VoCo: A Simple-yet-Effective Volume Contrastive Learning Framework for 3D Medical Image Analysis (Supplementary Materials)

Dataset	Modality	Task	Train	Valid.
Pre-training				
BTCV [10]	CT	pre-train	24	6
TCIA Covid19 [5]	CT	pre-train	722	49
LUNA [13]	CT	pre-train	843	45
Downstream				
BTCV [10]	CT	segmentation	24	6
LiTs [2]	CT	segmentation	100	31
MSD Spleen [1]	CT	segmentation	32	9
MM-WHS [24]	CT	segmentation	14	6
BraTs [14]	MRI	segmentation	387	97
CC-CCII [18]	CT	classification	2514	1664

Table 1. The details of pre-training and downstream datasets.

In the supplementary materials, we first introduce the pre-training and downstream datasets we use in our experiments. Then, we present the implementation details of VoCo, including the settings of pre-processing, pre-training, and finetuning. Finally, **additional experiments** are presented, including ablation studies and experiments on **2D medical dataset** [16].

1. Datasets

Pre-training and downstream datasets. The details of pre-training and downstream datasets are shown in Table 1. Specifically, we use BTCV [10] and TCIA Covid19 [5] totally about 0.8k CT scans for BTCV [10] downstream task, which aims to conduct fair comparison with previous works [4, 23]. And we further combine LUNA [13] to scale the size of pre-training datasets to 1.6k for the other four downstream tasks.

BTCV dataset. BTCV [10] dataset contains one background class and thirteen organ classes, *i.e.*, spleen, right kidney, left kidney, gallbladder, esophagus, liver, stomach, aorta, inferior vena cava, portal and splenic veins, pancreas, left and right adrenal glands. Following the previous works [4, 23, 22, 15], we split BTCV [10] dataset into 24 scans for training and 6 scans for validation. It is worth noting that the BTCV [10] dataset is used in pre-training.

LiTs dataset. LiTs [2] dataset releases 131 abdominal CT Volumes and associated annotations for training and validation. There are two types of labels in LiTs [2]: the liver and tumor. Following previous works [20, 21, 17], in this paper, we only utilize the ground truth masks of the liver to evaluate the effectiveness of various SSL algorithms.

Pre-process settings	
Spacing	[1.5, 1.5, 1.5]
Norm $[a_{min}, a_{max}]$	[-175.0, 250.0]
Norm $[b_{min}, b_{max}]$	[0.0, 1.0]
Roi-Size	64×64×64
Augmentation	Random rotate and flip
Pre-training settings	
Pre-training steps	100k
Optimizer	AdamW
Optimization LR	1e-3
LR schedule	warmup cosine
Warmup steps	100
Momentum	0.9
Regularization weight	1e-2
Batch size	4
Sw batch size	4
VoCo Resize	384×384×64
Resize after crop	64×64×64
VoCo n	4×4
VoCo λ	1.0
Finetuning settings	
Optimizer	AdamW
Optimization LR	3e-4
LR schedule	warmup cosine
Warmup steps	100
Momentum	0.9
Regularization weight	1e-5
Batch size	1
Sw batch size	4
Inference	sliding window
ROI size	96×96×96

Table 2. Pre-process and training settings in the experiments.

MSD Spleen dataset. MSD Spleen dataset is the 9_{th} challenge in MSD [1], which is developed for spleen segmentation. Specifically, aiming to conduct fair comparisons with previous state-of-the-art methods [9, 4, 23], we use 32 scans for training and 9 scans for validation, as shown in Table 1.

MM-WHS dataset. MM-WHS [24] dataset is also unseen in the pre-training, which contains 7 classes including Left Ventricle, whole aorta, Right Ventricle, Left Atrium, myocardium of Left Ventricle, Right Atrium, and Pulmonary Artery. The data splits are also shown in Table 1.

