Unpaired Cross-Modal Interaction Learning for COVID-19 Segmentation on Limited CT Images

Qingbiao Guan[1,2], Yutong Xie[3], Bing Yang[2], Jianpeng Zhang[2], Zhibin Liao[3], Qi Wu[3], and Yong Xia[1,2(✉)]

[1] Ningbo Institute of Northwestern Polytechnical University, Ningbo 315048, China
yxia@nwpu.edu.cn

[2] National Engineering Laboratory for Integrated Aero-Space-Ground-Ocean Big Data Application Technology, School of Computer Science and Engineering, Northwestern Polytechnical University, Xi'an 710072, China

[3] Australian Institute for Machine Learning, The University of Adelaide, Adelaide, SA, Australia

Abstract. Accurate automated segmentation of infected regions in CT images is crucial for predicting COVID-19's pathological stage and treatment response. Although deep learning has shown promise in medical image segmentation, the scarcity of pixel-level annotations due to their expense and time-consuming nature limits its application in COVID-19 segmentation. In this paper, we propose utilizing large-scale unpaired chest X-rays with classification labels as a means of compensating for the limited availability of densely annotated CT scans, aiming to learn robust representations for accurate COVID-19 segmentation. To achieve this, we design an Unpaired Cross-modal Interaction (UCI) learning framework. It comprises a multi-modal encoder, a knowledge condensation (KC) and knowledge-guided interaction (KI) module, and task-specific networks for final predictions. The encoder is built to capture optimal feature representations for both CT and X-ray images. To facilitate information interaction between unpaired cross-modal data, we propose the KC that introduces a momentum-updated prototype learning strategy to condense modality-specific knowledge. The condensed knowledge is fed into the KI module for interaction learning, enabling the UCI to capture critical features and relationships across modalities and enhance its representation ability for COVID-19 segmentation. The results on the public COVID-19 segmentation benchmark show that our UCI with the inclusion of chest X-rays can significantly improve segmentation performance, outperforming advanced segmentation approaches including nnUNet, CoTr, nnFormer, and Swin UNETR. Code is available at: https://github.com/GQBBBB/UCI.

Keywords: Covid-19 Segmentation · Unpaired data · Cross-modal

Q. Guan and Y. Xie—Contributed equally to this work.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-43898-1_58.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14222, pp. 603–613, 2023.
https://doi.org/10.1007/978-3-031-43898-1_58

1 Introduction

The COVID-19 pneumonia pandemic has posed an unprecedented global health crisis, with lung imaging as a crucial tool for identifying and managing affected individuals [16]. The commonly used imaging modalities for COVID-19 diagnosis are chest X-rays and chest computerized tomography (CT). The latter has been the preferred method for detecting acute lung manifestations of the virus due to its exceptional imaging quality and ability to produce a 3D view of the lungs. Effective segmentation of COVID-19 infections using CT can provide valuable insights into the disease's development, prediction of the pathological stage, and treatment response beyond just screening for COVID-19 cases. However, the current method of visual inspection by radiologists for segmentation is time-consuming, requires specialized skills, and is unsuitable for large-scale screening. Automated segmentation is crucial, but it is also challenging due to three factors: the infected regions often vary in shape, size, and location, appear similar to surrounding tissues, and can disperse within the lung cavity. The success of deep convolutional neural networks (DCNNs) in image segmentation has led researchers to apply this approach to COVID-19 segmentation using CT scans [7,14,17]. However, DCNNs require large-scale annotated data to explore feature representations effectively. Unfortunately, publicly available CT scans with pixel-wise annotations are relatively limited due to high imaging and annotation costs and data privacy concerns. This limited data scale currently constrains the potential of DCNNs for COVID-19 segmentation using CT scans.

In comparison to CT scans, 2D chest X-rays are a more accessible and cost-effective option due to their fast imaging speed, low radiation, and low cost, especially during the early stages of the pandemic [21]. For example, the ChestX-ray dataset [18] contains about 112,120 chest X-rays used to classify common thoracic diseases. ChestXR dataset [1] contains 17,955 chest X-rays used for COVID-19 recognition. We advocate using chest X-ray datasets such as ChestX-ray and ChestXR may benefit COVID-19 segmentation using CT scans because of three reasons: (1) supplement limited CT data and contribute to training a more accurate segmentation model; (2) provide large-scale chest X-rays with labeled features, including pneumonia, thus can help the segmentation model to recognize patterns and features specific to COVID-19 infections; and (3) help improve the generalization of the segmentation model by enabling it to learn from different populations and imaging facilities. Inspired by this, in this study, we propose a new learning paradigm for COVID-19 segmentation using CT scans, involving training the segmentation model using limited CT scans with pixel-wise annotations and unpaired chest X-ray images with image-level labels.

