnets.

First, the proposed 2D subnet was compared with three networks (i.e., the network without AIM, the one without Res2Net, and the baseline network without both modules), as shown in Table 4. Compared with the baseline network, the introduction of both modules can both help the 2D subnet achieve better performance. In terms of the Dice indicator, using the Res2Net alone improved

TABLE 4 Quantitative evaluation results of aggregate interaction module (AIM) and Res2Net blocks

Network	AIM	Res2Net	Dice	IoU	HD_{95}	Sen	Spe
Baseline	×	×	0.6124	0.4512	3.3811	0.7653	**0.9967**
2D subnet	×	√	0.6271	0.4692	3.2852	0.7709	**0.9967**
2D subnet	√	×	0.6237	0.4592	3.2295	0.7790	0.9966
2D subnet	√	√	**0.6379**	**0.4736**	**3.1780**	**0.7802**	0.9965

Abbreviations: Dice, Dice similarity coefficient; HD_{95}, 95th percentile of Hausdorff distance; IoU, intersection over union; Sen, sensitivity; Spe, specificity.
The bold values means the best performance.

TABLE 5 Quantitative evaluation results of filter number reduction and dilated residual (DilRes) block

Network	Filter reduction	DilRes block	Dice	IoU	HD_{95}	Sen	Spe	FLOPs	Param
Baseline	×	×	0.6858	0.5407	3.7606	0.8109	0.9975	1.895T	16.320M
3D subnet	√	×	0 .6679	0.5248	4.2179	0.7688	**0.9980**	1.560T	5.70M
3D subnet	√	√	**0.7087**	**0.5567**	**3.5191**	**0.8278**	0.9978	1.582T	11.012M

Abbreviations: Dice, Dice similarity coefficient; FLOPs, floating point operations; HD_{95}, 95th percentile of Hausdorff distance; IoU, intersection over union; Sen, sensitivity; Spe, specificity; Param, parameter number.
The bold values means the best performance.

TABLE 6 Results of ablation experiments with different loss functions

$L_{\text{reweighted-Dice}}$	L_{bce}	Adaptive strategy	Dice	IoU	HD_{95}	Sen	Spe
×	√	×	0.7308	0.5822	3.7189	0.7964	0.9982
√	×	×	0.7370	0.5985	3.5077	0.8097	0.9981
√	√	×	0 .7422	0.6034	3.3632	0.8243	**0.9990**
√	√	√	**0.7485**	**0.6068**	**3.1764**	**0.8358**	**0.9990**

Abbreviations: Dice, Dice similarity coefficient; HD_{95}, 95th percentile of Hausdorff distance; IoU, intersection over union; Sen, sensitivity; Spe, specificity.
The bold values means the best performance.

the score by 1.47%, and using the AIM module alone improved it by 1.13%. When combining both modules, the 2D subnet earned a considerable improvement of 2.55% compared to the baseline network.

Second, the proposed 3D subnet was compared with the baseline network without filter number adjustment and the DilRes block, and the 3D subnet with only filter number adjustment, as shown in Table 5. It shows that when reducing the filter number, the 3D subnet obtained a worse segmentation result compared to the baseline network. Then with the adding of the DilRes block, the 3D subnet improved the Dice by 2.29%, and also reduced the computation and parameter number compared to the baseline network.

3.5 | Effectiveness of imbalance-robust adaptive learning loss

To alleviate the imbalance problem, HFCF-Net proposed a new loss function that combined both the BCE function and reweighted Dice loss function, accompanied by an adaptive learning strategy. We conducted ablation experiments to quantitatively investigate the impact of the loss function and adaptive learning strategy. Table 6 shows the evaluation results with different loss functions. For the BCE function L_{bce}, the specificity is 0.01% higher than that of the reweighted Dice loss function $L_{\text{reweighted-Dice}}$, but the sensitivity is 1.33% lower than that of $L_{\text{reweighted-Dice}}$. It confirms that when the number of lesion pixels is far less than that of background pixels, L_{bce} made the prediction more biased toward the background, and $L_{\text{reweighted-Dice}}$ can effectively improve the performance degradation caused by this imbalance. Compared to only using L_{bce}, using both loss functions (L_{bce} and $L_{\text{reweighted-Dice}}$) improved the Dice score by 1.14%, the specificity by 0.08%, and the sensitivity by 1.79%. Finally, when combined with the adaptive learning strategy, the entire proposed network further achieved a Dice score of 74.85% and a sensitivity of 83.58%.

