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

Linshan Wu Jiaxin Zhuang Hao Chen *
Hong Kong University of Science and Technology

2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) | 979-8-3503-5300-6/24/$31.00 ©2024 IEEE | DOI: 10.1109/CVPR52733.2024.02158

Abstract

*Self-Supervised Learning (SSL) has demonstrated promising results in 3D medical image analysis. However, the lack of high-level semantics in pre-training still heavily hinders the performance of downstream tasks. We observe that 3D medical images contain relatively consistent contextual position information, i.e., consistent geometric relations between different organs, which leads to a potential way for us to learn consistent semantic representations in pre-training. In this paper, we propose a simple-yet-effective **Vo**lume **Co**ntrast (**VoCo**) framework to leverage the contextual position priors for pre-training. Specifically, we first generate a group of base crops from different regions while enforcing feature discrepancy among them, where we employ them as class assignments of different regions. Then, we randomly crop sub-volumes and predict them belonging to which class (located at which region) by contrasting their similarity to different base crops, which can be seen as predicting contextual positions of different sub-volumes. Through this pretext task, VoCo implicitly encodes the contextual position priors into model representations without the guidance of annotations, enabling us to effectively improve the performance of downstream tasks that require high-level semantics. Extensive experimental results on six downstream tasks demonstrate the superior effectiveness of VoCo. Code will be available at https://github.com/Luffy03/VoCo.*

1. Introduction

Deep learning has demonstrated outstanding achievements in 3D medical image analysis [53, 21, 40, 33, 39], yet is heavily hampered by the expensive cost of the required expert annotations [50, 23]. To address this problem, Self-Supervised Learning (SSL) has received significant attention due to its promising ability to learn representations without annotations [10, 11, 6, 28, 20], which has become an important label-efficient solution in 3D medical image analysis [71, 51, 32, 2, 34, 36].

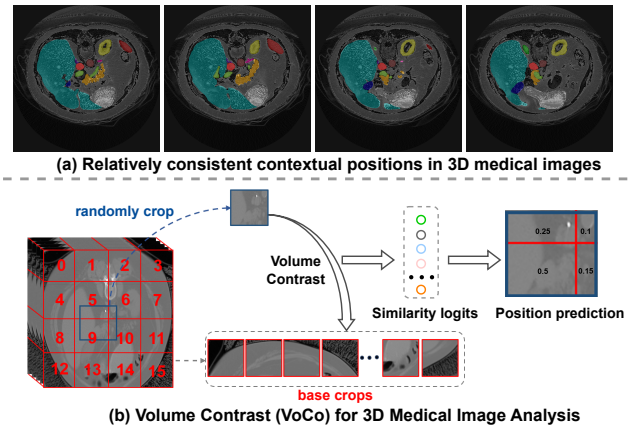

Figure 1. (a) In 3D medical images, the contextual positions, *i.e.*, the geometric relations between different organs are relatively consistent. (b) Aiming to leverage contextual position priors for pre-training, we proposed a Volume Contrast (VoCo) framework for 3D Medical Image Analysis.

Existing methods [50, 75, 71, 13] are mostly based on information reconstructions to learn augment-invariant representations of 3D medical images, which first employ strong data augmentation to the images and then reconstruct the raw information. Specifically, rotate-and-reconstruct [50, 51, 75, 52] proposed to randomly rotate the 3D volumetric images and learn to recover them, which encourages models to learn rotational invariant features. Recent methods [70, 71, 32, 25, 62] further proposed to restore information among different views of the image. PCRL [70, 71] cropped global and local patches then conducted multi-scale restorations. GVSL [32] further explored the geometric similarity between multi-scans by affine augmentation and matching. Mask-reconstruct methods [13, 73, 55] are also widely used, which are introduced from MAE [28] and aim to learn representations by masking images and reconstructing the missing pixels. Although promising results have been demonstrated, previous works [52, 32] have

*Corresponding author: jhc@cse.ust.hk

2575-7075/24/$31.00 ©2024 IEEE
DOI 10.1109/CVPR52733.2024.02158

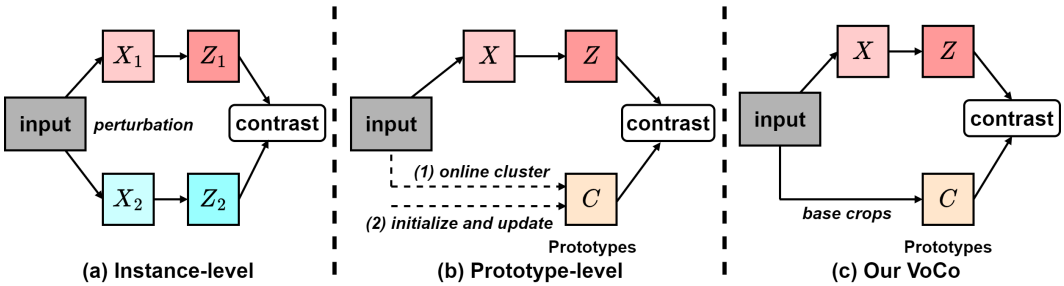

Figure 2. **Typical contrastive learning frameworks.** (a) Instance-level contrastive learning [10, 11, 29, 22, 7] employs strong data augmentation or model perturbation on input data to acquire different views of instance, then regularizes their consistency. (b) Prototype-level contrastive learning [5, 6, 56, 45, 15, 16] conducts (1) online clustering or (2) randomly initialize then online update process to obtain prototypes as class assignments, then leverage the prototypes to contrast each input image. (c) Our VoCo follows the idea of prototype-level contrastive learning. Specifically, instead of using time-consuming online clustering and updating procedures, we leverage the valuable contextual position priors of 3D medical images and leverage the base crops to generate prototypes (bases).

proved that the lack of high-level semantics in pre-training will heavily hinder the performance of downstream tasks. To address this challenge, we argue that stronger high-level semantics should be further involved into 3D medical image pre-training.

To this end, we argue that the contextual position priors of 3D medical images should be further exploited. As shown in Fig. 1(a), we observe that in 3D medical images, different organs (semantic regions) contain relatively consistent contextual positions with relatively consistent anatomic characteristics (shapes). Thus, the consistency of geometric relations between different organs leads to a potential way for us to learn consistent semantic representations for 3D medical images pre-training. In this paper, we propose a pretext task for contextual position predictions, which aims to encode contextual position priors into model representations and enables us to effectively improve the performance of downstream tasks that require high-level semantics.