To achieve this, an intuitive solution is building independent networks to learn features from each modality initially. Afterward, late feature fusion, co-attention or cross-attention modules are incorporated to transfer knowledge between CT and X-ray [12,13,22,23]. However, this solution faces two limitations. First, building modality-specific networks may cause insufficient interaction between CT and X-ray, limiting the model's ability to integrate information effectively. Although "Chilopod"-shaped multi-modal learning [6] has been

proposed to share all CNN kernels across modalities, it is still limited when the different modalities have a significant dimension gap. Second, the presence of unpaired data, specifically CT and X-ray data, in the feature fusion/cross-attention interaction can potentially cause the model to learn incorrect or irrelevant information due to the possible differences in their image distributions and objectives, leading to reduced COVID-19 segmentation accuracy. It's worth noting that the method using paired multimodal data [2] is not suitable for our application scenario, and the latest unpaired cross-modal [3] requires pixel-level annotations for both modalities, while our method can use X-ray images with image-level labels for training.

This paper proposes a novel Unpaired Cross-modal Interaction (UCI) learning framework for COVID-19 segmentation, which aims to learn strong representations from limited dense annotated CT scans and abundant image-level annotated X-ray images. The UCI framework learns representations from both segmentation and classification tasks. It includes three main components: a multi-modal encoder for image representations, a knowledge condensation and interaction module for unpaired cross-modal data, and task-specific networks. The encoder contains modality-specific patch embeddings and shared Transformer layers. This design enables the network to capture optimal feature representations for both CT and X-ray images while maintaining the ability to learn shared representations between the two modalities despite dimensional differences. To address the challenge of information interaction between unpaired cross-modal data, we introduce a momentum-updated prototype learning strategy to condense modality-specific knowledge. This strategy groups similar representations into the same prototype and iteratively updates the prototypes with a momentum term to capture essential information in each modality. Therewith, a knowledge-guided interaction module is developed that accepts the learned prototypes, enabling the UCI to better capture critical features and relationships between the two modalities. Finally, the task-specific networks, including the segmentation decoder and classification head, are presented to learn from all available labels. The proposed UCI framework has significantly improved performance on the public COVID-19 segmentation benchmark [15], thanks to the inclusion of chest X-rays.

The main contributions of this paper are three-fold: (1) we are the first to employ abundant X-ray images with image-level annotations to improve COVID-19 segmentation on limited CT scans, where the CT and X-ray data are unpaired and have potential distributional differences; (2) we introduce the knowledge condensation and interaction module, in which the momentum-updated prototype learning is offered to concentrate modality-specific knowledge, and a knowledge-guided interaction module is proposed to harness the learned knowledge for boosting the representations of each modality; and (3) our experimental results demonstrate our UCI learning method's effectiveness and strong generalizability in COVID-19 segmentation and the potential for related disease screening. This suggests that the proposed framework can be a valuable tool for medical practitioners in detecting and identifying COVID-19 and other associated diseases.

