Large-Scale 3D Medical Image Pre-training with Geometric Context Priors

Linshan Wu, Jiaxin Zhuang, Hao Chen, *Senior Member, IEEE*

Abstract—The scarcity of annotations poses a significant challenge in medical image analysis, which demands extensive efforts from radiologists, especially for high-dimension 3D medical images. Large-scale pre-training has emerged as a promising label-efficient solution, owing to the utilization of large-scale data, large models, and advanced pre-training techniques. However, its development in medical images remains underexplored. The primary challenge lies in harnessing large-scale unlabeled data and learning high-level semantics without annotations. We observe that 3D medical images exhibit consistent geometric context, *i.e.*, consistent geometric relations between different organs, which leads to a promising way for learning consistent representations. Motivated by this, we introduce a simple-yet-effective **Vo**lume **Co**ntrast (**VoCo**) framework to leverage geometric context priors for self-supervision. Given an input volume, we extract base crops from different regions to construct positive and negative pairs for contrastive learning. Then we predict the contextual position of a random crop by contrasting its similarity to the base crops. In this way, VoCo implicitly encodes the inherent geometric context into model representations, facilitating high-level semantic learning without annotations. To assess effectiveness, we **(1)** introduce PreCT-160K, the largest medical image pre-training dataset to date, which comprises 160K Computed Tomography (CT) volumes covering diverse anatomic structures; **(2)** investigate scaling laws and propose guidelines for tailoring different model sizes to various medical tasks; **(3)** build a comprehensive benchmark encompassing 48 medical tasks, including segmentation, classification, registration, and vision-language. Extensive experiments highlight the superiority of VoCo, showcasing promising transferability to unseen modalities and datasets. VoCo notably enhances performance on datasets with limited labeled cases and significantly expedites fine-tuning convergence. Codes, datasets, and models are available at https://github.com/Luffy03/Large-Scale-Medical.

Index Terms—Vision Pre-training, Foundation Models, Medical Image Analysis, Geometric Context Priors, Scalable Learners

1 Introduction

AI-driven medical image analysis has witnessed emerging development in recent years [2], [3], [4], [5], [6], [7], yet is heavily hampered by the high costs of the required expert annotations, especially for large-scale 3D medical images that with volumetric information [8], [9], [10], [11]. To address this dilemma, Self-Supervised Learning (SSL) [12], [13], [14], [15], [16] for pre-training foundation models have demonstrated the potential to learn feature representations without the guidance from annotations, offering a promising solution in addressing the annotation bottleneck in 3D medical image analysis [8], [9], [17], [18], [19].

Recent advances [12], [20], [21], [22], [23], [24] have highlighted the critical elements contributing to the success of vision foundation models, *i.e.*, large-scale data, large models, and advanced pre-training techniques. However, ***how well these solutions transfer to 3D medical image pre-training*** has not been thoroughly investigated. As shown in Fig. 1, **(1)** Data: previous methods [8], [9], [17], [18], [25], [26], [27], [28] are limited by the data scale (at most 10K volumes are used). Specifically, UniMiss [9], [27] innovatively proposed to boost chest CT pre-training by integrating 2D chest X-rays. However, the extendability to other anatomic regions remains under-explored. **(2)** Models: models trained in previous methods [8], [9], [17], [18], [25], [26], [27], [28] are still small-scale, with parameters only in the tens of millions. The scaling law of model capacity in medical image pre-training has not been well-explored. **(3)** Pre-training techniques: SuPreM [26] focused on supervised pre-training and annotated an abdomen segmentation dataset [29] for this purpose. Although showcasing state-of-the-art performance compared to previous methods, SuPreM [26] is still constrained by the scale of labeled data and fails to incorporate large-scale unlabeled data from diverse anatomical regions. In SSL, the majority of existing approaches [8], [9], [10], [17], [19], [25], [30], [31] relied on low-level information reconstructions to learn augment-invariant representations, which typically employ data augmentation to the images and then reconstruct the raw information. However, the lack of high-level semantics in pre-training still impedes the performance of various downstream tasks.

The primary challenge is to incorporate high-level semantics for pre-training large-scale unlabeled data. We highlight that the geometric context priors of 3D medical images can be exploited. As illustrated in Fig. 2, we observe that in 3D medical images, different organs (semantic regions) exhibit relatively consistent geometric relations with similar anatomic characteristics. Thus, the consistent geometric context between different organs offers a promising avenue for us to learn consistent semantic representations without the guidance of annotations in pre-training.

In this paper, we propose a simple-yet-effective Volume Contrast (VoCo) framework, aiming to leverage the geometric context priors for contrastive learning. VoCo introduces

• *Linshan Wu, Jiaxin Zhuang, and Hao Chen are with the Department of Computer Science and Engineering, The Hong Kong University of Science and Technology, Hong Kong, China. E-mail: linshan.wu@connect.ust.hk, jzhuangad@cse.ust.hk, jhc@cse.ust.hk.*

This paper is an extension of our CVPR 2024 paper [1].
Corresponding author: Hao Chen (jhc@cse.ust.hk).

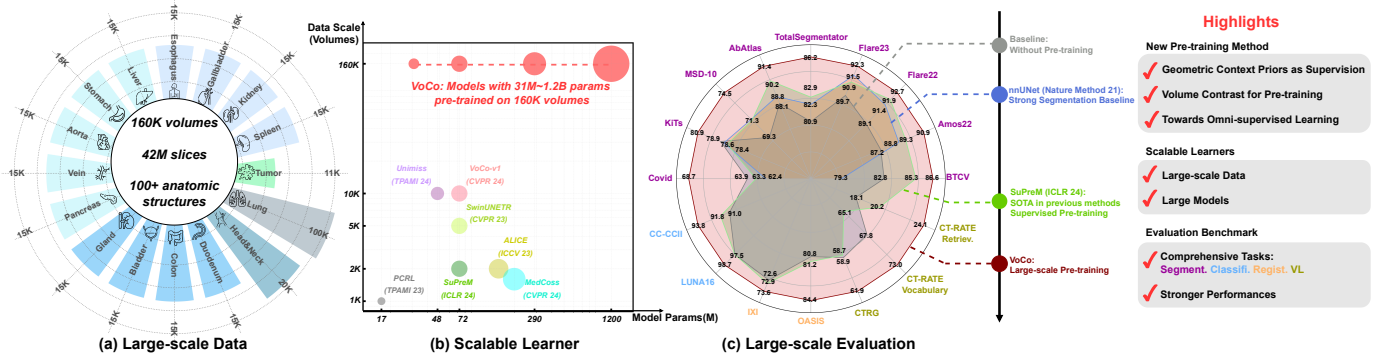

Fig. 1: **Overview: (a)** We curate a large-scale 3D medical dataset PreCT-160K for pre-training. To the best of our knowledge, it is the existing largest pre-training dataset in this field, comprising 160K CT volumes (42M slices). **(b)** We investigate the scaling law in medical image pre-training, where VoCo stands out from previous methods in both data scale and model capacity. **(c)** We build a comprehensive benchmark for evaluation, which contains **48** downstream datasets across different tasks, *i.e.*, segmentation, classification, registration, and vision-language (VL). Extensive experiments highlight the effectiveness of our proposed large-scale pre-training method.

a novel pretext task, *i.e.*, contextual position predictions, aiming to encode the geometric relation of different organs into model representations. First, VoCo extracts a group of non-overlap base crops from different regions within an input volume. The base crops are employed to construct positive and negative pairs with a random crop for contrastive learning, *i.e.*, base crops that overlap with the random crop are assigned as positive, otherwise negative. Then, we predict the contextual positions of a random crop by contrasting its similarity to the base crops. Intuitively, higher similarity indicates larger overlap areas, thus we can predict which region the random crop belongs to by calculating similarity. Specifically, we assign the overlap proportions between the random crop and base crops as position labels to supervise the position predictions. Through learning to predict contextual positions, VoCo implicitly encodes the inherent geometric contexts into the model representations without the guidance of annotations.