Figure 8 shows the validation Dice curves of the 2D subnet, 3D subnet, and our HFCF-Net. Obviously, the 2D subnet had a faster convergency speed but a lower accuracy than the 3D subnet. The HCHF-Net had a convergency speed similar to the 2D subnet but reached a higher accuracy than the 3D subnet. It demonstrates that ensembling both subnets and incorporating feature interaction can help the network promote the effectiveness and efficiency of training.

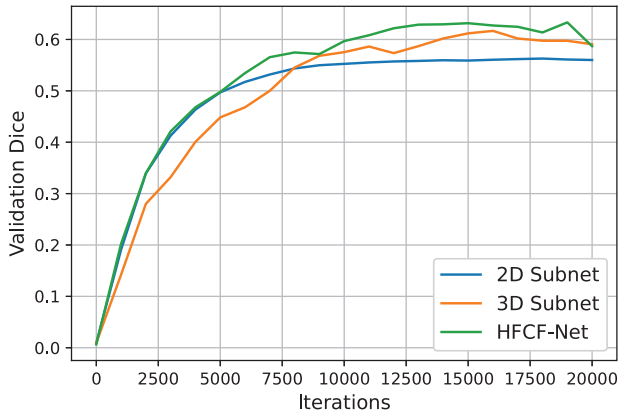

FIGURE 8 Validation Dice curves of our hybrid-feature cross fusion network (HFCF-Net) versus the 2D subnet and 3D subnet within 20k iterations

4 | DISCUSSION

4.1 | Brief summary and result analysis

Automatic lung lesion segmentation plays an essential role in clinical diagnosis to help doctors implement follow-up treatment. In this study, we propose an end-to-end COVID-19 lesion segmentation network that innovatively combines 2D and 3D networks to take their advantages to complement each other by fusing hybrid features in turn. We conduct extensive experiments to validate the effectiveness of our network. Compared with existing studies, the proposed HFCF-Net not only learns the feature representations of the intra-slice lesions via 2D multi-scale subnet, but also explores the inter-slice contextual information via 3D lightweight subnet, outperforming other state-of-the-art networks on the segmentation of lesions with different scales and shapes.

To better understand the performance improvement, we combine experimental results with clinical practice to analyze the efficacy of our network. The most commonly used clinical data are 3D volume data, such as CT and magnetic resonance imaging. The reason may be that 3D data contains more information than 2D data, which is more conducive to analyzing pathological features. In the diagnosis process, doctors often need to observe many consecutive slices to get the final diagnosis results. Otherwise, if doctors only observe a single slice image, it is easy to cause doctors to misdiagnose.

Our experimental results also indicate that 2D networks often mistakenly identify some small lesion regions as normal tissues since some lesions with low contrast often appear grayish-white and look very similar to normal tissues in the lung. But in the 3D case, the 3D networks can determine whether low contrast regions belonging to the lesions are true lesions based on the adjacent slices, as the lesions usually only occupy a small number of slices, while normal tissues run through the entire lung. Meanwhile, with the supplement of the consistent information between slices, the 3D networks can also improve the segmentation of some lesions with fuzzy boundaries.

Therefore, the performance gains are mainly contributed to the improvement of segmentation results of small lesions and fuzzy boundaries. Once the segmentation performance of these two types of objects is improved, the overall segmentation accuracy can be increased by a great margin.

4.2 | Applying segmentation network to non-COVID cases

Extensive experiments have been conducted to confirm the superior performance of our network in terms of COVID cases. To verify whether our HFCF-Net can be employed in clinical practice, we evaluated our network and other competitive networks on the normal cases and common pneumonia (CP) cases of the Covid-19-CT dataset,[41] which are constructed from cohorts from the China Consortium of Chest CT Image Investigation. Since non-COVID data do not contain any COVID-19 lesion annotations, conventional segmentation metrics cannot be used to evaluate these networks. Therefore, we adopted the pixel accuracy (PA) and area under the receiver operating characteristic curve (AUC) to assess the segmentation performance of the compared networks on non-COVID data. PA is a metric to explicitly gauge segmentation accuracy by calculating the proportion of correctly predicted pixels in the segmentation maps. AUC is used to implicitly measure the ability of segmentation networks to extract discriminative COVID-19 features. To calculate the AUC value, we employed a pre-trained COVID-19 classification network[42] to generate the class probability (normal and COVID) of the lung lesion maps produced by the compared segmentation networks on non-COVID data.