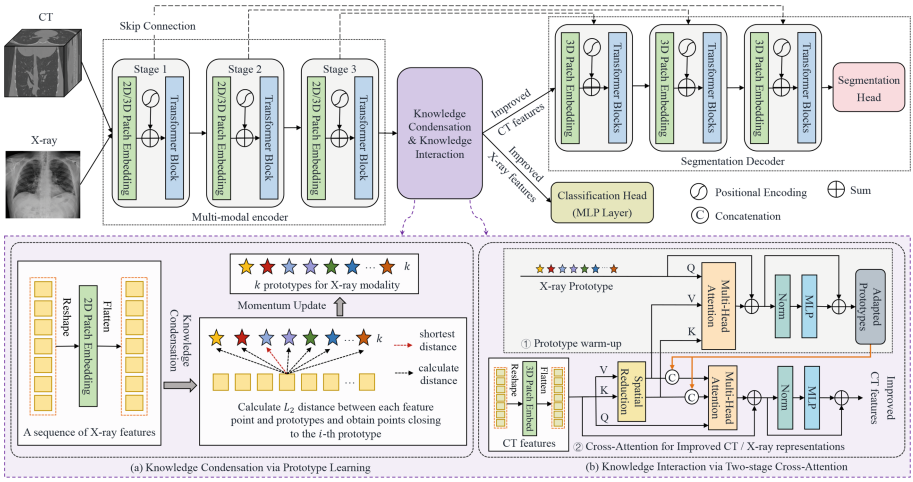

Fig. 1. Illustration of the proposed UCI learning framework.

2 Approach

The proposed UCI aims to explore effective representations for COVID-19 segmentation by leveraging both limited dense annotated CT scans and abundant image-level annotated X-rays. Figure 1 illustrates the three primary components of the UCI framework: a multi-modal encoder used to extract features from each modality, the knowledge condensation and interaction module used to model unpaired cross-modal dependencies, and task-specific heads designed for segmentation and classification purposes.

2.1 Multi-modal Encoder

The multi-modal encoder $\mathcal{F}(\cdot)$ consists of three stages of blocks, with modality-specific patch embedding layers and shared Transformer layers in each block, capturing modality-specific and shared patterns, which can be more robust and discriminative across modalities. Notice that due to the dimensional gap between CT and X-ray, we use the 2D convolution block as patch embedding for X-rays and the 3D convolution block as patch embedding for CTs. In each stage, the patch embedding layers down-sample the inputs and generate the sequence of modality-specific embedded tokens. The resultant tokens, combined with the learnable positional embedding, are fed into the shared Transformer layers for long-term dependency modeling and learning the common patterns. More details about architecture can be found in the Appendix.

Given a CT volume $\boldsymbol{x}^{ct}$, and a chest X-ray image $\boldsymbol{x}^{cxr}$, we denote the output feature sequence of the multi-modal encoder as

$$\begin{aligned} \boldsymbol{f}^{ct} &= \mathcal{F}(\boldsymbol{x}^{ct}; 3D) \in \mathbb{R}^{C^{ct} \times N^{ct}}, \\ \boldsymbol{f}^{cxr} &= \mathcal{F}(\boldsymbol{x}^{cxr}; 2D) \in \mathbb{R}^{C^{cxr} \times N^{cxr}} \end{aligned} \tag{1}$$

where C^{ct} and C^{cxr} represent the channels of CT and X-ray feature sequence. N^{ct} and N^{cxr} means the length of CT and X-ray feature sequence.

2.2 Knowledge Condensation and Interaction

Knowledge Condensation. It is difficult to directly learn cross-modal dependencies using the features obtained by the encoder because CT and X-ray data were collected from different patients. This means that the data may not have a direct correspondence between two modalities, making it challenging to capture their relationship. As shown in Fig. 1(a), we design a knowledge condensation (KC) module by introducing a momentum-updated prototype learning strategy to condensate valuable knowledge in each modality from the learned features. For the X-ray modality, given its prototypes $\mathcal{P}^{cxr} = \{\boldsymbol{p}_1^{cxr}, \boldsymbol{p}_2^{cxr}, ..., \boldsymbol{p}_k^{cxr}\}$ initialized randomly and the feature sequence $\boldsymbol{f}^{cxr}$, KC module first reduces the spatial resolution of $\boldsymbol{f}^{cxr}$ and groups the reduced $\boldsymbol{f}^{cxr}$ into k prototypes by calculating the distance between each feature point and prototypes, shown as follows

$$C_i^{cxr} = \left\{ \boldsymbol{m} \in \sigma(\boldsymbol{f}^{cxr}) : i = \arg\min_j \left\| \boldsymbol{m}, \boldsymbol{p}_j^{cxr} \right\|^2 \right\} \tag{2}$$

where C_i^{cxr} suggests the feature points closing to the i-th prototype. $\sigma(\cdot)$ represents a linear projection to reduce the feature sequence length to relieve the computational burden. Then we introduce a momentum learning function to update the prototypes with C_i^{cxr}, which means that the updates at each iteration not only depend on the current C_i^{cxr} but also consider the direction and magnitude of the previous updates, defined as