As shown in Fig. 1, existing works [8], [9], [17], [18], [19], [26] are still constrained by the size of data, resulting in a large gap towards powerful medical vision foundation models. To this end, we curate a large-scale dataset PreCT-160K from publicly available sources, which currently stands as the largest and most comprehensive dataset for medical image pre-training. As shown in Fig. 1(a), PreCT-160K comprises over 160K CT volumes with an excess of 42M slices, encompassing the 3D anatomical map of human bodies. PreCT-160K also includes a substantial portion of labeled data, enabling us to combine self- and semi-supervised learning for omni-supervised pre-training. In this paper, we propose an omni-supervised pre-training framework to effectively unleash the power of labeled and unlabeled medical images.

We further explore the scaling law of model capacity and develop guidelines for tailoring different model sizes to diverse medical tasks. Specifically, we build a large-scale evaluation benchmark for medical image pre-training. In contrast to previous studies [8], [9], [17], [18], [19], [26] that were limited in evaluation data and tasks, our benchmark encompasses 48 downstream datasets spanning various tasks such as segmentation, classification, registration, and vision-language. Extensive experimental results on 48 downstream datasets demonstrate that our proposed VoCo significantly outperforms existing methods by a clear margin and achieves new state-of-the-art performances.

The preliminary version of this study was presented in CVPR 2024 [1] and we named it VoCo-v1. In this paper, we made significant and substantial modifications, retaining the initial name as VoCo. The new contributions of this paper include but are not limited to:

- Compared to VoCo-v1 [1] that solely focused on intra-volume contrastive learning, we further introduce inter-volume contrastive learning with a momentum-based teacher-student module, enabling us to learn consistent representations between different volumes.
- We investigate the combination of self- and semi-supervised learning for omni-supervised pre-training, effectively leveraging both labeled and unlabeled data.
- We introduce the existing largest medical image pre-training dataset PreCT-160K and scale up the data scale from 10K [1] to 160K. Our PreCT-160K is poised to foster future research in medical image pre-training.
- We build the existing largest evaluation benchmark for medical image pre-training, encompassing diverse tasks across 48 downstream datasets. Our open-source implementation of various medical tasks will also benefit the following researchers in this field.
- We delve deeper into the scaling law and release pre-trained models with parameter sizes ranging from 31M to 1.2B. We also propose the guidelines for tailoring different model sizes to various medical tasks.
- We provide detailed and insightful analyses to underscore the core components of VoCo. These experiments further highlight the significance of large-scale pre-training, offering valuable insights that can inspire future research in the field of medical image pre-training.

2 Related Work

2.1 Large-scale Vision Pre-training

Vision pre-training opens up immense opportunities for harnessing large-scale vision data, playing a pivotal role in

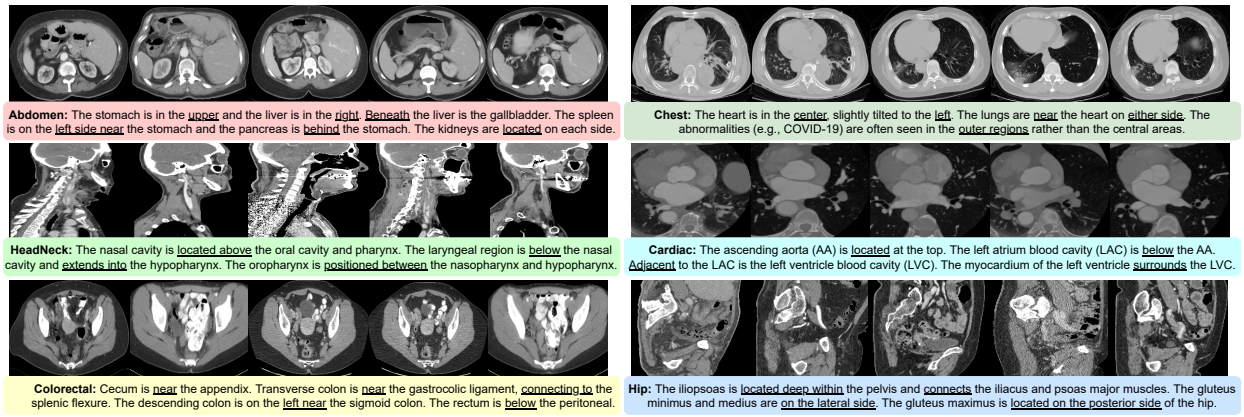

Fig. 2: **Motivation of VoCo**. In 3D medical images, the geometric relations between different organs are relatively consistent. We present some examples from PreCT-160K to illustrate these anatomical relationships across different regions. Motivated by this observation, we propose to leverage geometric context priors for learning consistent semantic representations and introduce a novel position prediction pretext task for pre-training.

the development of large vision foundation models [12], [20], [22], [23], [24], [32]. The primary challenge lies in devising effective pre-training methodologies. Although supervised pre-training stands as a straightforward approach, it grapples with the challenge of the lack of manual annotations, demanding substantial engineering efforts for building labeled datasets. Deng et al. [33] built the famous ImageNet and ImageNet pre-training has demonstrated its effectiveness in boosting downstream tasks. SAM [22] introduced the SA-1B dataset with over 1 billion segmentation masks for supervised pre-training, thus achieving a strong segmentation foundation model. However, the high costs of annotations and the neglect of large-scale unlabeled data still hinder the further development of supervised pre-training. To this end, SSL proposed to learn robust features without the guidance of annotations [12], [13], [14], [20], [21], which has garnered significant attention recently.

Typical SSL methods. SSL has showcased promising results across various vision tasks [12], [13], [14], [20], [21], [34]. DINO [12], [20] proposed to integrate advanced SSL methods and learn robust features without annotations, which has become a prevalent choice for pre-trained backbones in contemporary research. State-of-the-art SSL methods can broadly be classified into generative [21], [35], [36], [37] and contrastive [13], [14], [34], [38], [39], [40], [41], [42] learning methods. **(1)** Generative learning methods are mainly based on reconstructing raw information from augmented images. For example, MAE [21] proposed to mask random patches of the input image and reconstruct the missing pixels. **(2)** Contrastive learning methods aim to learn consistent representations by contrasting positive and negative pairs of samples.

