Supplementary Material

I. Details of 3DLSC-COVID database

Database establishment. This retrospective study was performed in accordance with the Declaration of Helsinki of the World Medical Association and was approved by the medical ethics committee of Liyuan Hospital, Tongji Medical College, Huazhong University of Science and Technology. Besides, all data were anonymized. More details about the patients and CT scans of the 3DLSC-COVID database are summarized in the Supplementary table I.

Data annotation. For lesion segmentation, the ground-truth lesions of the CT scans in our DLSC-COVID database were annotated via a professional software, ITK-SNAP [4]. Specifically, there were 3 stages for annotating the lesions in the CT scans. Firstly, we recruited 2 resident radiologists with above 2 years experience to annotate the areas and boundaries of the lesions in each 2D CT slice. Then, these two resident radiologists were asked to further refine the segmented lesions in 3D viewing mode. After that, both 2D and 3D lesions were reviewed and corrected by a senior radiologist with above 10 years experience in thoracic radiology. Finally, the ground-truth 3D lesions were obtained in the form of binary maps for our DLSC-COVID database.

TABLE I
THE PATIENT AND CT INFORMATION IN 3DLSC-COVID DATABASE.

	Characteristics	COVID-19	CAP	Non-pneumonia	Overall
Patient information	**Gender**	**Count | percent(%)**			
	Male	348 | 43.8	303 | 56.1	234 | 49.7	885 | 49.0
	Female	446 | 56.2	237 | 43.9	237 | 50.3	920 | 51.0
	Age	**Count | percent(%)**			
	≤ 18	0 | 0.0	11 | 2.0	16 | 3.4	27 | 1.5
	19-49	263 | 33.2	89 | 16.5	356 | 75.6	708 | 39.2
	50-64	273 | 34.3	142 | 26.3	91 | 19.3	506 | 28.1
	≥ 65	258 | 32.5	298 | 55.2	8 | 1.7	564 | 31.2
	Average age, mean $\pm$ std	57.5 $\pm$ 16.5	65.2 $\pm$ 19.0	37.0 $\pm$ 12.8	54.4 $\pm$ 19.7
	Comorbidity	**Count | percent(%)**			
	Chronic bronchitis	11 | 1.4	16 | 2.9	4 | 0.9	31 | 1.7
	Hypertension	111 | 13.9	51 | 9.4	56 | 11.8	218 | 12.1
	Coronary heart disease	15 | 1.9	17 | 3.1	10 | 2.1	42 | 2.3
	Liver dysfunction	11 | 1.4	10 | 1.8	3 | 0.7	24 | 1.3
	Chronic renal failure	15 | 1.9	11 | 2.1	7 | 1.4	33 | 1.8
	Diabetes	52 | 6.6	48 | 8.8	20 | 4.2	120 | 6.6
	Stroke	20 | 2.5	9 | 1.6	4 | 0.9	33 | 1.8
CT information	**Number of CTs**	794	540	471	1805
	Number of Slices, mean $\pm$ std	260 $\pm$ 30	240 $\pm$ 29	259 $\pm$ 33	254 $\pm$ 32
	Device				
	UIH uCT 510	794	177	302	1273
	Slice Thickness (mm)	1.5	1.5	1.5	1.5
	Spacing Between Slices (mm)	1.2	1.2	1.2	1.2
	GE Optima CT660	0	363	169	532
	Slice Thickness (mm)	1.25	1.25	1.25	1.25
	Spacing Between Slices (mm)	10	10	10	10
	Pixel Spacing, mean $\pm$ std (mm)	0.71 $\pm$ 0.05	0.70 $\pm$ 0.06	0.71 $\pm$ 0.06	0.71 $\pm$ 0.06
	Scan Length, mean $\pm$ std (mm)	313.4 $\pm$ 32.0	309.7 $\pm$ 32.1	325.8 $\pm$ 34.0	315.8 $\pm$ 34.2
	Tube voltage (KVp)	120	120	120	120
	Rows $\times$ Columns	512 $\times$ 512	512 $\times$ 512	512 $\times$ 512	512 $\times$ 512
	Rotation Direction	clockwise	clockwise	clockwise	clockwise

II. Details of the DeepSC-COVID model

As shown in Supplementary Fig. 1 , we provide the detailed structure of the proposed components, including the 3D inception block in the cross-task feature subnet, the down-transition and up-transition units in the 3D lesion subnet and the 3D encoder unit in the classification subnet.