We selected some competitive 2D and 3D networks, that is, Inf-Net, COPLE-Net, Eff-Net, and ConResNet, to perform the segmentation experiments on non-COVID data. Considering that the previous 2D networks may produce a higher false-positive rate when facing non-COVID data, we retrained 2D networks with all slices of data to conduct a comprehensive comparison, and these networks were denoted as Inf-Net′, COPLE-Net′, Eff-Net′.

The quantitative results listed in Table 7 shows that our network yielded relatively high performance. The PA of CP and normal cases, and AUC are 0.9993, 0.9997, and 0.8038, respectively. Our network obtained the best value among all networks in terms of the PA of CP cases and outperformed most compared networks in terms of the PA of normal cases. Since the PA values of all networks are relatively close, the performance difference among networks cannot be clearly reflected. The

TABLE 7 Comparison of segmentation results on non-COVID cases

	Pixel accuracy (CP)	Pixel accuracy (normal)	Diagnosis AUC
COPLE-Net[11]	0.9913	0.9992	0.7839
Inf-Net[17]	0.9919	0.9995	0.6646
Eff-Net[20]	0.9925	0.9992	0.6120
COPLE-Net′[11]	0.9939	**0.9998**	0.7817
Inf-Net′[17]	0.9945	0.9997	0.7533
Eff-Net′[20]	0.9960	0.9995	0.7028
ConResNet (3D)[14]	0.9924	**0.9998**	0.7958
U-Net (3D)[39]	0.9989	0.9992	0.7587
HFCF-Net	**0.9993**	0.9997	**0.8038**

Abbreviations: AUC, area under the receiver operating characteristic curve; CP, common pneumonia; HFCF-Net, hybrid-feature cross fusion network.
The bold values means the best performance.

AUC metric can effectively evaluate segmentation performance of each network. Compared with other networks, the AUC value of our HFCF-Net is the highest. It reveals the superiority of our network in the exploration of discriminative COVID-19 lesion features. Figure 9 visualized the distribution of segmentation results of all the compared networks. In contrast to other networks, our network offered a great advantage in identifying true COVID-19 lesion pixels.

In addition, as can be seen from Figure 9, when segmenting the CT image of a normal person, 3D networks would generate fewer wrongly classified pixels than 2D networks, thus achieving better segmentation results. This phenomenon is more apparent when we segment CT images of patients with pneumonia, since the image of the pneumonia cases contains much more noise than that of the normal ones. If there is no contextual information, the 2D networks would easily misjudge the lesion of pneumonia as the COVID lesion. The results also reveal that the retrained networks with all slices perform better than the networks trained with only lesion slices in segmenting non-COVID data. As shown in Figure 9, the false-positive rate of new 2D networks is much lower than that of previous 2D networks.

To conclude, although both quantitative and qualitative comparison results demonstrate the effectiveness of the proposed HFCF-Net in the segmentation task of non-COVID data, there is still room for improvement in our network. We will investigate a classification network to judge whether suspicious lesions are normal tissues for alleviating the false-positive rate in our future work.

4.3 | Application and limitation

In clinical practice, our segmentation network could combine quantitative analysis tools with user interactive display interfaces to build an interactive diagnostic system that can help doctors diagnose illness more quickly. According to the anti-epidemic situation in many nations, the detection speed of RT-PCR is significantly behind the increasing rate of suspicious cases during the epidemic breakout stage. Our segmentation network is intended to assist clinicians in serving as an effective screening tool to reduce patient wait times and shorten diagnostic workflow times, thus lowering radiologists' overall workload and allowing them to respond swiftly in emergencies. On the other hand, RT-PCR detection is unable to diagnose the severity of a patient's disease. Therefore, the integrated segmentation and analysis system can play a crucial role.

Regardless of whether people are confirmed to be COVID-19 pneumonia by RT-PCR or not, CT imaging can evaluate the lung condition for them. The lesion segmentation system can assist in locating suspicious lesions and performing additional pathological data analysis for the probable lesions so that doctors can make a definitive diagnosis. Furthermore, the system could assess the severity of patients diagnosed with COVID-19 pneumonia so that doctors can formulate reasonable treatment plans for them. In addition, if people do not receive nucleic acid testing or have not got the results yet, the system can check their CT scans to discover suspicious lesions, as extensive experiments have proved the accurate performance of our network in the segmentation task of CT scans without lesions.