Transfer to medical image analysis. Although the methods discussed above have achieved promising outcomes in natural images, directly applying these pre-trained models to medical images encounters challenges due to domain gaps [8], [9], [17], [19], [27], [28]. DINO [12], [20] pre-trained a series of strong 2D vision Transformers [43] and exhibited significant transferability in 2D medical images like X-ray and pathology images [44], [45], [46]. However, in the realm of challenging 3D medical tasks that necessitate volumetric information extraction, strong pre-trained 3D models are still under-explored [8], [9], [26], [47].

Most state-of-the-art SSL methods [14], [15], [21], [34], [38] often fall short in achieving competitive performances in 3D medical images, primarily caused by the ignorance of the unique characteristics of 3D medical images [8], [9], [18], [28]. Specifically: **(1)** Contrastive learning [15], [38] in natural images proposed to build positive and negative pairs of samples in a training batch, *i.e*, the augmented view of input is assigned as positive, and other images are negative. However, for 3D medical images that share similar anatomical structures, it is difficult to build negative pairs in this way [10], [11], [17], [28]. **(2)** Mask image modeling [13], [21] proposed to mask and reconstruct missing pixels. However, for 3D medical images characterized by high dimensions, large sizes, and a significant background proportion, these methods often encounter issues as models tend to converge towards reconstructing irrelevant background regions [16], [17], [31], [48], [49], [50], diminishing the understanding of semantic regions (*e.g.*, organs). Thus, the development of advanced SSL techniques for 3D medical images necessitates a meticulous consideration of the unique image information and the formulation of tailored strategies.

2.2 Large-scale Medical Image Pre-training

Medical image pre-training has been proven as an effective way to mitigate the scarcity of annotation in medical tasks [8], [9], [18], [26], [51], [52]. Early attempts [25], [53], [54] conducted pre-training on 2D X-ray images [55], [56], demonstrating improvements on chest pathology identification and pneumothorax segmentation. In comparison, 3D medical images, *e.g.*, CT and Magnetic Resonance Imaging (MRI), offer richer volumetric information for clinical diagnosis, which has received increasing attention in medical image analysis [10], [11], [16], [26], [47], [57]. Nonetheless, the complexity inherent in 3D medical images introduces significant challenges to pre-training. Although recent works [8], [9], [17], [18], [19], [26], [28] have demonstrated the effectiveness of 3D medical image pre-training, significant challenges still persist, particularly in the realms of data scale, model capacity, and pre-training methods.

2.2.1 Large-scale Data

Compared with 2D X-ray, collecting 3D medical images like CTs is more difficult, stemming from factors such as slower imaging speeds, heightened radiation exposure, and increased costs [58], [59]. As shown in Fig. 1(b), most existing methods [8], [17], [18], [19], [25], [26] leveraged limited scale of 3D data for pre-training. FreeTumor [60] first investigated the data-scaling law in tumor segmentation with 11K CTs. Wang et al. [50] built a dataset of 100K CTs for pre-training but it is not publicly available for research. To collect large-scale 3D data for pre-training, the necessity arises to ***aggregate datasets from diverse sources***, encompassing various hospitals across different regions and countries [29], [61]. This will lead to diverse image characteristics and inconsistent imaging quality in the dataset, introducing new challenges to pre-training.

Moreover, previous methods mainly collected data from specific body parts for pre-training, *e.g.*, PCRL [8], [25] and Unimiss [9], [27] on chest region, Alice [18] and SuPreM [26] on abdomen region, GVSL [28] on cardiac region. However, given the distinct characteristics present in various anatomical regions, the transferability of models pre-trained on one region to another may be constrained [1], [16], [48]. In this paper, we build a large-scale dataset PreCT-160K that encompasses diverse anatomic structures. However, data sourced from different anatomical regions exhibit varying imaging parameters, *i.e.*, different sizes, spacing, and intensities, posing new challenges for learning consistent representations in pre-training.

2.2.2 Large Model

Early works [25], [53], [54] in 3D medical image pre-training were constrained in model capacity, typically comprising only tens of millions of parameters. Recent advances [20], [21], [23], [43], [62] have demonstrated the astonishing effectiveness of scaling law, where large models trained on large-scale data exhibit remarkable intelligence. In this paper, we collect a large-scale 3D medical image dataset, which comprises diverse image characteristics from various sources. The availability of such extensive data unlocks new opportunities for us to train large models.

Given the diversity of various medical tasks, it is imperative to evaluate large models on comprehensive benchmarks. Previous methods [8], [17], [18], [26], [27], [53], [54] primarily evaluated the pre-trained models on only a few downstream tasks, typically focusing on segmentation or classification tasks. STU-Net [63] was the first to evaluate large models yet is limited in segmentation tasks. In this paper, we delve deeper into the scaling law in various medical tasks, providing insights into tailoring diverse model sizes to accommodate varying medical tasks effectively.

2.2.3 Advanced Pre-training Techniques

SSL for 3D medical images. Existing methods [8], [10], [19], [30], [31], [64] 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. Rotate-and-reconstruct [10], [17], [30], [65] proposed to randomly rotate the 3D volumetric images and learn to recover them, fostering the learning of rotation-invariant features. Recent methods [8], [18], [25], [27], [28], [66] delved into restoring low-level information across varied image perspectives. PCRL [8], [25] cropped global and local patches then conducted multi-scale restorations. GVSL [28] further explored the geometric similarities among multi-scans through affine augmentation and matching. Mask-reconstruct methods [16], [19], [31], [48], [49] were derived from MAE [21], aiming to learn representations by masking images and reconstructing the missing pixels. Although promising results have been demonstrated, the majority of these approaches often overlook the significance of integrating high-level semantics into model representations, thus impeding the further improvements in downstream tasks.

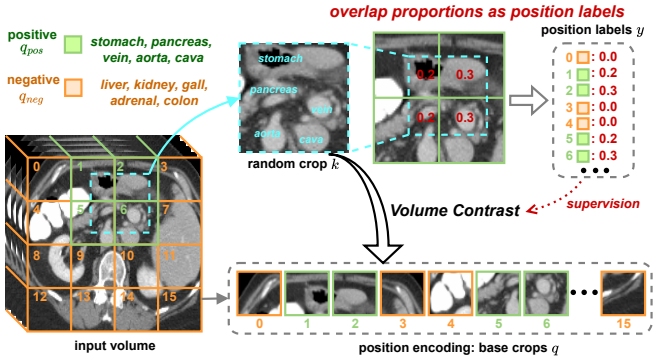

Fig. 3: **Generate position labels for supervision**. A pair of random crop k and base crop q are assigned as *positive* if they share overlap areas, otherwise as *negative*. We calculate the overlap proportions as position labels y, *e.g.*, y_1, y_2, y_5, y_6 are assigned as $0.2, 0.3, 0.2, 0.3$, respectively.

High-level semantics in pre-training. For medical images, high-level semantic information primarily stems from manual annotations, since it heavily relies on expert knowledge. Previous works [26], [63], [67], [68] proposed that supervised pre-training is more efficient and can achieve higher performances with less training time and labeled data [26]. However, the ongoing challenges persist in the scarcity of labeled data, impeding the transferability to various medical tasks, different anatomical structures, and extensive unseen datasets. In this paper, we aim to integrate large-scale unlabeled data into pre-training. Thus, we propose to leverage the inherent characteristics of medical images as high-level semantic priors for self-supervision.