III. Evaluation metrics

In this section, we introduce the metrics used for evaluating the model performance on the tasks of 3D lesion segmentation and classification, respectively.

(1) *3D lesion segmentation.* The evaluation metrics for segmentation include Dice similarity coefficient (DSC) [1], root mean square symmetric surface distance (RMSD) [2], normalized surface Dice (NSD) [3], sensitivity and specificity. The definitions of these 5 metrics are described as follows.

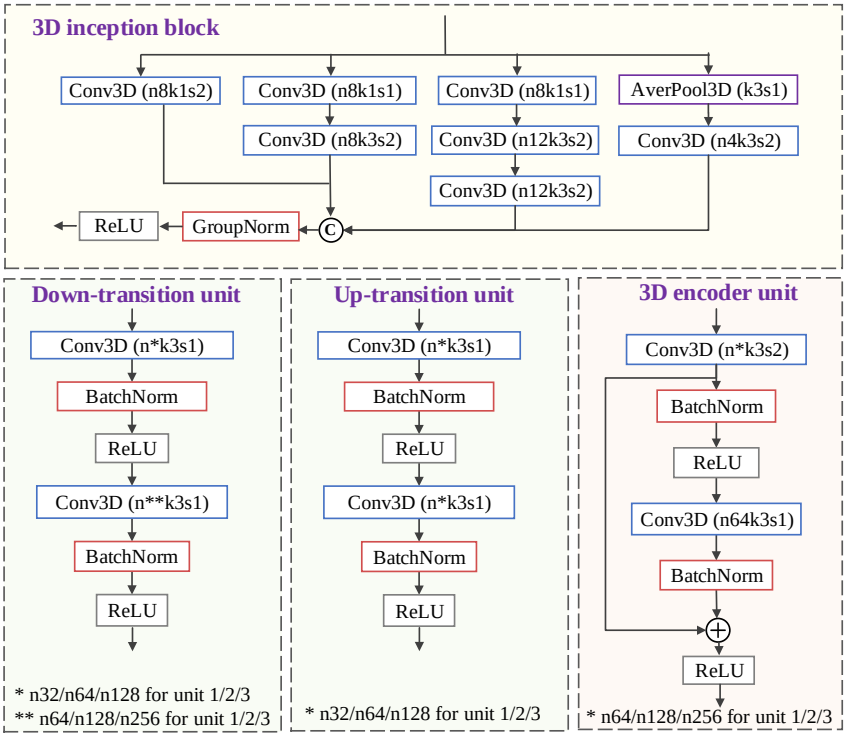

Fig. 1. Illustration of the detailed structure of the proposed components in DeepSC-COVID model, including the 3D inception block in the cross-task feature subnet, the down-transition and up-transition units in the 3D lesion subnet and the 3D encoder unit in the classification subnet.

- DSC is a region-based metric to measure the overlap degree between the ground-truth segmentation $\mathbf{S}$ and the generated segmentation $\hat{\mathbf{S}}$. Mathematically, DSC is calculated as follows,

$$\mathrm{DSC} = \frac{2\|\mathbf{S} \cap \hat{\mathbf{S}}\|_1}{\|\mathbf{S}\|_1 + \|\hat{\mathbf{S}}\|_1}. \tag{1}$$