Although our work has achieved outstanding results in lesion segmentation, the current network still has some limitations. First, our segmentation network only considered the COVID-19 lesions and may not perform well when dealing with CT images with non-COVID diseases. In real life, patients may have various lung diseases, such as CP. The non-COVID diseases will be taken into account in our follow-up work as well. Second, even if some data augmentation techniques were used to expand the sample number, such as random transformations, there still exists the risk of overfitting. For supervised image segmentation, the effectiveness of a network is largely determined by the training on a significant amount of annotated data. However, due to the complexity and time-consuming of annotating data, there are currently few publicly accessible datasets with a large number of annotated images. In the future, we will focus on the weakly supervised and unsupervised image segmentation to tackle the problem of less labeled data.

Furthermore, since the hospitals use different scanners with varied parameters to generate CT images, these data often have inconsistent distributions, such as variations in appearance, as seen in Figure 10. It could cause the networks to overfit on the training datasets and lack generalization ability on the unseen testing datasets, thereby degrading the segmentation performance and bringing challenges to the clinical application

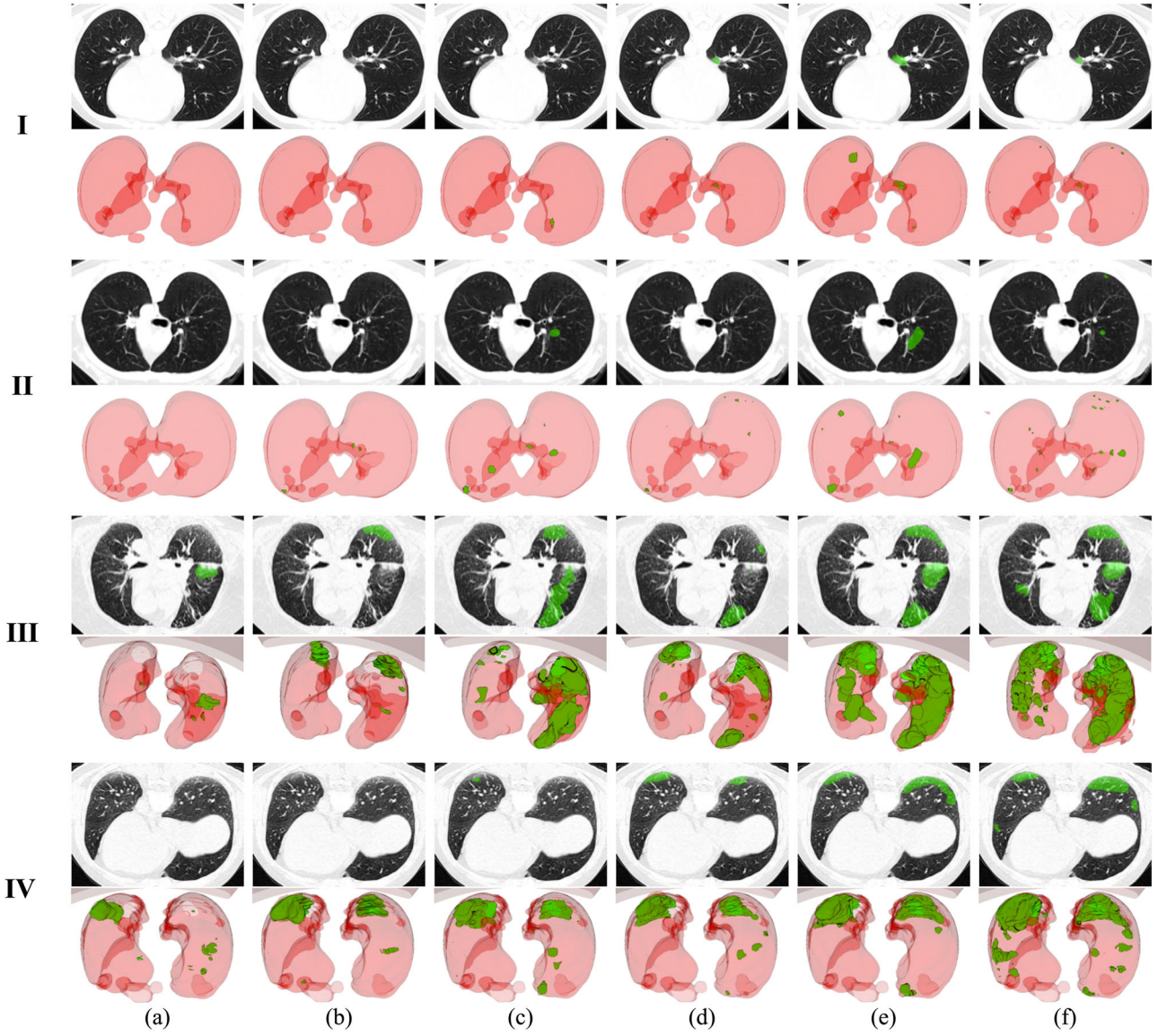

FIGURE 9 Visualization of segmentation results obtained by different networks on some non-COVID cases. From top to down: I and II are normal people's maps, and III and IV are common pneumonia patients' maps. From left to right: 3D segmentation results of (a) hybrid-feature cross fusion network (HFCF-Net) and (b) ConResNet, and 2D segmentation results of (c) Inf-Net′, (d) COPLE-Net′, (e) Inf-Net, and (f) COPLE-Net