Omni-supervised Learning. Although self-supervised learning enables us to involve large-scale unlabeled data in pre-training [1], [9], [16], it still overlooks the utilization of readily available labeled data. Omni-supervised learning [69], [70], [71] introduced the concept of leveraging diverse information for supervision. Specifically, semi-supervised learning [72], [73], [74] demonstrates powerful efficacy in leveraging both labeled and unlabeled data. In this paper, we propose a simple-yet-effective omni-supervised pre-training framework, which combines self- and semi-supervised learning to unleash the power of both labeled and unlabeled medical images.

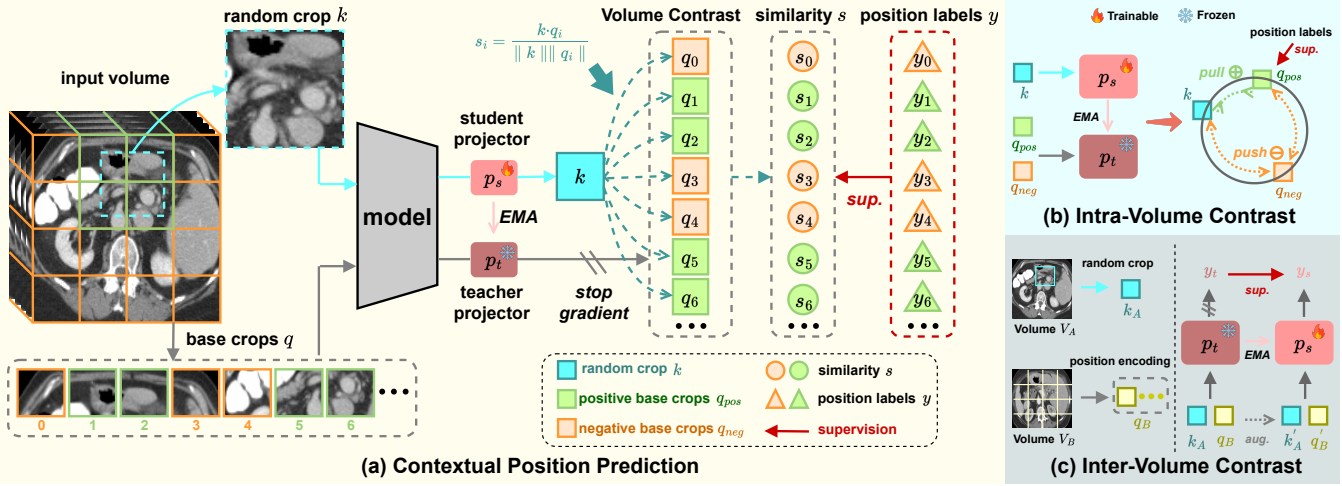

Fig. 4: **Overall framework of VoCo. (a)** First, we generate base crops q with corresponding position labels y (Sec. 3.1 & Fig. 3). Then we input the random crop k and base crops q for contextual position prediction. Specifically, we employ a student-teacher module to project k and q separately, where the teacher projector is frozen and updated from the student projector with Exponential Moving Average (EMA). Finally, we conduct volume contrast between k and q to predict similarity s (Eq. 2), where s is supervised by position labels y (Eq. 4). **(b)** We use the position labels to supervise the intra-volume contrast on k, q_{pos}, and q_{neg}, where k, q_{pos}, and q_{neg} are from the same volume. **(c)** We extract random crop k_A and base crops q_B from different volumes V_A and V_B for inter-volume contrast.

3 Method

3.1 Generate Position Labels for Supervision

The pivotal procedure is to generate position labels for self-supervision. We propose to leverage the inherent geometric context priors in 3D medical images. As illustrated in Fig. 3, given an input volume V, we first randomly crop a sub-volume k, with the objective of constructing positive and negative pairs with k for contrastive learning. Specifically, we propose to employ position encoding to generate n non-overlap base crops $q_i, i \in n$. For example, $n = 4 \times 4$ base crops are generated in Fig. 3, where each base crop q_i represents a distinct region of the input volume.

Within human body anatomy, various organs are situated in distinct regions, leading to a potential way for us to form positive and negative pairs. As shown in Fig. 3, the random crop k and the positive base crops q_{pos} exhibit overlap areas, whereas the negative base crops q_{neg}, lacking such overlaps, more likely encompass different organs (not absolutely). For example, in Fig. 3, k and q_{pos} both contain *stomach, pancreas, vein, aorta,* and *cava*, while k and q_{neg} exhibit different organ information. Thus, we can employ the position encoding to construct positive and negative pairs for contrastive learning.

Previous contrastive learning methods [14], [15], [38], [42] mainly employ InfoNCE loss [75] to maximize the mutual information of positive pairs. In this paper, we propose to generate labels with specific values to supervise the correlation extent of positive pairs, *i.e.*, with labels to reflect **how similar between k and q_{pos}**. It can be observed that the correlation between k and q_{pos} is associated with their overlap proportions. Intuitively, if a positive base crop q_{pos} shares more overlap areas with k, this q_{pos} would be more similar with k. Thus, as shown in Fig. 3, we propose to assign the overlap proportions as the values of position labels y, enabling us to measure the similarity between k and q_{pos}. In contrast, the position labels y of q_{neg} are assigned to 0. In this way, we leverage the overlap proportions between k and q to supervise the contextual position prediction results.

3.2 Volume Contrast for Contextual Position Prediction

The overall framework of VoCo is present in Fig. 4. Specifically, we propose a novel pretext task, *i.e.*, contextual position prediction, which employs volume contrast to predict the contextual positions of a random crop k. This pretext task includes: **(1)** intra-volume contrast among k, q_{pos}, and q_{neg}, where k, q_{pos}, and q_{neg} are from the same volume; **(2)** inter-volume contrast between different volumes V_A and V_B, which is established by consistency regularization with a typical student-teacher module [14], [15], [38], [42].

Contextual position prediction. As shown in Fig. 4(a), given an input volume, we first extract a random crop k and a group of base crops q, where the corresponding position labels y_i for q_i are generated as Sec. 3.1 and Fig. 3. Then we feed k and q into the model to extract high-dimension features. After extracting the features, we employ a typical momentum-based student-teacher module [15], [38] to project k and q separately. Specifically, the teacher projector p_t is frozen during training, where its parameters θ_t are updated from the parameters θ_s of the student projector p_s by Exponential Moving Average (EMA):

$$\theta_t = \rho\theta_t + (1-\rho)\theta_s, \tag{1}$$

where ρ is the momentum factor and is empirically set to 0.9. The momentum-based student-teacher module is effective in contrastive learning [15], [38], which enables stable training and avoids feature collapse [14], [34], [42].

With features extracted from the projectors, we conduct 3D adaptive average pooling to resize k and q as one dimension features, *i.e.*, $k \in \mathbb{R}^{1\times C}$ and $q \in \mathbb{R}^{1\times C}$, where C is

the number of feature dimensions. Then, we calculate the similarity s_i between random crop k and each base crop q_i. Specifically, we use cosine similarity to compute s_i as:

$$s_i = CosSim(k, q_i) = \frac{k \cdot q_i}{\parallel k \parallel \parallel q_i \parallel}, i \in n, \tag{2}$$

where s_i denotes the similarity between k and q_i, which ranges from 0 to 1.