In this paper, we propose a simple-yet-effective Volume Contrast (VoCo) framework for 3D medical image analysis, as shown in Fig. 1(b). Specifically, we first crop a group of non-overlap volumes from different positions while enforcing feature discrepancy among them. We represent these volumes as a group of bases in the learned high-dimension space, where we employ them as class assignments of different positions. Then, we randomly crop sub-volumes and predict them belonging to which class (located at which position) by contrasting their similarity to different bases, which can be seen as predicting contextual positions of different sub-volumes. In this way, we formulate a contextual position prediction pretext task for 3D medical image SSL. Through learning to predict contextual positions, we implicitly involve the high-level semantic priors into the model representations, which enables us to significantly improve the performance of downstream tasks. Extensive experimental results on six downstream tasks demonstrate that our proposed VoCo clearly outperforms existing state-of-the-art 3D medical image SSL methods.

2. Related Works

In this section, we first introduce the previous mainstream contrastive learning paradigms. Then, we survey the existing SSL methods for medical image analysis, especially for 3D medical images. Finally, we review the position-related SSL methods for comparisons with our method and highlight the differences.

Contrastive learning. Contrastive learning is one of the mainstream paradigms in SSL, which aims to learn consistent representations by contrasting positive and negative pairs of samples without extra annotations [10, 6, 29, 11]. According to [6], instance- and prototype-level contrastive learning are two typical types of contrastive learning, as shown in Fig. 2.

Instance-level contrastive learning [10, 11, 29, 22, 7] transforms input images with different augmentations or model perturbations, aiming to compare the features from each other. Prototype-level contrastive learning [5, 6, 56, 45, 15, 16] proposes to generate prototypes (also called clusters or bases) for contrasting each input image. Specifically, there are two typical ways to generate prototypes. First, Caron *et al.* proposed DeepCluster [5] to conduct online clustering on the whole dataset to generate prototypes. However, it is very time-consuming to calculate clusters on a large dataset. Thus, some recent works [6, 56, 15, 16] propose to randomly initialize a group of prototypes and then update them through back-propagation during training, which has demonstrated promising results. However, there is still no explicit guarantee that these randomly initialized prototypes can be updated well during training.

Our VoCo follows the primary idea of prototype-level constrasive learning. As shown in Fig. 2(c), to address the existing problems mentioned above, instead of randomly

initializing and updating prototypes, VoCo leverages the valuable contextual position priors of 3D medical images to generate base crops as prototypes, which also requires no time-consuming clustering on a large dataset.

SSL for medical image analysis. Due to the high potential in label-efficient learning [29, 59, 61, 58, 57, 60, 37], SSL has also received significant attention in the field of medical image analysis [70, 32, 31, 51, 19]. Existing methods are mainly based on comparative SSL [71]. Specifically, Zhou *et al.* [69] combined Mixup [66] into MoCo [29] to learn the diversity of positive and negative samples in InfoNCE [44]. Azizi *et al.* used multi-instance learning to compare multiple views of images from each patient. There are also a number of approaches [25, 70, 71] that supervising the models via restoring low-level information from raw images.

In 3D medical image analysis, reconstructing raw information is a popular pretext task for learning representations [50, 51, 71]. Existing methods are mainly based on reconstructing information from augmented images. These previous methods first conducted strong data augmentation, *e.g.*, rotate [51, 75, 52], multi-view crops [70, 71, 32], and mask [13, 73, 55], then supervised the model by reconstructing raw 3D information. Although promising results have been demonstrated, most of these methods still largely ignore the importance of integrating high-level semantics into model representations, which heavily hinders the performance of downstream tasks.

Position-related SSL. Position-related SSL methods are also explored in a number of previous works [8, 9, 41, 47, 43, 17, 64, 68] in the field of natural images. Noroozi *et al.* [43] proposed to predict the order of a set of shuffled patches. Zhai *et al.* [64] and Caron *et al.* [8] proposed to train a ViT [18] to predict the locations of each input patch. However, since the geometric relations of different objects are not very consistent in natural images, it is still difficult to effectively learn consistent position representations given visual appearance only (as stated in [68]). In addition, previous works [64, 8, 68] mainly trained a linear layer to output the positions directly, which works in a black-box manner.

In this paper, we introduce the pretext task of contextual position prediction into the field of 3D medical images, where the geometric relations between different organs are relatively consistent, which guides us to learn consistent semantic representations in pre-training. Different from the previous methods, in this paper, we introduce a totally different position prediction paradigm. Specifically, instead of using a linear layer to output positions directly, we predict the contextual positions based on volume contrast, which is more intuitive and effective.

3. Methodology

In this section, we first introduce the overall framework of our proposed VoCo in Section 3.1. After that, we present the process of contextual position prediction in Section 3.2. Then, the regularization process via volume contrast in our proposed VoCo framework is described in Section 3.3.

3.1. Overall Framework

The overall framework of our proposed VoCo is presented in Fig. 3, which contains a contextual position prediction branch and a regularization branch. The prediction branch is used to predict the contextual positions between different cropped volumes. Specifically, given an input volume, we first crop it into non-overlap base volumes, which cover the whole input volume. Then, we randomly crop a volume and transform it into the high-dimension feature space using a typical backbone (CNN [30] or Transformer [18]). The goal is to predict the contextual positions between the randomly cropped volumes and base volumes. In this paper, instead of training a linear classifier to predict positions as in previous works [64, 9, 8, 68], we propose to establish this goal by volume contrast. We develop a loss function L_{pred} to supervise the final predictions. In addition, we further use a loss function L_{reg} to regularize the feature discrepancy from different bases by enlarging their distance, aiming to learn more discriminative class assignments. The details are presented in Section 3.2 and 3.3.

3.2. Contextual Position Prediction

Base and random crops. Given an input volume, we first crop it into n non-overlap base volumes, which cover the whole input volume. We then employ the extracted features z as class assignments (we call them bases), which present the prototype-level features from different positions. Then, following previous SSL works [10, 11, 29], a projector with linear layers is used to project z into latent features q. Then, we randomly crop a volume and transform it into high-dimension feature space as p. The backbone and projector are also used to project the features from the randomly cropped volumes.