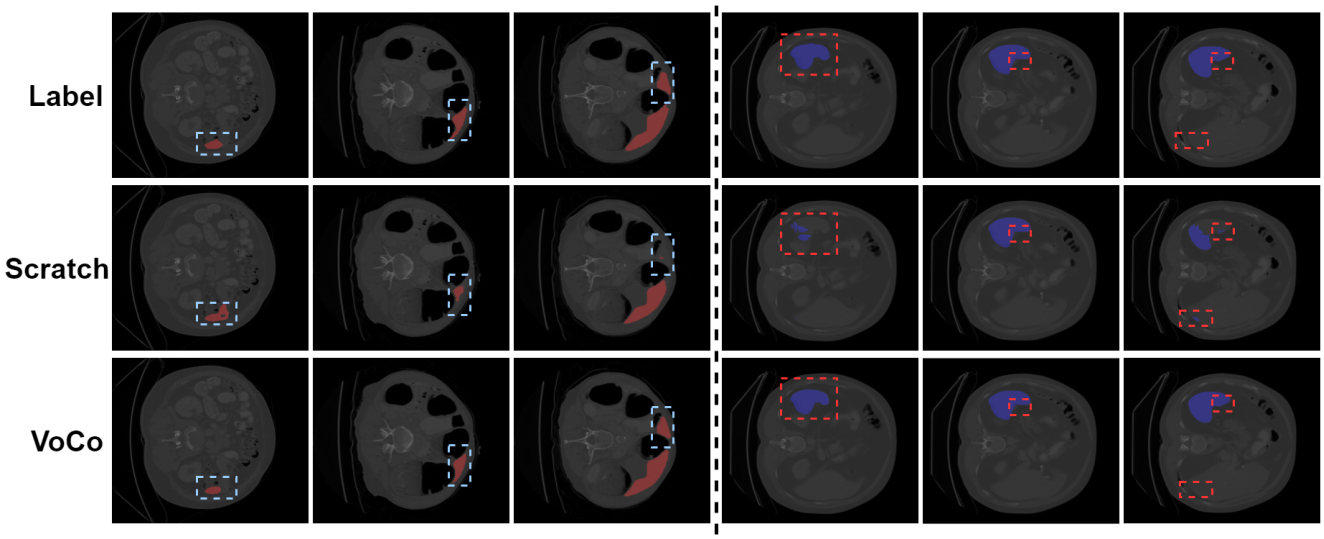

Figure 1. Qualitative visualization of segmentation results for the LiTS [2] and MSD Spleen [1] datasets. Scratch represents the results of 'from scratch'. The obvious differences are highlighted by blue and red dashed boxes, respectively.

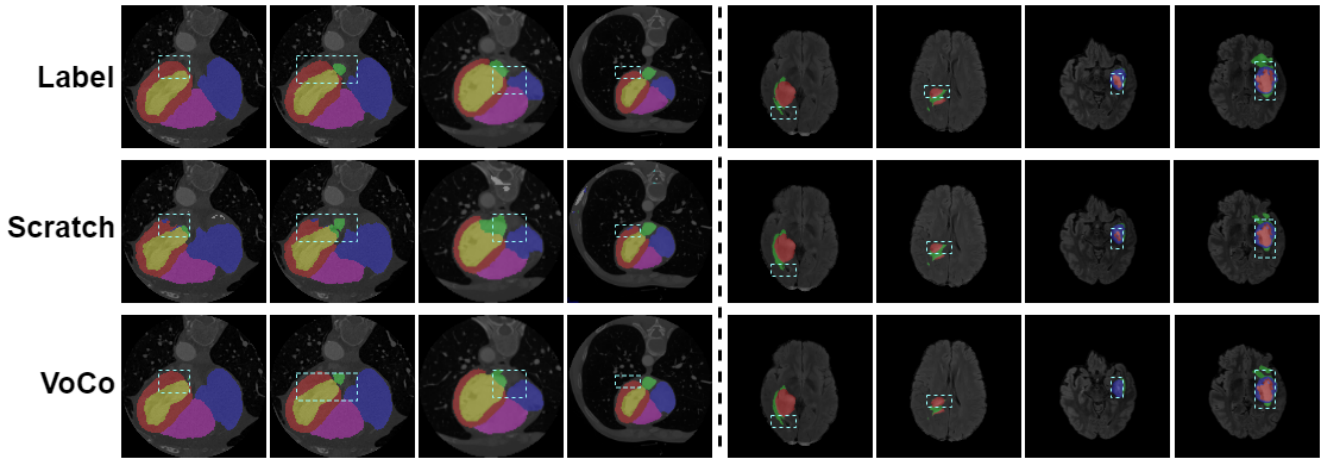

Figure 2. Qualitative visualization of segmentation results for the MM-WHS [24] and BraTs [14] datasets. Scratch represents the results of 'from scratch'. The obvious differences are highlighted by blue dashed boxes, respectively.

BraTs dataset. BraTs [14] dataset is an MRI dataset, which has been known as a series of challenges in brain tumor segmentation. In this paper, we evaluate the ability of model generalization on the BraTs [14] dataset, since we pre-train the model with only CT datasets. Specifically, we perform experiments on the released 387 scans of BraTS 2021 and evaluate the accuracy on the remained 97 scans. There are three classes in BraTS: whole tumor (WT), tumor core (TC), and enhancing tumor (ET).

