Appendix

1. Appendix

1.1. Datasets

BTCV/CT. The BTCV dataset[1] comprises 30 participants who had abdominal CT scans with 13 organs annotated by interpreters at Vanderbilt University Medical Center under the supervision of clinical radiologists. Each CT scan was performed in the portal venous phase with contrast enhancement and comprised of 80 to 225 slices with 512×512 pixels and a slice thickness of 1 to 6 mm. Each volume was pre-processed separately, with intensities in the range of [-175, 200] HU being normalized to [0, 1]. During preprocessing, all images are resampled to 1.5/2.0 mm (different resolutions for ablation study) isotropic voxel spacing. The multi-organ segmentation problem is formulated as a 13 class segmentation task with 1-channel input. We use the first 24 volumes for training and report on 6 validation images.

BraTS/MRI. Brain tumor segmentation is another important task. This BraTS dataset [3] contains a training set of 387 multi-modal multi-site MRI data (FLAIR, T1w, T1gd, T2w) with ground truth labels of gliomas segmentation necrotic/active tumor and edema is used for brain tumor segmentation. The voxel spacing of MRI images in this task is $1.0\times1.0\times1.0$ mm^3. The voxel intensities are pre-processed with z-score normalization. The problem of brain tumor segmentation is formulated as a 3 class segmentation task with 4-channel input. We report on 97 validation images.

TCIA-COVID19/CT. This is a public dataset [1] consisting of unenhanced chest CTs of patients with COVID19 infections. There are 771 volumes collected from 661 patients in total. All images are unannotated. We utilize this dataset as an extra unlabeled dataset for self-supervised learning in ablation study. All models in Tab. 1 are pretrained using a combination of this dataset and BTCV. In ablation study, we also compare the performance between pretraining with and without this dataset.

1.2. Experimental Settings

As can be seen from the reconstructed volumes, the large model has more restoration power than the tiny model, which supports the previous conclusion. The flattened dimensionality of 3D medical images is frequently very high, and a small model would unavoidably compress the original voxel space into a smaller voxel space, thereby losing a lot of information.

config	value
optimizer	AdamW[2]
base learning rate	3e-4
weight decay	0.005
optimizer momentum	beta1, beta2 = 0.9, 0.999
batch size	4
learning rate schedule	linear warmup cosine annealing
warmup epochs	300
total epochs	3000
augmentation	RangeScaleIntensity

Table 1: Sup. baseline setting for BTCV.

config	value
optimizer	AdamW[2]
base learning rate	3e-4
weight decay	0.005
optimizer momentum	beta1, beta2 = 0.9, 0.999
batch size	2
learning rate schedule	linear warmup cosine annealing
warmup epochs	100
total epochs	1000
augmentation	NormalizedIntensity

Table 2: Sup. baseline setting for BraTS.

config	value
optimizer	AdamW[2]
base learning rate	3e-4
weight decay	0.005
optimizer momentum	beta1, beta2 = 0.9, 0.999
batch size	4
learning rate schedule	linear warmup cosine annealing
warmup epochs	300
total epochs	3000
augmentation	RangeScaleIntensity

Table 3: Pretraining on CT 3D volumes.

[1] https://www.synapse.org/#!Synapse:syn3193805/wiki/89480

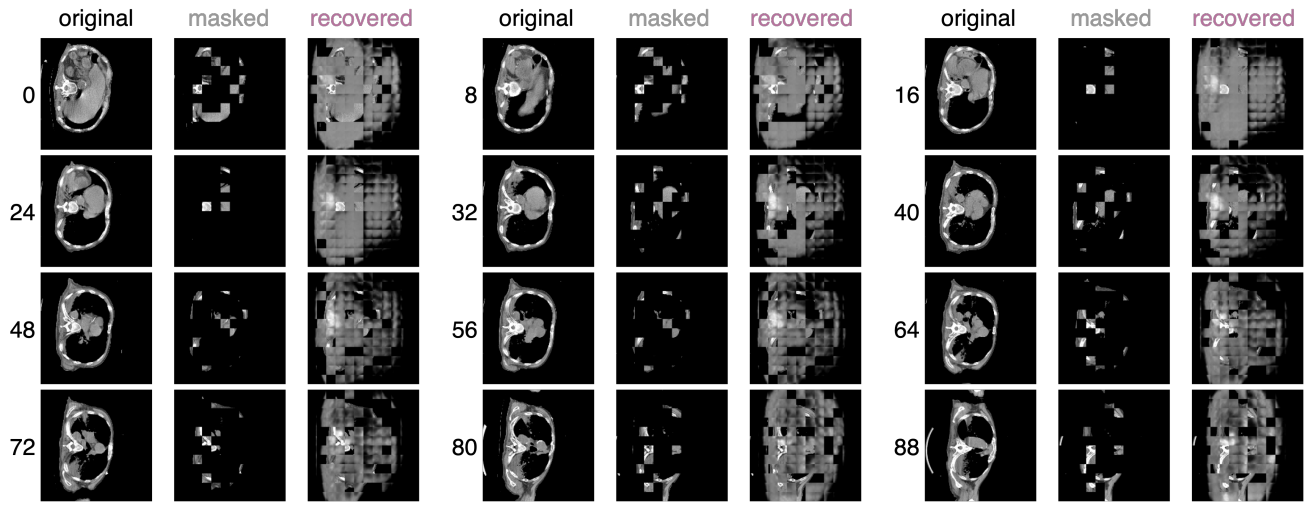

Figure 1: Example results of one CT scan from TCIA-COVID19. As the original images are all 3D volumes, we show the reconstructed images in a form of slices, where the indexing number represents the depth. For each triplet, we show the ground truth (left), the masked image (middle), and the SimMIM [4] reconstruction (right). In this case, a ViT-Base backbone is applied for the encoder, the masked patch size is 16 (for all dimensions), and the masking ratio is 75% following [4].