Intuitively, higher s_i represents that k has higher probabilities to share overlap proportions with q_i. In this way, we can predict the contextual position by calculating the similarity s. Then, we leverage the generated position labels y to supervise the predicted similarity s. The formulation of prediction loss function L_{pred} is based on entropy. Specifically, we first calculate the distance d between similarity s and position labels y:

$$d_i = |y_i - s_i|, i \in n, \tag{3}$$

where $|.|$ denotes the absolute value. Note that both s and y are ranging from 0 to 1. Then, L_{pred} is formulated as:

$$L_{pred} = -\frac{1}{n} \sum_{i \in n}^{n} log(1 - d_i). \tag{4}$$

Remark. Although we assign the position labels y as 0 for all negative base crops q_{neg}, there are instances where the random crop k may resemble some q_{neg}. Without labels during pre-training, constructing absolutely ideal negative pairs in contrastive learning remains challenging [15], [38], [76]. Nevertheless, the overall distances among negative pairs remain substantial. Thus, following previous methods [15], [38], [75], we adopt the average entropy of distances in Eq. 4.

Intra-volume contrast. As shown in Fig. 4(b), we conduct intra-volume contrast on a triplet: random crop k, positive base crop q_{pos}, and negative base crop q_{neg}. Specifically, we pull k and q_{pos} closer, push k and q_{pos}, q_{pos} and q_{neg} apart from each other. For random crop k, we use position labels y to supervise the process of contrastive learning. For q, we design a regularization loss L_{reg} to enforce the feature discrepancy between each pair of q_i and q_j:

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

where s_{ij} is the cosine similarity between different q_i and q_j as follows:

$$s_{ij} = CosSim(q_i, q_j) = \frac{q_i \cdot q_j}{\parallel q_i \parallel \parallel q_j \parallel}, i, j \in n, i \neq j. \tag{6}$$

Inter-volume contrast. As shown in Fig. 4(c), we extract random crop k_A from volume V_A and base crops q_B from volume V_B to establish inter-volume contrast, where volumes V_A and V_B are sampled from the same batch during training. It is worth noting that V_A and V_B are sampled from the same anatomical region, *e.g.*, both from the abdomen region or both from the chest region.

Specifically, we adopt a simple-yet-effective consistency regularization method [15], [38], [74] for inter-volume contrast. We first employ feature augmentation (aug. in Fig. 4) to k_A, q_B and get $k_A^{'}, q_B^{'}$. The augmentation here is a simple Dropout [77] as [74]. Then we fed the features before and after augmentation into p_s and p_t, respectively. After the projectors, we also compute the cosine similarity for contrastive learning:

$$y_s = CosSim(k_A, q_B), y_t = CosSim(k_A^{'}, q_B^{'}). \tag{7}$$

Then we formulate the loss function L_{inter} as:

$$L_{inter} = -\frac{1}{n} \sum_{i \in n}^{n} log(1 - |y_s - y_t|). \tag{8}$$

Overall loss function for SSL. Overall, the loss function L_{SSL} for SSL is the combination:

$$L_{SSL} = L_{pred} + L_{reg} + L_{inter}, \tag{9}$$

where we empirically set the same weights for three loss functions [1], since we consider their importance equally.

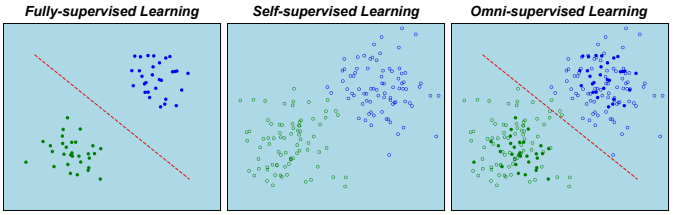

Fig. 5: Differences among fully-, self-, and omni-supervised learning. Solid and hollow markers denote labeled and unlabeled data, respectively. Dashed lines denote decision boundaries between different classes.

3.3 Towards Omni-supervised Pre-training

Algorithm 1: Omni-supervised Pre-training

Data: Labeled segmentation data: $(\boldsymbol{X_L}, \boldsymbol{Y_L})$.
Unlabeled data: $\boldsymbol{X_U}$
Result: Pre-trained model $\boldsymbol{M}$
First stage:
  Fully-supervised training $\boldsymbol{M} \leftarrow (\boldsymbol{X_L}, \boldsymbol{Y_L})$;
  Self-supervised training $\boldsymbol{M} \leftarrow \boldsymbol{X_U}$ [L_{SSL} Eq. 9];
Second stage:
  Generate pseudo labels $\boldsymbol{Y_U} \leftarrow (\boldsymbol{M}, \boldsymbol{X_U})$;
  Semi-supervised training $\boldsymbol{M} \leftarrow$ $(\boldsymbol{X_L}, \boldsymbol{Y_L}, \boldsymbol{X_U}, \boldsymbol{Y_U})$;
  Self-supervised training $\boldsymbol{M} \leftarrow \boldsymbol{X_U}$ [L_{SSL} Eq. 9];

As shown in Fig. 5, both fully- and self-supervised learning have specific merits and drawbacks. **(a)** Fully-supervised learning can learn discriminative decision boundaries with the guidance of labels yet it is constrained by the lack of labeled data. **(b)** SSL can leverage large-scale unlabeled data. However, lacking annotations for supervision, it struggles with learning clear decision boundaries between distinct classes. To this end, we propose omni-supervised pre-training to effectively leverage both labeled and unlabeled data, as described in Algorithm 1. Our omni-supervised learning amalgamates the strengths of both fully- and self-supervised learning and effectively unleashes the potential of labeled and unlabeled data.

Curating labeled segmentation dataset $\boldsymbol{X_L}, \boldsymbol{Y_L}$. The PreCT-160K dataset includes extensive labeled segmentation datasets. However, many of these datasets have incomplete labels [2], [6], [29], such as one dataset containing

TABLE 1: **PreCT-160K** contains **160K** CT from 30 public datasets, with more than **42M** slices covering the anatomical structures. 10K is used in our preliminary study [1].