$$p_i^{cxr} \leftarrow \lambda p_i^{cxr} + (1-\lambda)\frac{1}{C_i^{cxr}} \sum_{\boldsymbol{m} \in C_i^{cxr}} \boldsymbol{m}, \tag{3}$$

where λ is the momentum factor, which controls the influence of the previous update on the current update. Similarly, the prototypes $\mathcal{P}^{ct}$ for CT modality can be calculated and updated with the feature set $\boldsymbol{f}^{ct}$. The prototypes effectively integrate the informative features of each modality and can be considered modality-specific knowledge to improve the subsequent cross-modal interaction learning. The momentum term allows prototypes to move more smoothly and consistently towards the optimal position, even in the presence of noise or other factors that might cause the prototypes to fluctuate. This can result in a more stable learning process and more accurate prototypes, thus contributing to condensate the knowledge of each modality better.

Knowledge-Guided Interaction. The knowledge-guided interaction (KI) module is proposed for unpaired cross-modality learning, which accepts the learned prototypes from one modality and features from another modality as inputs. As shown in Fig. 1(b), the KI module contains two multi-head attention

(MHA) blocks. Take CT features $\boldsymbol{f}^{ct}$ and X-ray prototypes $\mathcal{P}^{cxr}$ as input example, the first block considers $\mathcal{P}^{cxr}$ as the query and reduced $\boldsymbol{f}^{ct}$ as the key and value of the attention. It embeds the X-ray prototypes through the calculated affinity map between $\boldsymbol{f}^{ct}$ and $\mathcal{P}^{cxr}$, resulting in the adapted prototype $\mathcal{P}^{cxr\prime}$. The first block can be seen as a warm-up to make the prototype adapt better to the features from another modality. The second block treats $\boldsymbol{f}^{ct}$ as the query and the concatenation of reduced $\boldsymbol{f}^{ct}$ and $\mathcal{P}^{cxr\prime}$ as the key and value, improving the $\boldsymbol{f}^{ct}$ through the adapted prototypes. Similarly, for the $\boldsymbol{f}^{cxr}$ and $\mathcal{P}^{ct}$ as inputs, the KI module is also used to boost the X-ray representations. Inspired by the knowledge prototypes, KI modules boost the interaction between the two modalities and allow for the learning of strong representations for COVID-19 segmentation and X-ray classification tasks.

2.3 Task-Specific Networks

The outputs of the KI module are fed into two multi-task heads - one decoder for segmentation and one prediction head for classification respectively. The segmentation decoder has a symmetric structure with the encoder, consisting of three stages. In each stage, the input feature map is first up-sampled by the 3D patch embedding layer, and then refined by the stacked Transformer layers. Besides, we also add skip connections between the encoder and decoder to keep more low-level but high-resolution information. The decoder includes a segmentation head for final prediction. This head includes a transposed convolutional layer, a Conv-IN-LeakyReLU, and a convolutional layer with a kernel size of 1 and the output channel as the number of classes. The classification head contains a linear layer with the output channel as the number of classes for prediction. We use the deep supervision strategy by adding auxiliary segmentation losses (*i.e.*, the sum of the Dice loss and cross-entropy loss) to the decoder at different scales. The cross-entropy loss is used to optimize the classification task.

3 Experiment

3.1 Materials

We used the public COVID-19 segmentation benchmark [15] to verify the proposed UCI. It is collected from two public resources [5,8] on chest CT images available on The Cancer Imaging Archive (TCIA) [4]. All CT images were acquired without intravenous contrast enhancement from patients with positive Reverse Transcription Polymerase Chain Reaction (RT-PCR) for SARS-CoV-2. In total, we used 199 CT images including 149 training images and 50 test images. We also used two chest x-ray-based classification datasets including ChestX-ray14 [18] and ChestXR [1] to assist the UCI training. The ChestX-ray14 dataset comprises 112,120 X-ray images showing positive cases from 30,805 patients, encompassing 14 disease image labels pertaining to thoracic and lung ailments. An image may contain multiple or no labels. The ChestXR dataset consists of 21,390 samples, with each sample classified as healthy, pneumonia, or COVID-19.