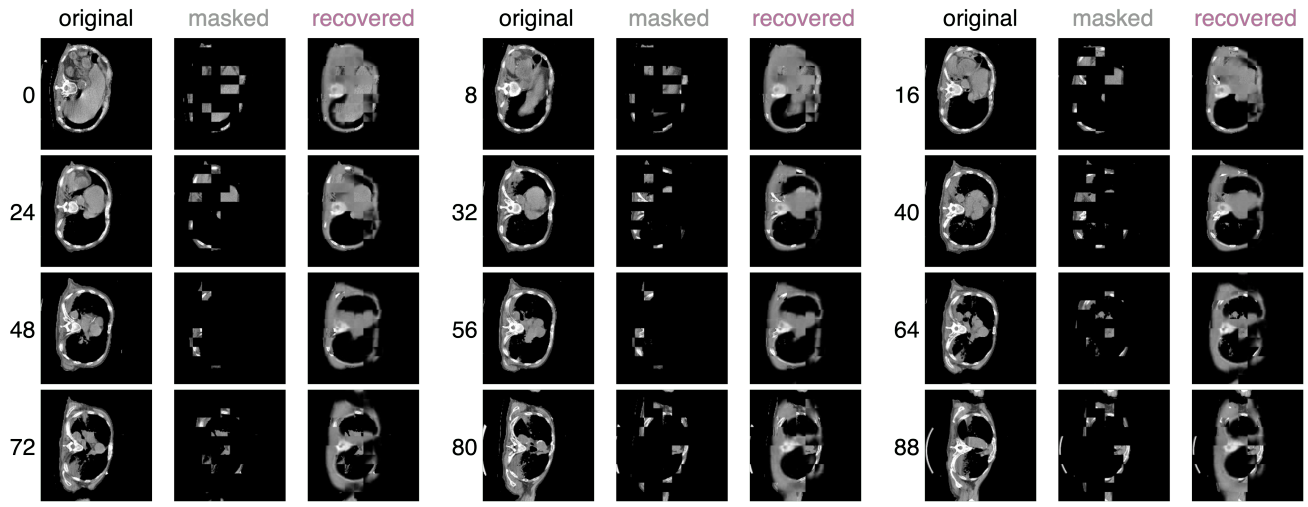

Figure 2: Example the same image as Fig. 1. As the original images are all 3D volumes, we show the reconstructed images in a form of slices, where the indexing number represents the depth. For each triplet, we show the ground truth (left), the masked image (middle), and the SimMIM [4] reconstruction (right). In this case, a ViT-Large backbone is applied for the encoder. All rest settings are consistent with Fig. 1.

References

[1] Stephanie A. Harmon, Thomas Sanford, Sheng Xu, Evrim B Turkbey, Holger R. Roth, Ziyue Xu, Dong Yang, Andriy Myronenko, Victoria L. Anderson, Amel Amalou, Maxime Blain, Michael T Kassin, Dilara Long, Nicole Varble, Stephanie M. Walker, Ulas Bagci, Anna Maria Ierardi, Elvira Stellato, Guido Giovanni Plensich, Giuseppe Franceschelli, Cristiano Girlando, Giovanni Irmici, Dominic LaBella, Dima A. Hammoud, Ashkan A. Malayeri, Elizabeth C. Jones, Ronald M. Summers, Peter L. Choyke, Daguang Xu, Mona G. Flores, Kaku Tamura, Hirofumi Obinata, Hitoshi Mori, F. Patella, Maurizio Cariati, Gianpaolo Carrafiello, Peng An, Bradford J. Wood, and Baris Turkbey. Artificial intelligence for the detection of covid-19 pneumonia on chest ct using multinational datasets. *Nature Communications*, 11, 2020. 1

[2] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019. 1, 3

[3] Amber L. Simpson, Michela Antonelli, Spyridon Bakas, Michel Bilello, Keyvan Farahani, Bram van Ginneken, An-

config	value
optimizer	AdamW[2]
base learning rate	3e-4
weight decay	0.005
optimizer momentum	beta1, beta2 = 0.9, 0.999
batch size	2
learning rate schedule	linear warmup cosine annealing
warmup epochs	100
total epochs	1000
augmentation	RangeScaleIntensity

Table 4: Pretraining setting on MRI 3D volumes.

nette Kopp-Schneider, Bennett A. Landman, Geert J. S. Litjens, Bjoern H. Menze, Olaf Ronneberger, Ronald M. Summers, Patrick Bilic, Patrick Ferdinand Christ, Richard Kinh Gian Do, Marc J. Gollub, Jennifer Golia-Pernicka, Stephan Heckers, William R. Jarnagin, Maureen McHugo, Sandy Napel, Eugene Vorontsov, Lena Maier-Hein, and M. Jorge Cardoso. A large annotated medical image dataset for the development and evaluation of segmentation algorithms. *ArXiv*, abs/1902.09063, 2019. 1

[4] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. Simmim: A simple framework for masked image modeling. *ArXiv*, abs/2111.09886, 2021. 2