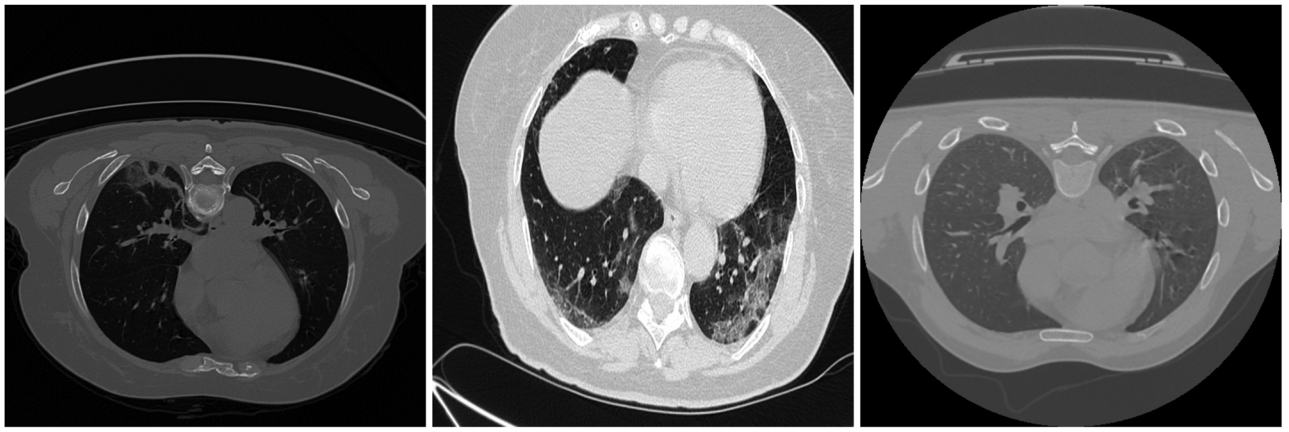

FIGURE 10 Heterogeneity between data from different hospitals

of our network. Applying domain adaptation techniques for reducing the heterogeneity of the multi-center data could be another research direction to further improve the clinical segmentation performance of the networks.

5 | CONCLUSION

In this paper, we proposed an end-to-end HFCF-Net for COVID-19 lung lesion segmentation in CT volume data. It first explored abundant information between and within slices, followed by the effective cross fusion of hybrid features to jointly optimize 2D and 3D branches for achieving competitive performance. To better train the proposed network, we designed a novel loss function with an adaptive learning strategy to effectively tackle the imbalance problem between the proportions of lesion and non-lesion voxels. The proposed network innovatively utilized the advantages that 2D subnet requires less computation overhead and 3D subnet contains rich spatial information. Extensive experiments conducted on the publicly available dataset have proved that the proposed network reached the segmentation performance of 74.85% on the Dice score, superior to the state-of-the-art networks. The visual comparison of segmentation performance also demonstrates that our network outperformed the other networks.

ACKNOWLEDGMENTS

Our work was supported by Natural Science Foundation of Beijing (no. M21012).

CONFLICT OF INTEREST

The authors have no relevant conflicts of interest to disclose.

DATA AVAILABILITY STATEMENT

All data comes from the public dataset.