Volume contrast for contextual position prediction. With features extracted from the backbone and projector, following previous SSL works [10, 11, 29], we first conduct 3D adaptive average pooling to resize them to one dimension, *i.e.*, $p \in \mathbb{R}^{1\times C}$ and $q \in \mathbb{R}^{1\times C}$, where C is the number of channels. Specifically, we empirically set C to 2048 as in [10, 11, 29].

Then, we calculate the similarity logits l between p and q_i. Specifically, we use cosine similarity to compute l as follows:

$$l_i = CosSim(p, q_i) = \frac{p \cdot q_i}{\| p \| \| q_i \|}, i \in n \tag{1}$$

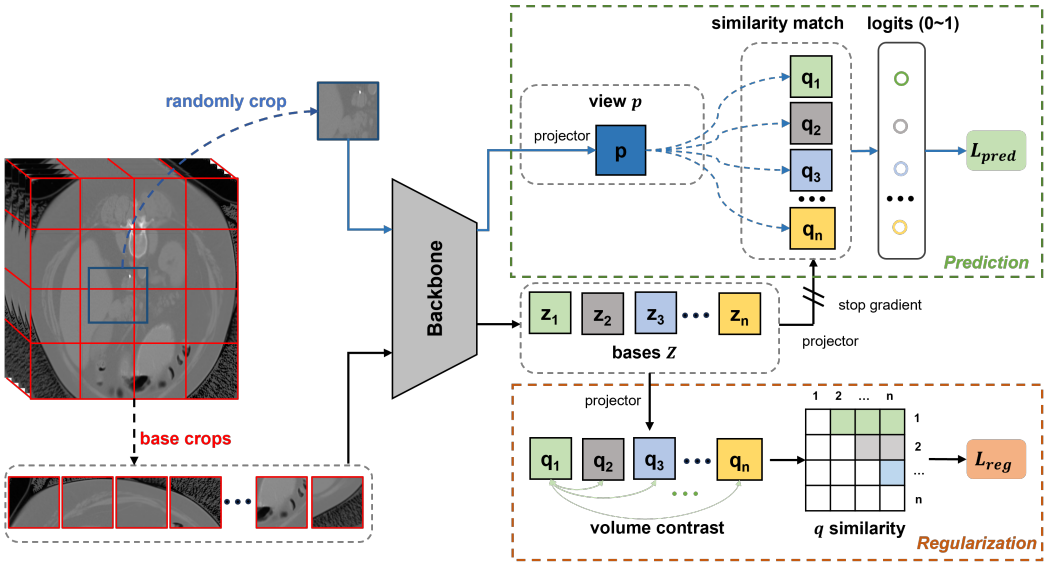

Figure 3. **The overall framework of VoCo**. VoCo contains a prediction branch and a regularization branch. The prediction branch is responsible for predicting contextual positions between different sub-volumes. The regularization branch is employed to enforce the feature discrepancy between different bases, which aims to learn more discriminative class assignments.

where q_i is the projected feature of each base crop. l_i denotes the similarity between p in q_i, which ranges from 0 to 1. It is worth noting that, we stop the gradients of q when computing Eq. 1, which aims to avoid feature collapse [10, 11, 6].

Intuitively, higher l_i represents that p has higher probabilities to share overlap regions with q_i. In this way, we can explicitly associate the similarity value with the position information, *i.e.*, p with higher l_i is more likely to be located in the region of the i_{th} base. Thus, instead of training a black-box linear layer, we predict the contextual positions by volume contrast, which is more intuitive and effective.

Position labels generation. The process of generating position labels is shown in Fig. 4. As shown in Fig. 4, when we generate $n = 4\times4$ base crops, there will be n class assignments. Then we calculate the overlap area between a randomly cropped volume and n base crops. The proportions of the overlap area are then assigned as position labels y, which also range from 0 to 1. Thus, we can easily supervise the model by calculating the distance between the prediction logits l and position labels y. The setting of the number n of base crops will be discussed in Section 4.4.

Loss function for contextual position prediction. The formulation of prediction loss function L_{pred} is based on entropy. Specifically, we first calculate the distance d between prediction logits l and position labels y:

$$d_i = |y_i - l_i|, i \in n, \tag{2}$$

where $|.|$ denotes the absolute value. Then, L_{pred} is formulated as follows:

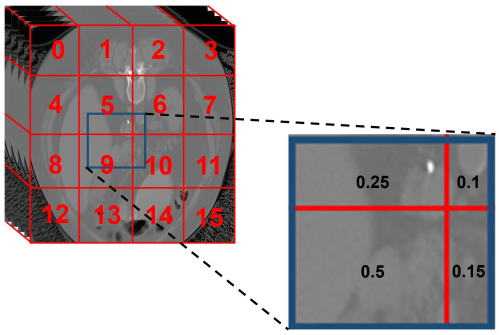

Figure 4. **The process of generating position labels**. We calculate the proportions of overlap area as position labels y, *e.g.*, the randomly cropped volume in the figure is assigned to class 5, 6, 9, and 10 with probabilities of 0.25, 0.1, 0.5, and 0.15, respectively.

$$L_{pred} = -\frac{1}{n}\sum_{i \in n}^{n} log(1 - d_i). \tag{3}$$

It is worth noting that VoCo predicts contextual positions of a volume (high similarities with all its contextual overlapped volumes), thus don't need one-to-one correspondence: *e.g.*, in Fig. 4, high-value l_i pertain to $y_i{>}0, i{=}5, 6, 9, 10$ simultaneously. Then we calculate the distance between l_i and y_i (Eq. 2).

3.3. Volume Contrast for Regularization

We aim to learn more discriminative class assignments (bases) for volume contrast. Since intuitively, different sub-volumes tend to contain different organs (semantic discrep-

ancy). Thus, we aim to enlarge the high-dimension feature discrepancy between different bases. To this end, we design a regularization loss L_{reg} to enlarge the feature discrepancy between different bases z.