Dataset	Anatomical Region	Pre-training Scale		Number of Volumes
		10K	160K	
BTCV [78]	Abdomen	✔	✔	24
TCIA-Covid19 [79]	Chest	✔	✔	722
LUNA16 [80]	Chest	✔	✔	843
FLARE23 [81]	Abdomen	✔	✔	4000
HNSCC [82]	Head/Neck	✔	✔	1071
STOIC 2021 [83]	Chest	✔	✔	2000
LIDC [84]	Chest	✔	✔	1018
TotalSegmentator [85]	104 Anatomic Structures	✔	✔	1203
Tumor datasets [2], [86], [87], [88], [89], [90]	Abdomen		✔	1334
WORD [91]	Abdomen		✔	120
AMOS22 [92]	Abdomen		✔	300
DeepLesion [93]	Abdomen		✔	1618
PANORAMA [94]	Abdomen		✔	2238
AbdomenAtlas1.0 [29]	Abdomen		✔	5195
OPC-Radiomics [95]	Oropharyngeal		✔	606
HeadNeckCT [96]	Head/Neck		✔	504
Qin-Headneck [97]	Head/Neck		✔	892
TCGA-HNSC [98]	Head/Neck		✔	1274
CT COLONOGRAPHY [99]	Chest, Abdomen, Colon cancer		✔	1730
MELA [100]	Chest		✔	770
StonyBrookChestCT [101]	Chest		✔	2316
CT-RATE [102]	Chest		✔	47149
NLST [103]	Chest		✔	84830
Total				**160167**

solely liver labels and another with only pancreas labels. Thus, we first ensemble various models to generate complete labels Y_L for X_L and curate a small subset of labeled data from PreCT-160K. This subset, which we named VoComni, contains 20K volumes spanning 20 different organ and tumor classes, which will be released with PreCT-160K for fostering future research. All the val and test sets are unseen in PreCT-160K and VoComni.

Semi-supervised learning is a scalable learner. To effectively leverage labeled and unlabeled data, we propose to conduct semi-supervised learning [72], [73], [74] to borrow the knowledge from labeled data to large-scale unlabeled data. Notably, segmentation emerges as a pivotal technique in supervised training [26], [67], [68], given that many medical tasks demand a granular understanding at the pixel level for accurate diagnosis. Previous works [7], [67], [104] only leveraged a few hundred cases for semi-supervised segmentation. However, complex designs of semi-supervised segmentation will significantly increase the burden of training, which is not feasible to our large-scale data. In this paper, we adopt a simple semi-supervised learning baseline and scale up the data to 160K volumes. We find that incorporated with VoCo, the simplest semi-supervised baseline can already achieve competitive results. As shown in Algorithm 1, we first curate a labeled segmentation dataset (X_L, Y_L) from PreCT-160K and perform supervised segmentation training in the first stage. Then in the second stage, we generate pseudo labels Y_U for unlabeled data X_U, aiming to perform semi-supervised learning on both X_L and X_U. **Note that** SSL is collaboratively integrated with semi-supervised training in both two stages. In this way, we amalgamate the strengths of self- and semi-supervised learning, advancing towards omni-supervised pre-training.

4 Experiments

4.1 Dataset and Implementation Details

Pre-training dataset[1]. In this paper, we curate the existing largest dataset medical image pre-training PreCT-160K, as shown in Table 1. PreCT-160K is collected from diverse sources and underwent thorough pre-processing to ensure a consistent data format for training. Specifically, to address variations in sizes, spacing, and intensity across volumes from different anatomical regions, we have devised tailored pre-processing protocols. Since in PreCT-160K, data from chest regions cover larger proportions, we simply balance the sampling during pre-training. For VoComni dataset[2], we use model ensembling to generate pseudo labels, where we discard the volumes with low prediction confidence. Consequently, we have created a segmentation dataset comprising 20K pseudo-labeled volumes (encompassing 20 organ and tumor classes) for our omni-supervised pre-training.

Evaluation benchmark. We build a large-scale evaluation benchmark as shown in Table 2, which includes 48 downstream datasets for various tasks. It can seen in Table 3 that our evaluation benchmark is more comprehensive than that of previous works [8], [9], [17], [18], [19], [26], [28]. A number of datasets [2], [81], [87], [92] are evaluated on the public leaderboards. If the test sets and public leaderboards are not available, we report the offline val sets results with the same data splits for fair comparisons.

Experiment settings. We first conduct pre-training on PreCT-160K then finetune the pre-trained models on 48 downstream datasets (Table 2) for evaluation. We adopt both SwinUNETR [126] and nnUNet [47] as the backbones

1. https://huggingface.co/datasets/Luffy503/PreCT-160K
2. https://huggingface.co/datasets/Luffy503/VoComni

TABLE 2: **48 downstream datasets in our benchmark**. **28** of them are unseen in pre-training (denoted with †). **18** of them are with less than 50 cases for finetuning (denoted with *). The labels of val sets are unseen in pre-training. Test sets are evaluated on public leaderboards (if available).

Dataset	Modality	Task
BTCV* [78]	CT	Abdomen Seg.
AMOS22 [92]	CT	Abdomen Seg.
WORD [91]	CT	Abdomen Seg.
FLARE22* [81]	CT	Abdomen Seg.
FLARE23 [81]	CT	Abdomen Seg.
Abdomenct1k [2]	CT	Abdomen Seg.
AbdomenAtlas [29]	CT	Abdomen Seg.
TotalSegmentator [85]	CT	104 Structures Seg.
MM-WHS†* [105]	CT	Heart Seg.
AVT†* [106]	CT	Aorta Seg.
CHAOS* [88]	CT	Liver Seg.
Sliver07†* [107]	CT	Liver Seg.
IRCADb†* [108]	CT	Liver Tumor Seg.
KiTS [87]	CT	Kidney Tumor Seg.
KiPA22† [109]	CT	Kidney Tumor Seg.
TCIA-Panc.* [89]	CT	Panc. Seg.
PANORAMA [94]	CT	Panc. Tumor Seg.
SegThor†* [110]	CT	Thoracic Risk Seg.
BHSD† [111]	CT	Brain Bleed Seg.
StructSeg19†* [112]	CT	Nasopharynx Cancer Seg.
Verse20† [113]	CT	Vertebrae Seg.
Covid-19-20† [114]	CT	Covid Seg.
FUMPE†* [115]	CT	Pulmonary Embolism Seg.
Parse22† [116]	CT	Pulmonary Artery Seg.
AIIB23† [117]	CT	Fibrotic Lung Seg.
CC-CCII† [118]	CT	Covid Classi.
LUNA16 [80]	CT	Lung Nodule Classi.
AutoPET-II23† [119]	PET-CT	HeadNeck Lesion Seg.
AMOS-MRI†* [92]	MRI	Abdomen Seg.
MM-WHS-MRI†* [105]	MRI	Heart Seg.
ACDC† [120]	MRI	Heart Seg.
ATLAS-MRI†* [121]	MRI	Liver Tumor Seg.
BraTs21† [122]	MRI	Brain Tumor Seg.
IXI† [123]	MRI	Brain MRI Registration
OASIS† [124]	MRI	Brain MRI Registration
CTRG-Chest† [125]	VLP	Report Generation
CT-RATE [102]	VLP	Vocabulary Classi.
CT-RATE [102]	VLP	Report-Volume Retrieval
MSD Challenge [86]		
Task01 Brain†	MRI	Brain Tumor Seg.
Task02 Heart†*	MRI	Heart Seg.
Task03 Liver	CT	Liver Tumor Seg.
Task04 Hip.†	MRI	Hip. Seg.
Task05 Pros.†*	MRI	Prostate Seg.
Task06 Lung*	CT	Lung Cancer Seg.
Task07 Panc.	CT	Pancreas Tumor Seg.
Task08 Vessel†	CT	Vessel Tumor Seg.
Task09 Spleen*	CT	Spleen Seg.
Task10 Colon	CT	Colon Cancer Seg.

TABLE 3: **Our benchmark is more comprehensive**, with more tasks and data for evaluation. **Eval Sets** denote the number of downstream datasets.