REFERENCES

1. Hui DS, Azhar EI, Madani TA, et al. The continuing 2019-nCoV epidemic threat of novel coronaviruses to global health—the latest 2019 novel coronavirus outbreak in Wuhan, China. *Int J Infect Dis.* 2020;91:264-266.
2. Zhu N, Zhang D, Wang W, et al. A novel coronavirus from patients with pneumonia in China, 2019. *N Engl J Med.* 2020;382(8):727-733.
3. Wang C, Horby PW, Hayden FG, Gao GF. A novel coronavirus outbreak of global health concern. *Lancet.* 2020;395(10223):470-473.
4. Van Doremalen N, Bushmaker T, Morris DH, et al. Aerosol and surface stability of SARS-CoV-2 as compared with SARS-CoV-1. *N Engl J Med.* 2020;382(16):1564-1567.
5. The Center for Systems Science and Engineering at Johns Hopkins University. *Coronavirus COVID-19 Global Cases*. 2020. https://coronavirus.jhu.edu/map.html
6. Ng M-Y, Lee EYP, Yang J, et al. Imaging profile of the COVID-19 infection: radiologic findings and literature review. *Radiol Cardiothorac Imaging.* 2020;2(1):e200034.
7. Zu ZY, Di Jiang M, Xu PP, et al. Coronavirus disease 2019 (COVID-19): a perspective from China. *Radiology.* 2020;296(2):E15-E25.
8. Ai T, Yang Z, Hou H, et al. Correlation of chest CT and RT-PCR testing for coronavirus disease 2019 (COVID-19) in China: a report of 1014 cases. *Radiology.* 2020;296(2):E32-E40.
9. Chung M, Bernheim A, Mei X, et al. CT imaging features of 2019 novel coronavirus (2019-nCoV). *Radiology.* 2020;295(1):202-207.
10. Xie X, Zhong Z, Zhao W, Zheng C, Wang F, Liu J. Chest CT for typical coronavirus disease 2019 (COVID-19) pneumonia: relationship to negative RT-PCR testing. *Radiology.* 2020;296(2):E41-E45.
11. Wang G, Liu X, Li C, et al. A noise-robust framework for automatic segmentation of COVID-19 pneumonia lesions from CT images. *IEEE Trans Med Imaging.* 2020;39(8):2653-2663.
12. Huang L, Han R, Ai T, et al. Serial quantitative chest CT assessment of COVID-19: a deep learning approach. *Radiol Cardiothorac Imaging.* 2020;2(2):e200075.
13. Cowan IA, MacDonald SLS, Floyd RA. Measuring and managing radiologist workload: measuring radiologist reporting times using data from a radiology information system. *J Med Imaging Radiat Oncol.* 2013;57(5):558-566.
14. Zhang J, Xie Y, Wang Y, Xia Y. Inter-slice context residual learning for 3D medical image segmentation. *IEEE Trans Med Imaging.* 2021;40(2):661-672. https://doi.org/10.1109/TMI.2020.3034995
15. Litjens G, Kooi T, Bejnordi BE, et al. A survey on deep learning in medical image analysis. *Med Image Anal.* 2017;42:60-88.
16. Oulefki A, Agaian S, Trongtirakul T, Laouar AK. Automatic COVID-19 lung infected region segmentation and measurement using CT-scans images. *Pattern Recognit.* 2020;114:107747.
17. Fan D-P, Zhou T, Ji G-P, et al. Inf-Net: automatic COVID-19 lung infection segmentation from CT images. *IEEE Trans Med Imaging.* 2020;39(8):2626-2637.
18. Laradji I, Rodriguez P, Manas O, et al. A weakly supervised consistency-based learning method for COVID-19 segmentation in CT images. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2021:2453-2462.
19. Yao Q, Xiao L, Liu P, Zhou SK. Label-Free Segmentation of COVID-19 Lesions in Lung CT. arXiv Prepr arXiv200906456. 2020.
20. Feng Y, Liu S, Cheng Z, et al. Severity Assessment and Progression Prediction of COVID-19 Patients Based on the LesionEncoder Framework and Chest CT. medRxiv. 2020.
21. Chen X, Yao L, Zhang Y. Residual Attention U-Net for Automated Multi-Class Segmentation of COVID-19 Chest CT Images. arXiv Prepr arXiv200405645. 2020.
22. Zhou T, Canu S, Ruan S. An Automatic COVID-19 CT Segmentation Network Using Spatial and Channel Attention Mechanism. arXiv Prepr arXiv200406673. 2020.
23. Shan F, Gao Y, Wang J, et al. Lung Infection Quantification of COVID-19 in CT Images with Deep Learning. arXiv Prepr arXiv200304655. 2020.
24. Yan Q, Wang B, Gong D, et al. COVID-19 Chest CT Image Segmentation–A Deep Convolutional Neural Network Solution. arXiv Prepr arXiv200410987. 2020.
25. Wang Y, Zhang Y, Liu Y, et al. Does non-COVID-19 lung lesion help? Investigating transferability in COVID-19 CT image segmentation. *Comput Methods Programs Biomed.* 2021;202:106004.
26. Ronneberger O, Fischer P, Brox T. U-net: convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer; 2015:234-241.
27. Pang Y, Zhao X, Zhang L, Lu H. Multi-scale interactive network for salient object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020:9413-9422.