First, given projected bases q, we also calculate the cosine similarity s_{ij} between different q_i and q_j as follows:

$$s_{ij} = CosSim(q_i, q_j) = \frac{q_i \cdot q_j}{\| q_i \| \| q_j \|}, i, j \in n, i \neq j, \quad (4)$$

where we aim to regularize s_{ij} to 0, enforcing feature discrepancy between different bases. Thus, the loss function L_{reg} is formulated as:

$$L_{reg} = \frac{2}{n(n-1)} \sum_{i,j \in n, i \neq j}^{n} |s_{ij}|, \quad (5)$$

where $|.|$ denotes the absolute value. With loss L_{reg}, we aim to optimize q as linearly independent bases:

$$q_i \perp q_j, i \neq j, i, j \in n. \quad (6)$$

With the regularization loss function L_{reg}, we aim to learn a group of linearly independent bases to represent all directions of high-dimension features [6]. In this way, we can learn a group of more discriminative class assignments to supervise the final position predictions.

Overall loss function. Thus, the total loss function L is the combination of L_{reg} and L_{pred}:

$$L = L_{pred} + \lambda L_{reg}, \quad (7)$$

where λ is used to balance the relative contributions of these two loss terms and set to 1.0 in experiments empirically since we consider their importance equally. The ablation studies of λ are provided in the supplementary materials.

4. Experiments

In this section, we first describe the datasets used in the pre-training and downstream tasks. Then, we briefly introduce the implementation details of VoCo. Finally, we report detailed experiment results of our proposed VoCo compared with other state-of-the-art SSL methods in 3D medical images. More details are in the supplementary materials.

4.1. Datasets

Pre-training datasets. Aiming to conduct fair comparisons with the previous works [51, 55, 70, 71, 13, 73], we also carry out pre-training experiments on the same three public datasets, *i.e.*, BTCV [35], TCIA Covid19 [14], and LUNA [48] datasets, including about totally 1.6k CT scans for pre-training. It is worth noting that, aiming to conduct fair comparisons with previous works [73, 13], we only use BTCV [35] and TCIA Covid-19 [14] for pre-training in the downstream experiments of BTCV [35]. For the other downstream tasks, we use all three datasets for pre-training. Details are provided in the supplementary materials.

Downstream datasets. To evaluate the effectiveness of our VoCo, we conduct downstream experiments on six public datasets, *i.e.*, BTCV [35], LiTs [4], MSD Spleen [1], MM-WHS [74], BraTS 21 [49], and CC-CCII [67], including segmentation and classification tasks. The first five datasets are developed for segmentation, while CC-CCII [67] is for COVID-19 classification. Note that only BTCV [35] is used in pre-training, the other datasets are unseen in pre-training. In addition, to evaluate the cross-modality generalization ability, we transfer the model pre-trained on the CT dataset to the MRI dataset BraTS 21 [49]. We adopt consistent settings as previous works [13, 73, 26, 55, 32]. We also evaluate the performance on 2D medical dataset [54]. Details are provided in the supplementary materials.

4.2. Implementation details

Following previous works [51, 55], we use Swin-UNETR [26] for both pre-training and downstream tasks. We use AdamW [38] optimizer and cosine learning rate scheduler for all experiments. We set 100K training steps in the pre-training process and applied a slicing window inference for fair comparisons with previous works [51, 55, 13, 73]. Aiming to evaluate the pure effectiveness, we do not use foundation models or post-processing [34, 36]. Details are provided in the supplementary materials.

Comparison methods. We compare our VoCo with both General and Medical SSL methods. First, we compare with the typical SSL methods MAE [28, 13] and MoCo v3 [29, 12], since they represent the two mainstream SSL paradigms, *i.e.*, mask-autoencoder and constrastive learning. Since it is not practical to set a large batch size for 3D medical images due to computation cost, for fair comparisons, we adopt consistent settings with other methods in MAE [28, 13] and MoCo v3 [29, 12]. We also report the results of SimCLR [10] and SimMIM [63] according to [13]. We further evaluate the performance of Jiasaw [9] and PosiitionLabel [68], since they are related to our position-aware methods. Most existing state-of-the-art medical SSL methods are compared in our experiments.

4.3. Experiments on downstream tasks

Outperform existing methods on the BTCV dataset. We first conduct experiments on BTCV [35], as shown in Table 1. Specifically, among the comparison methods, MAE3D [28, 13], SimCLR [10], SimMIM [63], MoCo v3 [29, 12], and GL-MAE [73] use UNETR [27]. Other methods including our VoCo adopt Swin-UNETR [26] as the default settings of the previous work [51].

Remark. It can be seen in Table 1 that the general SSL methods perform worse than most medical SSL methods.

Method	Dice Score(%)													AVG
	Spl	RKid	LKid	Gall	Eso	Liv	Sto	Aor	IVC	Veins	Pan	RAG	LAG	
From Scratch														
UNETR [27]	93.02	94.13	94.12	66.99	70.87	96.11	77.27	89.22	82.10	70.16	76.65	65.32	59.21	79.82
Swin-UNETR† [26]	94.06	93.54	93.80	65.51	74.60	97.09	75.94	91.80	82.36	73.63	75.19	68.00	61.11	80.53
With General SSL														
MAE3D [28, 13]	93.98	94.37	94.18	69.86	74.65	96.66	80.40	90.30	83.13	72.65	77.11	67.34	60.54	81.33
SimCLR [10]	92.79	93.04	91.41	49.65	50.99	98.49	77.92	85.56	80.58	64.37	67.16	59.04	48.99	73.85
SimMIM [63]	95.56	95.82	94.14	52.06	53.52	**98.98**	80.25	88.11	82.98	66.49	69.16	60.88	50.45	76.03
MoCo v3† [29, 12]	91.96	92.85	92.42	68.25	72.77	94.91	78.82	88.21	81.59	71.15	75.76	66.48	58.81	79.54
Jigsaw† [9]	94.62	93.41	93.55	75.63	73.21	95.71	80.80	89.41	84.78	71.02	79.57	65.68	60.22	81.35
PositionLabel† [68]	94.35	93.15	93.21	75.39	73.24	95.76	80.69	88.80	84.04	71.18	79.02	65.11	60.07	81.09
With Medical SSL														
MG [72]	91.99	93.52	91.81	65.11	76.14	95.98	**86.88**	89.29	83.59	71.79	81.62	67.97	63.18	81.45
ROT [50]	91.75	93.13	91.62	65.09	**76.55**	94.21	86.16	89.74	83.08	71.13	81.55	67.90	63.72	81.20
Vicregl [3]	90.32	94.15	91.30	65.12	75.41	94.76	86.00	89.13	82.54	71.26	81.01	67.66	63.08	80.89
Rubik++† [52]	**96.21**	90.41	89.33	75.22	72.64	97.44	79.25	89.65	83.76	74.74	78.35	67.14	61.97	81.38
PCRLv1† [70]	95.73	89.66	88.53	75.41	72.33	96.20	78.99	89.11	83.06	74.47	77.88	67.02	61.85	80.78
PCRLv2† [71]	95.50	91.43	89.52	**76.15**	73.54	97.28	79.64	90.16	84.17	75.20	78.71	68.74	62.93	81.74
Swin-UNETR [26, 51]	95.25	93.16	92.97	63.62	73.96	96.21	79.32	89.98	83.19	76.11	**82.25**	68.99	65.11	81.54
SwinMM [55]	94.33	94.18	94.16	72.97	74.75	96.37	83.23	89.56	82.91	70.65	75.52	69.17	62.90	81.81
GL-MAE [73]	94.54	94.39	94.37	73.19	74.93	96.51	83.49	89.74	83.11	70.80	75.71	69.39	63.12	82.01
GVSL† [32]	95.27	91.22	92.25	72.69	73.56	96.44	82.40	88.90	84.22	70.84	76.42	67.48	63.25	81.87
VoCo	95.73	**96.53**	**94.48**	76.02	75.60	97.41	78.43	**91.21**	**86.12**	**78.19**	80.88	**71.47**	**67.88**	**83.85**