  According to equation (1), a higher score of DSC means that the generated segmentation is more close to the ground-truth, indicating better segmentation performance.
- RMSD is a boundary-based metrics to measure the distance between the ground-truth surface $\partial\mathbf{S}$ and the generated segmentation surface $\partial\hat{\mathbf{S}}$. Mathematically, RMSD is defined by the following equation,

$$\mathrm{RMSD} = \sqrt{\frac{\sum_{\tilde{\mathrm{x}} \in \partial\mathbf{S}} \|\mathrm{d}(\tilde{\mathrm{x}}, \partial\hat{\mathbf{S}})\|_2^2 + \sum_{\tilde{\mathrm{y}} \in \partial\hat{\mathbf{S}}} \|\mathrm{d}(\tilde{\mathrm{y}}, \partial\mathbf{S})\|_2^2}{\|\partial\mathbf{S}\|_1 + \|\partial\hat{\mathbf{S}}\|_1}}, \tag{2}$$

  where $\mathrm{d}(\tilde{\mathrm{x}}, \partial\hat{\mathbf{S}})$ denotes the distance from the pixel $\tilde{x}$ to the surface $\partial\hat{\mathbf{S}}$.
- NSD is a metric that measures the similarity between the ground-truth surface $\partial\mathbf{S}$ and the generated segmentation surface $\partial\hat{\mathbf{S}}$, which is defined as follows

$$\mathrm{NSD} = \frac{\|\partial\mathbf{S} \cap \mathrm{D}(\partial\hat{\mathbf{S}}, \tau)\|_1 + \|\partial\hat{\mathbf{S}} \cap \mathrm{D}(\partial\mathbf{S}, \tau)\|_1}{\|\partial\mathbf{S}\|_1 + \|\partial\hat{\mathbf{S}}\|_1}, \tag{3}$$
$$\text{where } \mathrm{D}(\mathbf{B}, \tau) = \{\mathrm{x} \in \mathbb{R}^3 \,|\, \exists \tilde{\mathrm{x}} \in \mathbf{B},\ \|\mathrm{x} - \tilde{\mathrm{x}}\|_1 \leq \tau\}.$$

  In equation (3), $\mathrm{D}(\mathbf{B}, \tau)$ is the 3D dilation operation for the surface $\mathbf{B}$, at tolerance τ (=10 mm in this paper). Note that lower RMSD and higher NSD values indicate better segmentation performance.
- Since 3D lesion segmentation is a voxel-wise binary classification problem (i.e., lesion or normal), sensitivity and specificity are also calculated to evaluate the classification accuracy of each voxel in the generated segmentation mask. To be more specific, sensitivity measures the percentage of actual positive (i.e., lesion) voxels that are correctly identified. On the other hand, specificity measures the percentage of negative (i.e., normal) voxels that are correctly identified. Mathematically, the definitions of sensitivity and specificity are formulated as follows,

$$\mathrm{Sensitivity}_{\mathrm{seg}} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}; \quad \mathrm{Specificity}_{\mathrm{seg}} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}} \tag{4}$$

In the above 2 equations, taking lesion and normal voxels as positive and negative samples, TP, TN, FP and FN represent the true positive, true negative, false positive and false negative rates, respectively.

(2) *Disease classification.* The evaluation metrics for disease classification include accuracy, sensitivity, specificity, F_1-score and area under the receiver operating characteristic (ROC) curve (AUC).

- Accuracy stands for the ratio of true predictions over all predictions. Sensitivity measures the percentage of actual positives samples which are correctly identified. Specificity measures the percentage of negative samples which are correctly identified. They are defined in the following equations,

$$\text{Accuracy} = \frac{\sum_{i=1}^{C} \mathrm{TP}_i}{\sum_{i=1}^{C} \mathrm{N}_i}; \quad \text{Sensitivity}_{\text{cls}} = \frac{1}{C}\sum_{i=1}^{C} \frac{\mathrm{TP}_i}{\mathrm{TP}_i + \mathrm{FN}_i}; \quad \text{Specificity}_{\text{cls}} = \frac{1}{C}\sum_{i=1}^{C} \frac{\mathrm{TN}_i}{\mathrm{TN}_i + \mathrm{FP}_i}. \tag{5}$$

Here, C (=3) is the number of classes; $i \in \{1, 2, 3\}$ represents the class of COVID-19, CAP and non-pneumonia, respectively; N_i is the number of samples in class i; TP_i, TN_i, FP_i and FN_i stand for the true positive, true negative, false positive and false negative rates for class i, respectively.