28. Gao S, Cheng M-M, Zhao K, Zhang X-Y, Yang M-H, Torr PHS. Res2net: a new multi-scale backbone architecture. *IEEE Trans Pattern Anal Mach Intell.* 2021;43:652-662.
29. Simonyan K, Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv Prepr arXiv14091556. 2014.
30. Yu F, Koltun V. Multi-Scale Context Aggregation by Dilated Convolutions. arXiv Prepr arXiv151107122. 2015.
31. Lin T-Y, Goyal P, Girshick R, He K, Dollár P. Focal loss for dense object detection. In: Proceedings of the IEEE International Conference on Computer Vision. 2017:2980-2988.
32. Shirokikh B, Shevtsov A, Kurmukov A, et al. Universal loss reweighting to balance lesion size inequality in 3D medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer; 2020:523-532.
33. An P, Xu S, Harmon S, et al. CT Images in COVID-19 [Data set]. *The Cancer Imaging Archive (TCIA) Public Access*. 2020. https://doi.org/10.7937/tcia.2020.gqry-nc81
34. Clark K, Vendt B, Smith K, et al. The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository. *J Digit Imaging.* 2013;26(6):1045-1057. https://doi.org/10.1007/s10278-013-9622-7
35. Roth H, Xu Z, Diez CT, et al. Rapid artificial intelligence solutions in a pandemic – the COVID-19-20 lung CT lesion segmentation challenge. *Res Sq.* 2022; Published online 2022. https://doi.org/10.21203/rs.3.rs-571332/v1
36. He K, Zhang X, Ren S, Sun J. Delving deep into rectifiers: surpassing human-level performance on ImageNet classification. In: Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV). ICCV '15. IEEE Computer Society; 2015:1026-1034. https://doi.org/10.1109/ICCV.2015.123
37. Zhou L, Li Z, Zhou J, et al. A rapid, accurate and machine-agnostic segmentation and quantification method for CT-based COVID-19 diagnosis. *IEEE Trans Med Imaging.* 2020;39(8):2638-2652.
38. Qin X, Zhang Z, Huang C, Dehghan M, Zaiane OR, Jagersand M. U2-Net: going deeper with nested U-structure for salient object detection. *Pattern Recognit.* 2020;106:107404.
39. Çiçek Ö, Abdulkadir A, Lienkamp SS, Brox T, Ronneberger O. 3D U-Net: learning dense volumetric segmentation from sparse annotation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer; 2016:424-432.
40. Milletari F, Navab N, Ahmadi S-A. V-net: fully convolutional neural networks for volumetric medical image segmentation. In: 2016 Fourth International Conference on 3D Vision (3DV). IEEE; 2016:565-571.
41. Zhang K, Liu X, Shen J, et al. Clinically applicable AI system for accurate diagnosis, quantitative measurements, and prognosis of COVID-19 pneumonia using computed tomography. *Cell.* 2020;181(6):1423-1433.
42. Wang X, Deng X, Fu Q, et al. A weakly-supervised framework for COVID-19 classification and lesion localization from chest CT. *IEEE Trans Med Imaging.* 2020;39(8):2615-2625. https://doi.org/10.1109/TMI.2020.2995965

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Wang Y, Yang Q, Tian L, Zhou X, Rekik I, Huang H. HFCF-Net: A hybrid-feature cross fusion network for COVID-19 lesion segmentation from CT volumetric images. *Med Phys.* 2022;49:3797–3815. https://doi.org/10.1002/mp.15600