Table 1. Experimental results on BTCV [35]. The best results are **bolded**. 'From Scratch' denotes the supervised baseline without self-supervised pre-training. † denotes we re-implement the approach. Most results are drawn from [13, 65, 73] or their own papers.

Method	Network	Dice Score(%)
From Scratch		
3D UNet [46]	-	90.70
UNETR† [27]	-	93.25
Swin-UNETR† [26]	-	93.42
With General SSL		
Jigsaw [9]	3D UNet	94.36
MAE3D [28, 13]	UNETR	94.02
MoCo v3† [29, 12]	UNETR	93.86
Jigsaw† [9]	Swin-UNETR	95.24
PositionLabel† [68]	Swin-UNETR	94.13
With Medical SSL		
MG [72]	3D UNet	91.30
TransVW [24]	3D UNet	91.42
ROT [50]	3D UNet	94.49
PCRLv1 [70]	3D UNet	93.87
PCRLv2 [71]	3D UNet	94.50
Rubik [75]	3D UNet	94.93
Rubik++ [52]	3D UNet	95.46
Rubik++† [52]	Swin-UNETR	95.72
Swin-UNETR [51, 26]	Swin-UNETR	95.33
SwinMM [55]	Swin-UNETR	95.52
VoCo	3D UNet	**96.03**
VoCo	Swin-UNETR	**96.52**

Table 2. Experimental results on LiTs [4]. We report the Dice Scores of liver segmentation. † denotes we re-implement the approach. Most results are drawn from [65, 71].

Method	MSD Spleen	MM-WHS
From Scratch		
3D UNet [46]	93.71	83.09
UNETR† [27]	94.20	85.85
Swin-UNETR† [26]	94.63	86.11
With General SSL		
MAE3D [28, 13]	95.20	86.03
MoCo v3† [29, 12]	94.32	84.16
Jiasaw† [9]	94.29	85.88
PositionLabel† [68]	94.16	85.52
With Medical SSL		
MG [72]	94.40	86.36
VicRegl [3]	94.12	84.72
UniMiss [62]	95.09	84.68
PCRLv1† [70]	94.32	86.58
PCRLv2† [71]	94.94	86.82
Rubik++† [52]	95.11	87.02
Swin-UNETR [51, 26]	95.02	87.06
SwinMM [55]	95.34	86.98
JSSL [42]	94.92	84.89
GVSL [32]	95.47	88.27
VoCo	**96.34**	**90.54**

Table 3. Experimental results on MSD Spleen [1] and MM-WHS [74]. We report the Dice Scores of segmentation prediction. † denotes we re-implement the approach.

Specifically, MoCo v3 [29, 12] can only achieve 79.54% Dice Score. Since MoCo v3 [29, 12] heavily relies on a large batch size to acquire adequate negative samples, which is not practical in 3D medical images due to the huge computation burden. In addition, the negative relation between different images used in MoCo [29, 12] is not appropriate in medical images. MAE [28, 13], SimCLR [10], and SimMIM [63] (results from [13]) also gain limited performance. Our VoCo also outperforms the position-based methods Jigsaw [9] and PositionLabel [68] by a clear margin. Thus, we conclude that general SSL methods are not very suitable for 3D medical images. It is crucial to consider the characteristics of medical images in medical SSL.

The scratch Swin-UNETR [26] only achieves 80.53%

Method	Net.	Dice Score(%)			
		TC	WT	ET	AVG
From Scratch					
UNETR [27]	-	81.62	87.81	57.34	75.58
Swin-UNETR [26]	-	81.28	88.67	57.73	75.89
With General SSL					
MAE3D [28, 13]	UNETR	82.34	90.35	59.18	77.29
SimMIM [63]	UNETR	84.06	90.43	59.07	77.85
SimCLR [10]	UNETR	83.13	89.44	58.42	76.99
MoCo v3† [29, 12]	UNETR	82.60	88.89	57.69	76.39
Jigsaw† [9]	Sw-UNE.	81.62	89.45	59.10	76.72
PositionLabel† [68]	Sw-UNE.	81.35	89.62	58.73	76.64
With Medical SSL					
PCRLv1†[70]	Sw-UNE.	81.96	88.83	57.58	76.12
PCRLv2†[71]	Sw-UNE.	82.13	90.06	57.70	76.63
Rubik++†[52]	Sw-UNE.	84.32	90.23	58.01	77.51
Swin-UNETR [51, 26]	Sw-UNE.	82.51	89.08	58.15	76.58
SwinMM [55]	Sw-UNE.	83.48	**90.47**	58.72	77.56
VoCo	Sw-UNE.	**85.27**	90.45	**59.87**	**78.53**

Table 4. Experimental results on BraTS 21 [49]. WT, TC, and ET denote the whole tumor, tumor core, and enhancing tumor, respectively. † denotes we re-implement the approach.