3.2 Implementation Details

For CT data, we first truncated the HU values of each scan using the range of $[-958, 327]$ to filter irrelevant regions, and then normalized truncated voxel values by subtracting 82.92 and dividing by 136.97. We randomly cropped sub-volumes of size $32 \times 256 \times 256$ as the input and employed the online data augmentation like [10] to diversify the CT training set. For chest X-ray data, we set the size of input patches to 224×224. We employ the online data argumentation, including random cropping and zooming, random rotation, and horizontal/vertical flip, to enlarge the X-ray training dataset. We follow the extension of [20] for weight initialization and use the AdamW optimizer [11] and empirically set the initial learning rate to 0.0001, batch size to 2 and 32 for segmentation and classification, maximum iterations to 25w, momentum factor λ to 0.99, and the number of prototypes k to 256.

To evaluate the COVID-19 segmentation performance, we utilized six metrics, including the Dice similarity coefficient (DSC), intersection over union (IoU), sensitivity (SEN), specificity (SPE), Hausdorff distance (HD), and average surface distance (ASD). These metrics provide a comprehensive assessment of the segmentation quality. The overlap-based metrics, namely DSC, IoU, SEN, and SPE, range from 0 to 1, with a higher score indicating better performance. On the other hand, HD and ASD are shape distance-based metrics that measure the dissimilarity between the surfaces or boundaries of the segmentation output and the ground truth. For HD and ASD, a lower value indicates better segmentation results.

3.3 Compared with Advanced Segmentation Approaches

Table 1 gives the performance of our models and four advanced competing ones, including nnUNet [10], CoTr [19], nnformer [24], and Swin UNETR [9] in COVID-19 lesion segmentation. The results demonstrate that our UCI, which utilizes inexpensive chest X-rays, outperforms all other methods consistently and significantly, as evidenced by higher Dice and IoU scores. This suggests that the segmentation outcomes generated by our models are in good agreement with the ground truth. Notably, despite ChestXR being more focused on COVID-19 recognition, the UCI model aided by the ChestX-ray14 dataset containing 80k images performs better than the UCI model using the ChestXR dataset with only 16k images. This suggests that having a larger auxiliary dataset can improve the segmentation performance even if it is not directly related to the target task. The results also further prove the effectiveness of using a wealth of chest X-rays to assist the COVID-19 segmentation under limited CTs. Finally, our UCI significantly reduces HD and ASD values compared to competing approaches. This reduction demonstrates that our segmentation results provide highly accurate boundaries that closely match the ground-truth boundaries.

Table 1. Quantitative results of advanced segmentation approaches on the test set. '16k' and '80k' mean the number of auxiliary Chest X-rays during training.

Methods	DSC↑	IoU↑	SEN↑	SPE↑	HD↓	ASD↓
nnUNet [10]	0.6794	0.5404	0.7661	0.9981	132.5493	31.2794
CoTr [19]	0.6668	0.5265	0.7494	0.9984	118.0828	29.2167
nnFormer [24]	0.6649	0.5250	**0.7696**	0.9980	136.6311	34.9980
Swin UNETR [9]	0.5726	0.4279	0.6230	0.9784	155.8780	46.7789
UCI with ChestXR (16k)	0.6825	0.5424	0.7388	0.9984	132.1020	29.3694
UCI with ChestX-ray14 (80k)	**0.6922**	**0.5524**	0.7308	**0.9987**	**81.1366**	**16.6171**

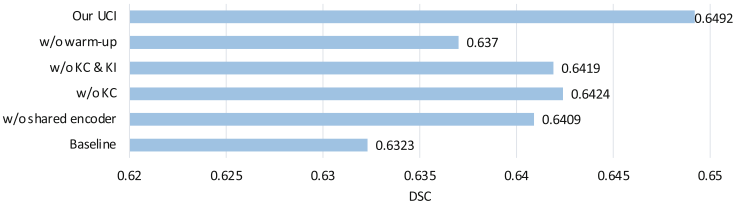

Fig. 2. Effectiveness of each module in UCI.