Method (Publication)	Downstream Tasks				Eval Sets
	Seg.	Cls.	Reg.	VL	
PCRL [8] (TPAMI23)	✔	✔			4
GVSL [28](CVPR23)	✔	✔			4
Swin. [17] (CVPR23)	✔				11
Alice [18] (ICCV23)	✔				3
Univer. [68] (ICCV23)	✔				14
Unim. [9] (TPAMI24)	✔	✔			10
MedC. [19] (CVPR24)	✔				9
VoCo-v1 [1] (CVPR24)	✔	✔			6
SuPrem [26] (ICLR24)	✔				3
VoCo	✔	✔	✔	✔	**48**

for pre-training. Specifically, Swin-Base (B), Swin-Large (L), and Swin-Huge (H) are all adopted for training, with feature sizes of 48, 96, and 192 in SwinUNETR [126], respectively. This project is supported by NVIDIA SuperPOD hardware. 8 × NVIDIA H800 GPUs are used for pre-training and all the downstream tasks can be done with one H800 or 3090 GPU. It spent over 10,000 GPU hours in downstream evaluation. Our implementation codes are all open-source and support both MONAI [127] and nnUNet [47] frameworks.

4.2 Comparison with State-of-the-Art Methods

We perform in-depth comparisons with previous methods [8], [9], [17], [26], [27], [47], [53], [65], [67], [68] that **have released their codes and checkpoints**. **Note that** in instances where certain datasets necessitate extensive computational resources or involve limited cases, we exclusively report the results of methods [17], [26], [47] that with better performances. Our evaluations span across segmentation, classification, registration, and vision-language tasks. In following discussions, the term ***baseline*** **denotes adopting the same backbones but without pre-training (from scratch)**.

4.2.1 *Medical Image Segmentation*

Seven widely-used segmentation datasets. As shown in Table 4, on seven widely-used segmentation datasets, VoCo demonstrates leading performances and surpass previous methods [9], [17], [26], [27], [47], [53], [67], [68] by a clear margin. It can be seen that the general method MoCo-v3 [15], [38] did not perform well on medical tasks. Since MoCo v3 [15], [38] heavily relies on a large batch size to acquire adequate negative samples, which is not feasible in 3D medical images. Moreover, the negative relation between different images used in MoCo v3 [15], [38] is not appropriate in medical images.

Notably, VoCo outperforms the baseline by average $\mathbf{3.12}\%$ DSC. SuPreM [26] achieved the best results among the previous pre-training methods since it used an abdomen dataset [29] for supervised pre-training and the datasets in Table 4 are almost abdomen datasets. VoCo surpasses SuPreM [26] and achieves new state-of-the-art performances. Specifically, for the challenging ToTalSegmentor [85] dataset, VoCo (Swin-H) outperforms SuPreM [26] by $\mathbf{3.22}\%$ DSC. The overall results in Table 4 vividly underscore the effectiveness of our method.

24 organ/tumor segmentation tasks. As shown in Table 5, we report the results on 24 organ and tumor segmentation datasets, across different modalities and anatomical regions as shown in Table 2. Notably, models with VoCo pre-training outperform those without pre-training by average $\mathbf{4.42}\%$ DSC. It is worth noting that a majority of these datasets contain fewer than 50 annotated cases for fine-tuning, highlighting the effectiveness of pre-training as a label-efficient solution. The overall improvements observed across these 24 datasets serve as compelling evidence for the efficacy of our proposed large-scale pre-training method.

MSD Challenge. Table 6 reports the results on the MSD 10-Task [86] dataset. We adopt the settings of nnUNet [47] for fair comparisons. Notably, with VoCo pre-training, the segmentation DSC is improved by average $\mathbf{2.98}\%$.

TABLE 4: The DSC (%) of seven widely-used segmentation datasets, *i.e.*, BTCV [78], AMOS22 [92], WORD [91], FLARE22 [81], FLARE23 [81], TotalSegmentator [85], and AbdomenAtlas [29]. The state-of-the-art results among previous methods are underlined while the best results are **bolded**. Note that [26], [63], [68] are fully-supervised pre-training methods. Since [29], [81], [85] require huge computation costs, we only report the results of advanced methods for comparisons. Compared with models without pre-training, VoCo pre-training brings average **+3.12**% DSC improvements.

Method	Model (Params)	Data Scale	BTCV	AMOS	WORD	FLA22	FLA23	Total	Atlas	△%(AVG)
From Scratch	3D-UNet (19M)		80.98	84.02	83.21	87.58	-	-	-	
	UNETR [128] (115M)		79.82	82.52	79.77	89.02	-	-	-	
	nnUNet [47] (31M)		79.29	88.79	86.04	91.38	91.54	82.26	88.77	
	Swin-B [126] (72M)		82.79	87.19	84.56	89.18	89.72	80.97	88.12	
	Swin-L [126] (290M)		83.52	87.53	83.17	89.49	89.33	82.04	88.56	
	Swin-H [126] (1.2B)		79.36	86.14	82.46	87.96	89.71	82.78	88.97	
General Pre-training										
MAE3D [16], [21], [31]	UNETR	10k	82.48	82.71	74.27	89.31	-	-	-	
MoCo v3 [15], [38]	Swin-B	1.6k	79.54	80.95	71.16	83.22	-	-	-	
Medical Pre-training										
MG [53]	3D-UNet	0.6k	81.45	81.27	85.50	85.02	-	-	-	
DoDNet [67]	3D-UNet	1k	81.10	79.63	85.90	86.19	-	-	-	
Unimiss [9], [27]	MiT(48M)	5k	82.05	86.26	83.37	89.17	-	-	-	
SwinUNETR [17]	Swin-B	5k	82.58	85.68	84.88	89.31	-	-	-	
Universal Model [68]	Swin-B	2.1k	83.74	88.01	85.19	91.11	-	-	-	
SuPreM [26]	Swin-B	2.1k	85.32	88.14	85.97	91.37	89.98	82.96	89.16	
VoCo-v1 [1]	Swin-B	10k	84.51	88.06	86.11	91.29	90.07	80.46	89.13	
VoCo	Swin-B	160k	**86.64**	**90.86**	**86.88**	92.17	90.34	84.84	90.38	↑2.79
VoCo	Swin-L	160k	86.05	89.43	86.77	92.37	91.56	85.27	90.90	↑2.67
VoCo	Swin-H	160k	86.21	88.79	86.12	**92.65**	**92.30**	**86.18**	**91.41**	↑3.75

TABLE 5: The DSC (%) of **24** downstream segmentation tasks. nnUNet [47] and Swin-B [126] are from scratch, others are with pre-training. **Note that** the improvements are more significant for the datasets with fewer cases for finetuning (refer to Table 2). Compared with models without pre-training, VoCo pre-training brings average **+4.42**% DSC improvements.