Method	Network	Accuracy(%)
From Scratch		
UNETR [27]	-	88.92
Swin-UNETR [26]	-	88.04
With General SSL		
MAE3D [28, 13]	UNETR	89.47
MoCo v3 [29, 12]	UNETR	84.95
Jiasaw [9]	Swin-UNETR	86.88
PositionLabel [68]	Swin-UNETR	87.54
With Medical SSL		
PCRLv1 [70]	Swin-UNETR	88.72
PCRLv2 [71]	Swin-UNETR	89.15
Rubik++ [52]	Swin-UNETR	89.23
Swin-UNETR [51, 26]	Swin-UNETR	89.45
SwinMM [55]	Swin-UNETR	89.61
VoCo	Swin-UNETR	**90.83**

Table 5. Experimental results of CC-CCII [67] classification.

Dice Score. With VoCo pre-training, we gain 3.32% improvements with 83.85% Dice Score, which also outperforms existing methods by a clear margin. Among the compared methods, GL-MAE [73] achieves the highest Dice Score (82.01%). Our VoCo surpasses it by 1.84% Dice Score, which is a clear improvement in this dataset.

Promising performance on Unseen datasets. We further conduct experiments on unseen datasets in pre-training, *i.e.*, LiTs [4], MSD Spleen [1], and MM-WHS [74]. The results on LiTs [4] are shown in Table 2. We report the results of compared methods according to [70, 71, 65]. Since the scratch Swin-UNETR [26] can obtain a higher Dice Score (93.42%), we further pre-train a 3D UNet [46] based on VoCo, aiming to conduct fair comparisons. It can be seen that with VoCo pre-training, Swin-UNETR [26] gains 3.10% improvements and achieves 96.52% Dice Score. With 3D UNet [46] as the backbone, VoCo also achieves 96.03% Dice Score, proving the effectiveness of VoCo with different network architectures.

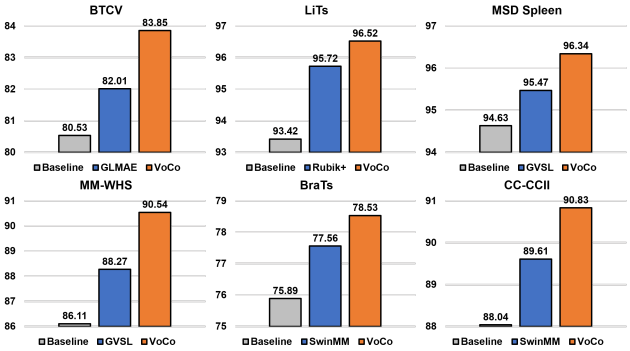

Figure 5. Overall comparisons with state-of-the-art methods on six different datasets.

Loss Functions		BTCV	MM-WHS
L_{pred}	L_{reg}		
✗	✗	80.53	86.11
✓	✗	82.96	88.82
✓	✓	**83.85**	**90.54**

Table 6. Evaluation of loss functions L_{pred} and L_{reg}. We report the average Dice Score on BTCV [35] and MM-WHS [74].

Number of bases n	BTCV	MM-WHS
(x, y, z)		
2×2×1	81.56	86.73
3×3×1	82.77	89.31
4×4×1	**83.85**	**90.54**
5×5×1	83.60	90.49
3×3×2	82.86	89.14
4×4×2	83.47	90.52

Table 7. Evaluation of the value of bases n. We report the average Dice Score on BTCV [35] and MM-WHS [74].

The results on MSD Spleen [1] and MM-WHS [74] datasets are shown in Table 3. In previous methods, GVSL [32] achieves the best performance with 95.47% and 88.27% Dice Score, while our VoCo surpasses all previous methods with 96.34% and 90.54% Dice Score on MSD Spleen [1] and MM-WHS [74] datasets, respectively.

Generalization capacity on MRI dataset. To verify the generalization capacity on the MRI dataset, we further evaluate the performance of VoCo on BraTS 21 [49]. As shown in Table 4, VoCo achieves 78.53% Dice Score and outperforms existing state-of-the-art methods, demonstrating the cross-model generalization capacity of VoCo.

Evaluation of COVID-19 classification. We further evaluate the performance of the classification task on the CC-CCII [67] dataset in Table 5. Since existing SSL methods did not conduct experiments on this dataset, we reproduce the related methods for comparisons. It can be seen that VoCo can also achieve superior results with 90.83% accuracy, proving its effectiveness in the classification task.

Overall comparisons on six downstream datasets. The

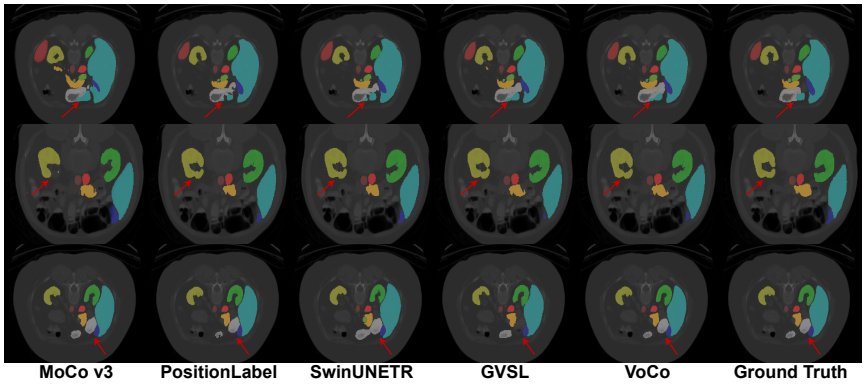

Figure 6. Qualitative visualization of segmentation results for the BTCV [35] dataset. We compare VoCo with MoCo v3 [12], PositionLabel [68], SwinUNETR [51, 26], and GVSL [32].

overall comparisons are shown in Fig. 5. Specifically, we compare with the existing state-of-the-art methods on six different downstream datasets. It can be seen that our VoCo outperforms them by a clear margin.

4.4. Ablation Study

We further conduct ablation studies to evaluate the loss functions and the settings in VoCo, which are verified on the BTCV [35] and MM-WHS [74] datasets.

Loss functions. We first study the importance of the two loss functions, *i.e.*, L_{pred} and L_{reg}, as shown in Table 6. It can be seen that with our proposed L_{pred} loss function, the performance is significantly improved, *i.e.*, 80.53% to 82.96% on BTCV, 86.11% to 88.82% on MM-WHS. These results demonstrate the effectiveness of our proposed position prediction pretext task. In addition, with the proposed regularization loss L_{reg}, the performance can be further improved. Thus, we can see that it is crucial to learn discriminative bases in VoCo.