3.4 Discussions

Ablations. We perform ablation studies over each component of UCI, including the multi-modal encoder, Knowledge Condensation (KC) and Knowledge Interaction (KI) models, as listed in Fig. 2. We set the maximum iterations to 8w and use ChestX-ray14 as auxiliary data for all ablation experiments. We compare five variants of our UCI: (1) baseline: trained solely on densely annotated CT images; (2) w/o shared encoder: replacing the multi-modal encoder with two independent encoders, each designed to learn features from a separate modality; (3) w/o KC: removing the prototype and using the features before KC for interaction; (4) w/o KC & KI: only with encoder to share multi-modal information; and (5) w/o warm-up: removing the prototype warm-up in KI. Figure 2 reveals several noteworthy conclusions. Firstly, our UCI model, which jointly uses Chest X-rays, outperforms the baseline segmentation results by up to 1.69%, highlighting the effectiveness of using cheap large-scale auxiliary images. Secondly, using only a shared encoder for multi-modal learning (UCI w/o KC & KI) can still bring a segmentation gain of 0.96%, and the multi-modal encoder outperforms building independent modality-specific networks (UCI w/o shared encoder), underscoring the importance of shared networks. Finally, our results demonstrate the effectiveness of the prototype learning and prototype warm-up steps.

Hyper-Parameter Settings. To evaluate the impact of hyper-parameter settings on COVID-19 segmentation, we conducted an investigation of the number of prototypes (k) and the number of momentum factors (λ). Figure 3 illustrates

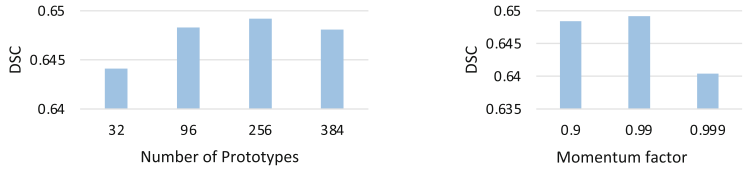

Fig. 3. Dice scores of UCI versus **Left**: the number of prototypes k and **right** the number of momentum factors λ.

the Dice scores obtained on the test set for different values of k and λ, providing insights into the optimal settings for these hyper-parameters.

4 Conclusion

Our study introduces UCI, a novel method for improving COVID-19 segmentation under limited CT images by leveraging unpaired X-ray images with image-level annotations. Especially, UCI includes a multi-modal shared encoder to capture optimal feature representations for CT and X-ray images while also learning shared representations between the two modalities. To address the challenge of information interaction between unpaired cross-modal data, UCI further develops a KC and KI module to condense modality-specific knowledge and facilitates cross-modal interaction, thereby enhancing segmentation training. Our experiments demonstrate that the UCI method outperforms existing segmentation models for COVID-19 segmentation.

Acknowledgment. This work was supported in part by the Ningbo Clinical Research Center for Medical Imaging under Grant 2021L003 (Open Project 2022LYKFZD06), in part by the Natural Science Foundation of Ningbo City, China, under Grant 2021J052, and in part by the National Natural Science Foundation of China under Grant 62171377.

References

1. Akhloufi, M.A., Chetoui, M.: Chest XR COVID-19 detection (2021). https://cxr-covid19.grand-challenge.org/. Accessed Sept 2021
2. Cao, X., Yang, J., Wang, L., Xue, Z., Wang, Q., Shen, D.: Deep learning based inter-modality image registration supervised by intra-modality similarity. In: Shi, Y., Suk, H.-I., Liu, M. (eds.) MLMI 2018. LNCS, vol. 11046, pp. 55–63. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00919-9_7
3. Chen, X., Zhou, H.Y., Liu, F., Guo, J., Wang, L., Yu, Y.: Mass: modality-collaborative semi-supervised segmentation by exploiting cross-modal consistency from unpaired ct and mri images. Med. Image Anal. **80**, 102506 (2022)
4. Clark, K., et al.: The cancer imaging archive (tcia): maintaining and operating a public information repository. J. Digit. Imaging **26**, 1045–1057 (2013)
5. Desai, S., et al.: Chest imaging representing a covid-19 positive rural us population. Sci. Data **7**(1), 414 (2020)