CC-CCII dataset. CC-CCII [18] dataset is designed for COVID-19 detection, which can be seen as a classification task. CC-CCII [18] dataset contains 2516 scans for training and 1644 scans for validation, which includes three classes, *i.e.*, novel coronavirus pneumonia (NCP), common pneumonia (CP), and normal controls (Normal).

2. Implementation Details

Aiming to conduct fair comparisons with previous methods [21, 17, 9, 23, 8], we adopt comparatively consistent settings in the experiments. The details of pre-process and training settings are shown in Table 2. Our implementation is mainly based on the open-source platform Monai [1] and Pytorch [11]. We use one NVIDIA A100 GPU for all the experiments.

Fine-tuning on downstream datasets. The fine-tuning settings are almost consistent with the pre-training settings, except for the number of training epochs. Specifically, the training epochs are set to 3000, 1000, 1000, 1000, 500, and 100 for BTCV [10], LiTs [2], MSD Spleen [1], MM-

[1] https://monai.io/

λ	BTCV	MM-WHS
0.5	83.52	90.16
1.0	**83.85**	**90.54**
1.5	83.80	90.48

Table 3. Ablation studies of λ on BTCV [10] and MM-WHS [24].

Methods	NIH ChestX-ray
From scratch	75.4
MG [22]	77.3
TransVW [6]	77.6
C2L [19]	79.0
SimSiam [3]	79.4
PCRLv1 [20]	79.9
PCRLv2 [21]	81.5
VoCo	**82.02**

Table 4. Experimental results on the NIH ChestX-ray [16] dataset. The results are drawn from [21].

Organs	Dice Scores(%)
Left Ventricle	91.32
Whole aorta	91.30
Right Ventricle	94.64
Left Atrium	86.89
Myocardium of Left Ventricle	89.16
Right Atrium	96.35
Pulmonary Artery	84.13
Average	**90.54**

Table 5. Dice Scores of 7 organs on MM-WHS [24].

WHS [24], BraTs [14], and CC-CCII [18], respectively.

3. Experiments

We provide some experiments that are not presented in the main paper due to the limitation of pages, including ablation studies, 2D medical image analysis, and others.

3.1. Ablation Studies

We further evaluate the settings of the balance parameter λ for the loss functions, as shown in Table. 3. We also report the Dice Score on the BTCV [10] and MM-WHS [24] datasets for evaluation. We set λ as 0.5, 1.0, and 1.5 for ablation studies. As shown in Table. 3, we find that the settings of λ do not matter a lot. Thus, in VoCo, we consider the importance of loss functions equal and set λ as 1.

3.2. 2D Medical Image Analysis

In the main paper, we evaluate the effectiveness of VoCo on 3D medical images. To further verify its performance on 2D medical images, we also conduct experiments on the NIH ChestX-ray [16] datasets. We follow the consistent settings of previous works [20, 21], *i.e.*, pre-train on NIH ChestX-ray and fine-tune on NIH ChestX-ray. Specifically, for fair comparisons with [20, 21], 60% of data are used for pre-training and the remaining is used for finetuning. 3D-UNet [12] is used for experiments. As shown in Table 4, VoCo can also achieve competitive results on the 2D medical dataset. We conclude that although the 2D images contain less information than 3D scans, the position priors still exist, which benefits the training of VoCo.

3.3. Dice Scores of MM-WHS dataset

The Dice Scores of 7 organs on the MM-WHS [24] dataset are shown in Table 5.

3.4. Validation results on the leaderboard

We have verified the ***BTCV*** test results and further evaluated the test sets of ***Flare23*** and ***Amos22*** in the public leaderboard. **Note that** aiming to verify the pure effectiveness, we did not use model ensembling, extra data, or other tricks. We compare with the strong baseline SwinUNETR[7] (since with the same network and settings) in Table 6.

The MSD leaderboard has not been updated for a long time. Due to the rebuttal emergency, we provide the results of the offline validation set instead. We strictly follow the settings of SwinUNETR[7] and the results are shown in Table 7.

Method	BTCV	Flare23	Amos22 (DSC/NSD)
SwinUNETR[7]	†84.72	87.84	†88.00/76.15
VoCo	**86.44**	**90.07**	**89.06/78.86**

Table 6. **Online test results**. †: drawn from previous papers.