Number of bases. We further evaluate different settings of the number of bases n in VoCo. We compare with different settings of n in the ablation studies, as shown in Table 7. It is worth noting that due to the ROI size inconsistency in the Z direction, it is not practical to crop multiple bases in the Z direction, since we have to resize the volume after crops, which will result in inconsistent volume scales. In addition, due to the computation limitation, it is costly to increase the values of n. As shown in Table 7, with $n = 2\times2\times1$, the VoCo only achieves 81.56% and 86.73% Dice Score on BTCV and MM-WHS, respectively. When we increase the values of n to $3\times3\times1$ and $4\times4\times1$, the performances are improved obviously. Specifically, with n as $4\times4\times1$, we achieve 83.85% and 90.54% on BTCV and MM-WHS, respectively. However, we observe that higher n ($5\times5\times1$) cannot further bring higher performance. We further verify the performance of generating base crops in the Z direction. It can be seen that $3\times3\times2$ and $4\times4\times2$ cannot yield improvements. Thus, aiming to balance the performance and efficiency, we set n as $4\times4\times1$ in VoCo. It can be seen that the setting of n is crucial to VoCo.

Visualization results on BTCV [35] are shown in Fig. 6. It can be seen that VoCo can yield improved segmentation accuracy and completeness. More visualization results are in the supplementary materials.

5. Conclusion and Future Directions

In this paper, we develop a simple-yet-effective SSL framework VoCo for 3D medical image analysis. Motivated by the observation that 3D medical images contain relatively consistent contextual positions between different organs, we propose to leverage the contextual position priors to learn consistent semantic representations in pre-training. Specifically, we crop volumes from different positions in an input volume and represent them as a group of bases to represent features in different directions. Then, we predict the contextual position of a randomly cropped volume by contrasting its similarity to different bases. In this way, VoCo effectively encodes the contextual position priors into model representations, enabling us to effectively improve the performance of downstream tasks that require high-level semantics. Extensive experiments demonstrate that VoCo achieves superior performance.

We will further consider several ways of extension: (1) Scale up the pre-training dataset to evaluate the upper performance of VoCo. (2) Experiments on more downstream datasets. (3) Evaluate the label-efficient performance of VoCo (*e.g.*, semi-supervised learning). (4) Explore the capacity of VoCo in mining inter-volume relationships.

Acknowledgments

This work was supported by Hong Kong Innovation and Technology Fund (Project No. ITS/028/21FP and No. MHP/002/22), and Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. T45-401/22-N).

References

[1] Michela Antonelli et al. The medical segmentation decathlon. *Nature Commun.*, 13(1):4128, 2022. 5, 6, 7

[2] Shekoofeh Azizi et al. Big self-supervised models advance medical image classification. In *ICCV*, pages 3478–3488, 2021. 1

[3] Adrien Bardes, Jean Ponce, and Yann LeCun. Vicregl: Self-supervised learning of local visual features. *NIPS*, 35:8799–8810, 2022. 6

[4] Patrick Bilic et al. The liver tumor segmentation benchmark (lits). *Medical Image Analy.*, 84:102680, 2023. 5, 6, 7

[5] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *ECCV*, pages 132–149, 2018. 2

[6] Mathilde Caron et al. Unsupervised learning of visual features by contrasting cluster assignments. *NIPS*, 33:9912–9924, 2020. 1, 2, 4, 5

[7] Mathilde Caron et al. Emerging properties in self-supervised vision transformers. In *ICCV*, pages 9650–9660, 2021. 2

[8] Mathilde Caron, Neil Houlsby, and Cordelia Schmid. Location-aware self-supervised transformers. *arXiv preprint arXiv:2212.02400*, 2022. 3

[9] Pengguang Chen, Shu Liu, and Jiaya Jia. Jigsaw clustering for unsupervised visual representation learning. In *CVPR*, pages 11526–11535, 2021. 3, 5, 6, 7

[10] Ting Chen et al. A simple framework for contrastive learning of visual representations. In *ICML*, pages 1597–1607, 2020. 1, 2, 3, 4, 5, 6, 7

[11] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *CVPR*, pages 15750–15758, 2021. 1, 2, 3, 4

[12] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. *arXiv preprint arXiv:2104.02057*, 2021. 5, 6, 7, 8

[13] Zekai Chen et al. Masked image modeling advances 3d medical image analysis. In *WACV*, pages 1970–1980, 2023. 1, 3, 5, 6, 7

[14] 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. 5

[15] Jiequan Cui et al. Parametric contrastive learning. In *ICCV*, pages 715–724, 2021. 2

[16] Jiequan Cui et al. Generalized parametric contrastive learning. *IEEE Trans. Pattern Analy. Mach. Intell.*, 2023. 2

[17] Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by context prediction. In *ICCV*, pages 1422–1430, 2015. 3

[18] Alexey Dosovitskiy et al. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2020. 3

[19] Hao Du, Qihua Dong, Yan Xu, and Jing Liao. Weakly-supervised 3d medical image segmentation using geometric prior and contrastive similarity. *IEEE Trans. Medi. Imag.*, 2023. 3

[20] Yuting Gao, Jia-Xin Zhuang, et al. Disco: Remedying self-supervised learning on lightweight models with distilled contrastive learning. In *ECCV*, pages 237–253, 2022. 1

[21] Florin-Cristian Ghesu et al. Multi-scale deep reinforcement learning for real-time 3d-landmark detection in ct scans. *IEEE Trans. Pattern Anal. Mach. Intell.*, 41(1):176–189, 2017. 1

[22] Jean-Bastien Grill et al. Bootstrap your own latent-a new approach to self-supervised learning. *NIPS*, 33:21271–21284, 2020. 2

[23] Katharina Grünberg et al. Annotating medical image data. *Medical Image Analy.*, pages 45–67, 2017. 1

[24] 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. 6

[25] Fatemeh Haghighi et al. Dira: Discriminative, restorative, and adversarial learning for self-supervised medical image analysis. In *CVPR*, pages 20824–20834, 2022. 1, 3

[26] Ali Hatamizadeh et al. Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In *MICCAIW*, pages 272–284, 2021. 5, 6, 7, 8

[27] Ali Hatamizadeh et al. Unetr: Transformers for 3d medical image segmentation. In *WACV*, pages 574–584, 2022. 5, 6, 7