6. Dou, Q., Liu, Q., Heng, P.A., Glocker, B.: Unpaired multi-modal segmentation via knowledge distillation. IEEE Trans. Med. Imaging **39**(7), 2415–2425 (2020)
7. Fan, D.P., et al.: Inf-net: automatic covid-19 lung infection segmentation from ct images. IEEE Trans. Med. Imaging **39**(8), 2626–2637 (2020)
8. Harmon, S.A., et al.: Artificial intelligence for the detection of covid-19 pneumonia on chest ct using multinational datasets. Nat. Commun. **11**(1), 4080 (2020)
9. Hatamizadeh, A., Nath, V., Tang, Y., Yang, D., Roth, H.R., Xu, D.: Swin unetr: swin transformers for semantic segmentation of brain tumors in mri images. In: Crimi, A., Bakas, S. (eds) Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 7th International Workshop, BrainLes 2021, Held in Conjunction with MICCAI 2021, Virtual Event, 27 September 2021, Revised Selected Papers, Part I, pp. 272–284. Springer, Heidelberg (2022). https://doi.org/10.1007/978-3-031-08999-2_22
10. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. Nat. Methods **18**(2), 203–211 (2021)
11. Loshchilov, I., Hutter, F.: Fixing weight decay regularization in adam (2018)
12. Lyu, J., Sui, B., Wang, C., Tian, Y., Dou, Q., Qin, J.: Dudocaf: dual-domain cross-attention fusion with recurrent transformer for fast multi-contrast mr imaging. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) Medical Image Computing and Computer Assisted Intervention-MICCAI 2022: 25th International Conference, Singapore, 18–22 September 2022, Proceedings, Part VI, pp. 474–484. Springer, Heidelberg (2022). DOI: https://doi.org/10.1007/978-3-031-16446-0_45
13. Mo, S., et al.: Multimodal priors guided segmentation of liver lesions in MRI using mutual information based graph co-attention networks. In: Martel, A.L., et al. (eds.) MICCAI 2020. LNCS, vol. 12264, pp. 429–438. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59719-1_42
14. Qiu, Y., Liu, Y., Li, S., Xu, J.: Miniseg: an extremely minimum network for efficient covid-19 segmentation. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, pp. 4846–4854 (2021)
15. Roth, H.R., et al.: Rapid artificial intelligence solutions in a pandemic-the covid-19-20 lung ct lesion segmentation challenge. Med. Image Anal. **82**, 102605 (2022)
16. Shi, F., et al.: Review of artificial intelligence techniques in imaging data acquisition, segmentation, and diagnosis for covid-19. IEEE Rev. Biomed. Eng. **14**, 4–15 (2020)
17. Wang, G., et al.: A noise-robust framework for automatic segmentation of covid-19 pneumonia lesions from ct images. IEEE Trans. Med. Imaging **39**(8), 2653–2663 (2020)
18. Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., Summers, R.M.: Chestx-ray8: hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2097–2106 (2017)
19. Xie, Y., Zhang, J., Shen, C., Xia, Y.: CoTr: efficiently bridging CNN and transformer for 3D medical image segmentation. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12903, pp. 171–180. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87199-4_16
20. Xie, Y., Zhang, J., Xia, Y., Wu, Q.: Unimiss: universal medical self-supervised learning via breaking dimensionality barrier. In: Avidan, S., Brostow, G., Cisse, M., Farinella, G.M., Hassner, T. (eds.) ECCV 2022. LNCS, vol. 13681, pp. 558–575. Springer, Heidelberg (2022). https://doi.org/10.1007/978-3-031-19803-8_33

21. Zhang, J., et al.: Viral pneumonia screening on chest x-rays using confidence-aware anomaly detection. IEEE Trans. Med. Imaging **40**(3), 879–890 (2020)
22. Zhang, Y., He, N., Yang, J., Li, Y., Wei, D., Huang, Y., Zhang, Y., He, Z., Zheng, Y.: mmformer: Multimodal medical transformer for incomplete multimodal learning of brain tumor segmentation. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) Medical Image Computing and Computer Assisted Intervention-MICCAI 2022: 25th International Conference, Singapore, 18–22 September 2022, Proceedings, Part V, pp. 107–117. Springer, Heidelberg (2022). https://doi.org/10.1007/978-3-031-16443-9_11
23. Zhang, Y., et al.: Modality-aware mutual learning for multi-modal medical image segmentation. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12901, pp. 589–599. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87193-2_56
24. Zhou, H.Y., Guo, J., Zhang, Y., Yu, L., Wang, L., Yu, Y.: nnformer: interleaved transformer for volumetric segmentation. arXiv preprint arXiv:2109.03201 (2021)