Method	Task1	Task2	Task3	Task4	Task5
SwinUNETR†[47]	75.13	95.89	81.72	91.98	80.23
VoCo	**76.26**	**96.93**	**84.98**	**92.09**	**82.16**
Method	**Task6**	**Task7**	**Task8**	**Task9**	**Task10**
SwinUNETR†[7]	63.46	64.32	70.54	94.63	44.57
VoCo	**67.74**	**67.85**	**70.92**	**96.34**	**45.17**

Table 7. MSD Decathlon. More results will be in the revision.

3.5. More Visualization Results

Visualization results on LiTs [2], MSD Spleen [1], MM-WHS [24], and BraTs [14] are shown in Fig. 1 and Fig. 2.

References

[1] Michela Antonelli et al. The medical segmentation decathlon. *Nature Commun.*, 13(1):4128, 2022. 1, 2, 3

[2] Patrick Bilic et al. The liver tumor segmentation benchmark (lits). *Medical Image Analy.*, 84:102680, 2023. 1, 2, 3

[3] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *CVPR*, pages 15750–15758, 2021. 3

[4] Zekai Chen et al. Masked image modeling advances 3d medical image analysis. In *WACV*, pages 1970–1980, 2023. 1

[5] Kenneth Clark and Bruceand others Vendt. The cancer imaging archive (tcia): maintaining and operating a public information repository. *Jour. of Dig. Imag.*, 26:1045–1057, 2013. 1

[6] Fatemeh Haghighi et al. Transferable visual words: Exploiting the semantics of anatomical patterns for self-supervised learning. *IEEE Trans. Medical Imag.*, 40(10):2857–2868, 2021. 3

[7] Ali Hatamizadeh et al. Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In *MICCAIW*, pages 272–284, 2021. 3

[8] Kaiming He et al. Masked autoencoders are scalable vision learners. In *CVPR*, pages 16000–16009, 2022. 2

[9] Yuting He et al. Geometric visual similarity learning in 3d medical image self-supervised pre-training. In *CVPR*, pages 9538–9547, 2023. 1, 2

[10] Bennett Landman et al. Miccai multi-atlas labeling beyond the cranial vault–workshop and challenge. In *MICCAIW*, volume 5, page 12, 2015. 1, 2, 3

[11] Adam Paszke et al. Pytorch: An imperative style, high-performance deep learning library. *NIPS*, 32, 2019. 2

[12] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, pages 234–241, 2015. 3

[13] Arnaud Arindra Adiyoso Setio et al. Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: the luna16 challenge. *Medical Image Analy.*, 42:1–13, 2017. 1

[14] Amber L Simpson et al. A large annotated medical image dataset for the development and evaluation of segmentation algorithms. *arXiv preprint arXiv:1902.09063*, 2019. 1, 2, 3

[15] Yucheng Tang et al. Self-supervised pre-training of swin transformers for 3d medical image analysis. In *CVPR*, pages 20730–20740, 2022. 1

[16] Xiaosong Wang et al. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *CVPR*, pages 2097–2106, 2017. 1, 3

[17] Chuyan Zhang, Hao Zheng, and Yun Gu. Dive into the details of self-supervised learning for medical image analysis. *Medical Image Anal.*, 89:102879, 2023. 1, 2

[18] Kang Zhang et al. Clinically applicable ai system for accurate diagnosis, quantitative measurements, and prognosis of covid-19 pneumonia using computed tomography. *Cell*, 181(6):1423–1433, 2020. 1, 2, 3

[19] Hong-Yu Zhou et al. Comparing to learn: Surpassing imagenet pretraining on radiographs by comparing image representations. In *MICCAI*, pages 398–407, 2020. 3

[20] Hong-Yu Zhou et al. Preservational learning improves self-supervised medical image models by reconstructing diverse contexts. In *ICCV*, pages 3499–3509, 2021. 1, 3

[21] Hong-Yu Zhou et al. A unified visual information preservation framework for self-supervised pre-training in medical image analysis. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2023. 1, 2, 3

[22] Zongwei Zhou et al. Models genesis. *Medical Image Analy.*, 67:101840, 2021. 1, 3

[23] Jia-Xin Zhuang, Luyang Luo, and Hao Chen. Advancing volumetric medical image segmentation via global-local masked autoencoder. *arXiv preprint arXiv:2306.08913*, 2023. 1, 2

[24] Xiahai Zhuang. Multivariate mixture model for myocardial segmentation combining multi-source images. *IEEE Trans. Pattern Analy. Mach. Intell.*, 41(12):2933–2946, 2018. 1, 2, 3