[28] Kaiming He et al. Masked autoencoders are scalable vision learners. In *CVPR*, pages 16000–16009, 2022. 1, 5, 6, 7

[29] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, pages 9729–9738, 2020. 2, 3, 5, 6, 7

[30] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 3

[31] Xingxin He, Leyuan Fang, Mingkui Tan, and Xiangdong Chen. Intra-and inter-slice contrastive learning for point supervised oct fluid segmentation. *IEEE Trans. Image Process.*, 31:1870–1881, 2022. 3

[32] Yuting He et al. Geometric visual similarity learning in 3d medical image self-supervised pre-training. In *CVPR*, pages 9538–9547, 2023. 1, 3, 5, 6, 7, 8

[33] Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18(2):203–211, 2021. 1

[34] Yankai Jiang et al. Anatomical invariance modeling and semantic alignment for self-supervised learning in 3d medical image analysis. In *ICCV*, pages 15859–15869, 2023. 1, 5

[35] Bennett Landman et al. Miccai multi-atlas labeling beyond the cranial vault–workshop and challenge. In *MICCAIW*, volume 5, page 12, 2015. 5, 6, 7, 8

[36] Jie Liu et al. Clip-driven universal model for organ segmentation and tumor detection. In *ICCV*, pages 21152–21164, 2023. 1, 5

[37] Qiang Liu et al. A multi-level label-aware semi-supervised framework for remote sensing scene classification. *IEEE Trans. Geosci. Remote Sens.*, 2023. 3

[38] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 5

[39] Xiangde Luo et al. WORD: A large scale dataset, benchmark and clinical applicable study for abdominal organ segmentation from ct image. *Medical Image Analysis*, 82:102642, 2022. 1

[40] Jun Ma et al. Abdomenct-1k: Is abdominal organ segmentation a solved problem? *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(10):6695–6714, 2021. 1

[41] T Nathan Mundhenk, Daniel Ho, and Barry Y Chen. Improvements to context based self-supervised learning. In *CVPR*, pages 9339–9348, 2018. 3

[42] Nguyen et al. Joint self-supervised image-volume representation learning with intra-inter contrastive clustering. In *AAAI*, pages 14426–14435, 2023. 6

[43] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *ECCV*, pages 69–84. Springer, 2016. 3

[44] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 3

[45] Maxime Oquab et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 2

[46] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, pages 234–241, 2015. 6, 7

[47] Rodrigo Santa Cruz et al. Deeppermnet: Visual permutation learning. In *CVPR*, pages 3949–3957, 2017. 3

[48] 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. 5

[49] 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. 5, 7

[50] Aiham Taleb et al. 3d self-supervised methods for medical imaging. *NIPS*, 33:18158–18172, 2020. 1, 3, 6

[51] Yucheng Tang et al. Self-supervised pre-training of swin transformers for 3d medical image analysis. In *CVPR*, pages 20730–20740, 2022. 1, 3, 5, 6, 7, 8

[52] Xing Tao et al. Revisiting rubik's cube: self-supervised learning with volume-wise transformation for 3d medical image segmentation. In *MICCAI*, pages 238–248, 2020. 1, 3, 6, 7

[53] Guotai Wang et al. Deepigeos: a deep interactive geodesic framework for medical image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 41(7):1559–1572, 2018. 1

[54] 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. 5

[55] Yiqing Wang et al. Swinmm: masked multi-view with swin transformers for 3d medical image segmentation. In *MICCAI*, 2023. 1, 3, 5, 6, 7

[56] Xin Wen et al. Self-supervised visual representation learning with semantic grouping. *NIPS*, 35:16423–16438, 2022. 2

[57] Linshan Wu et al. Deep bilateral filtering network for point-supervised semantic segmentation in remote sensing images. *IEEE Trans. Image Process.*, 31:7419–7434, 2022. 3

[58] Linshan Wu et al. Modeling the label distributions for weakly-supervised semantic segmentation, 2024. 3

[59] Linshan Wu, Leyuan Fang, Xingxin He, Min He, Jiayi Ma, and Zhun Zhong. Querying labeled for unlabeled: Cross-image semantic consistency guided semi-supervised semantic segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(7):8827–8844, Jul. 2023. 3

[60] Linshan Wu, Ming Lu, and Leyuan Fang. Deep covariance alignment for domain adaptive remote sensing image segmentation. *IEEE Trans. Geosci. Remote Sens.*, 60:1–11, 2022. 3

[61] Linshan Wu, Zhun Zhong, Leyuan Fang, Xingxin He, Qiang Liu, Jiayi Ma, and Hao Chen. Sparsely annotated semantic segmentation with adaptive gaussian mixtures. In *CVPR*, pages 15454–15464, 2023. 3

[62] Yutong Xie, Jianpeng Zhang, Yong Xia, and Qi Wu. Unimiss: Universal medical self-supervised learning via breaking dimensionality barrier. In *ECCV*, pages 558–575, 2022. 1, 6

[63] Zhenda Xie et al. Simmim: A simple framework for masked image modeling. In *CVPR*, pages 9653–9663, 2022. 5, 6, 7

[64] Shuangfei Zhai et al. Position prediction as an effective pretraining strategy. *arXiv preprint arXiv:2207.07611*, 2022. 3

[65] 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. 6, 7

[66] Hongyi Zhang et al. Mixup: Beyond empirical risk minimization. In *ICLR*, 2018. 3

[67] 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. 5, 7

[68] Zhemin Zhang and Xun Gong. Positional label for self-supervised vision transformer. In *AAAI*, pages 3516–3524, 2023. 3, 5, 6, 7, 8

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

[70] 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, 5, 6, 7

[71] 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, 3, 5, 6, 7

[72] Zongwei Zhou et al. Models genesis. *Medical Image Analy.*, 67:101840, 2021. 6

[73] 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, 3, 5, 6, 7

[74] Xiahai Zhuang. Multivariate mixture model for myocardial segmentation combining multi-source images. *IEEE Trans. Pattern Analy. Mach. Intell.*, 41(12):2933–2946, 2018. 5, 6, 7, 8

[75] Xinrui Zhuang et al. Self-supervised feature learning for 3d medical images by playing a rubik's cube. In *MICCAI*, pages 420–428, 2019. 1, 3, 6