- F_1-score is the trade-off metric that considers the classification accuracy of both positives and negative samples. AUC comprehensively measures the classification effect of various threshold settings for classification. The definitions of these 2 metrics are formulated as follows,

$$\mathrm{F}_1\text{-score} = \frac{1}{C}\sum_{i=1}^{C} \frac{2\mathrm{TP}_i}{2\mathrm{TP}_i + \mathrm{FN}_i + \mathrm{FP}_i}, \tag{6}$$

$$\mathrm{AUC} = \frac{1}{C}\sum_{i=1}^{C} \mathrm{AREA}(\mathrm{ROC}_i). \tag{7}$$

In the above equation, $\mathrm{AREA}(\mathrm{ROC_i})$ is the area under the ROC curve for the i-th class. Note that for all these 5 metrics, higher scores admit better classification performance.

IV. Hyper-parameter setting

The hyper-parameters in our DeepSC-COVID model are tuned for the optimal performance over the validation set in the 3DLSC-COVID database. The training process consists of two stages. At the first stage, we separately pre-train the corresponding subsets for segmentation and classification tasks, respectively. At the second stage, all subnets are simultaneously fine-tuned based on the pre-trained models, over both tasks of segmentation and classification. Some important hyper-parameters of these two stages are listed in Supplementary Table II.

TABLE II
VALUES OF SOME KEY HYPER-PARAMETERS IN THE TWO TRAINING STAGES.

Stage I	Batch size	4
	Initial learning rate	1×10^{-3}
	λ_{dice} for $\mathcal{L}_{\text{seg}}$ in equation (8)	1
	λ_{focal} for $\mathcal{L}_{\text{seg}}$ in equation (8)	2
	Hyper-parameters of focal loss in equation (7)	$\alpha = 0.25, \gamma = 2$
	Combination weight in the 3D lesion subnet	$\eta_1 = 0.50, \eta_2 = 0.25, \eta_3 = 0.25$
Stage II	Batch size	1
	Initial learning rate	1×10^{-4}
	λ_{dice} for $\mathcal{L}_{\text{seg}}$ in equation (8)	1
	λ_{focal} for $\mathcal{L}_{\text{seg}}$ in equation (8)	2
	λ_{ta} for $\mathcal{L}$ in equation (11)	10
	λ_{cls} for $\mathcal{L}$ in equation (11)	2
	λ_{seg} for $\mathcal{L}$ in equation (11)	2
	Hyper-parameters of focal loss in equation (7)	$\alpha = 0.25, \gamma = 2$
	Combination weight in the 3D lesion subnet	$\eta_1 = 0.50, \eta_2 = 0.25, \eta_3 = 0.25$

* All the equations in this table are from the main text.

References

[1] L. R. Dice. Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302, 1945.

[2] T. Heimann, B. Van Ginneken, M. A. Styner, Y. Arzhaeva, V. Aurich, C. Bauer, A. Beck, C. Becker, R. Beichel, G. Bekes, et al. Comparison and evaluation of methods for liver segmentation from CT datasets. *IEEE transactions on medical imaging*, 28(8):1251–1265, 2009.

[3] S. Nikolov, S. Blackwell, R. Mendes, J. De Fauw, C. Meyer, C. Hughes, H. Askham, B. Romera-Paredes, A. Karthikesalingam, C. Chu, et al. Deep learning to achieve clinically applicable segmentation of head and neck anatomy for radiotherapy. *arXiv preprint in https://arxiv.org/abs/1809.04430*, 2018.

[4] P. A. Yushkevich, J. Piven, H. Cody Hazlett, R. Gimpel Smith, S. Ho, J. C. Gee, and G. Gerig. User-guided 3D active contour segmentation of anatomical structures: Significantly improved efficiency and reliability. *Neuroimage*, 31(3):1116–1128, 2006.