Method	Ab1k [2]	WHS [105]	AVT [106]	CHAOS [88]	Sliver. [107]	IR. [108]	KiTs [87]	Kipa. [109]
nnUNet [47]	85.74	88.72	50.19	94.53	94.87	51.26	78.92	88.99
Swin-B [126]	85.76	89.11	46.76	94.10	94.96	57.19	78.61	85.18
SwinUNETR [17]	86.32	89.06	46.18	94.98	94.67	55.69	76.82	85.14
SuPrem [26]	86.40	90.88	58.85	96.42	96.72	68.48	78.38	85.76
VoCo (nnUNet)	86.75	89.53	58.23	96.01	95.98	60.84	80.80	**90.31**
VoCo (Swin-B)	**87.77**	**91.22**	**69.64**	**96.68**	**97.75**	**74.27**	**80.81**	87.54
△(nnUNet)	↑1.01	↑0.81	↑8.04	↑1.48	↑1.11	↑9.58	↑1.88	↑1.32
△(Swin-B)	↑2.01	↑2.11	↑22.88	↑2.58	↑1.79	↑17.08	↑2.20	↑2.36

Method	Panc. [89]	PANO. [88]	Segthor [110]	BHSD [111]	Struct. [112]	Verse. [113]	Covid. [114]	FUMPE [115]
nnUNet [47]	84.68	78.06	88.15	35.02	70.60	65.13	62.42	48.62
Swin-B [126]	84.38	78.40	87.90	36.40	**76.42**	62.01	63.91	50.31
SwinUNETR [17]	84.53	78.34	87.23	35.97	53.36	87.33	65.90	51.72
SuPrem [26]	85.19	79.92	89.70	32.82	59.85	89.54	63.29	51.98
VoCo (nnUNet)	**87.59**	79.52	88.82	37.04	72.74	**67.82**	65.35	49.50
VoCo (Swin-B)	86.57	**80.13**	**90.17**	**38.38**	75.58	63.72	**68.72**	**55.32**
△(nnUNet)	↑2.91	↑1.48	↑1.46	↑2.02	↑2.15	↑2.69	↑2.93	↑0.89
△(Swin-B)	↑2.19	↑1.73	↑2.38	↑1.98	↓1.14	↑1.71	↑4.82	↑5.01

Method	Parse. [116]	AIIB. [117]	Auto. [119]	AM-MR. [92]	WHS-MR [105]	ACDC [120]	At-MR [121]	BraTs. [122]
nnUNet [47]	80.55	88.72	35.84	72.56	85.36	92.12	63.23	**91.02**
Swin-B [126]	82.78	89.09	25.25	72.46	86.13	87.22	60.40	89.05
SwinUNETR [17]	81.66	89.05	22.09	72.89	86.29	89.47	60.51	87.33
SuPrem [26]	82.88	89.96	24.68	75.69	85.79	89.10	64.64	89.54
VoCo (nnUNet)	81.60	90.12	33.02	74.38	86.26	**92.41**	68.19	90.51
VoCo (Swin-B)	**83.87**	**90.44**	32.61	**79.24**	**87.71**	89.51	**69.80**	90.23
△(nnUNet)	↑1.11	↑1.41	↓2.82	↑1.82	↑0.90	↑0.29	↑4.96	↓0.51
△(Swin-B)	↑1.10	↑1.35	↑7.36	↑6.78	↑1.58	↑2.28	↑9.40	↑1.18

TABLE 6: The DSC (%) of MSD 10-Task Challenge [86]. We report the results of the same folds defined by nnUNet [47] for fair comparison. We report the tumor DSC for Task03 and Task07. Compared with models without pre-training, VoCo pre-training brings average **+2.98**% DSC improvements.

Method	Task01	Task02	Task03	Task04	Task05	Task06	Task07	Task08	Task09	Task10
nnUNet [47]	71.25	91.84	67.43	87.08	79.16	64.73	46.08	68.92	92.57	41.69
Swin-B [126]	71.74	92.28	67.85	88.66	72.64	70.28	47.88	64.80	94.90	35.13
SwinUNETR [17]	72.79	92.06	64.72	87.01	73.76	71.64	48.31	60.72	95.02	26.19
SuPrem [26]	70.07	92.55	68.20	87.40	72.92	72.55	50.02	64.71	96.01	38.78
VoCo (nnUNet)	73.56	**93.92**	70.09	**89.10**	**80.09**	68.99	50.43	**70.61**	95.70	41.93
VoCo (Swin-B)	**73.94**	93.72	**71.22**	88.55	75.57	**75.74**	**51.34**	67.25	**96.12**	**42.57**
△(nnUNet)	↑2.30	↑2.08	↑2.66	↑2.02	↑0.93	↑4.26	↑4.35	↑1.69	↑3.13	↑0.24
△(Swin-B)	↑2.21	↑2.44	↑3.37	↓1.11	↑2.93	↑5.46	↑3.46	↑2.45	↑1.22	↑7.44

TABLE 7: The accuracy (%) of medical image classification on CC-CCII [118] and LUNA16 [80] datasets. Underline are the baseline performances.

Method	CC-CCII [118]	LUNA16 [80]
From Scratch		
3D UNet [129]	89.07	98.27
Swin-B [126]	91.04	97.54
With Pre-training		
Jigsaw [130]	87.18	98.07
Rubik++ [65]	89.93	98.18
PCRLv2 [8]	89.35	98.30
SwinUNETR [17]	91.22	97.41
SuPreM [26]	91.83	97.53
VoCo	**93.80** (↑2.76)	**98.67** (↑1.13)

TABLE 8: The DSC (%) of medical image registration on IXI [123] and OASIS [124] datasets. Some preliminary results are drawn from TransMorph [131]. Underline are the baseline performances.

Method	IXI [123]	OASIS [124]
From Scratch		
VoxelMorph [132]	71.5	78.6
Siebert et al [133]	73.1	81.0
Mok et al [134]	73.5	82.0
TransMorph [131]	74.5	81.6
Swin-B [126]	72.6	81.8
With Pre-training		
SwinUNETR [17]	67.7	81.5
SuPreM [26]	72.9	81.2
VoCo	**73.6** (↑1.0)	**84.4** (↑2.6)

TABLE 9: The performances of the vision-language task: Report Generation on the CTRG dataset [125]. We use BLEU [135] to measure the accuracy. Some preliminary results only reporting BLEU-4 are drawn from CTRG [125].

Method	BLEU-1	BLEU-2	BLEU-3	BLEU-4
From Scratch				
Mesh-Memor [136]	-	-	-	21.0
RSTNet [137]	-	-	-	18.3
GSKET [138]	-	-	-	23.5
SL-DG [125]	-	-	-	23.7
Swin-B [126]	58.90	47.84	40.71	35.63
Swin-L [126]	59.15	48.87	40.89	36.04
Swin-H [126]	59.98	49.00	41.28	36.95
With Pre-training				
TransVW [54]	54.98	44.62	38.34	33.97
PCRLv2 [8]	58.11	46.96	40.08	35.26
SwinUNETR [17]	57.15	46.73	40.72	34.28
SuPreM [26]	58.70	48.13	39.96	35.23
VoCo (Swin-B)	60.45	49.35	42.30	37.42
VoCo (Swin-L)	61.79	**49.89**	42.25	36.85
VoCo (Swin-H)	**61.88**	49.82	**42.60**	**37.91**

TABLE 10: Performances of vocabulary classification and image-text retrieval on CT-RATE [102] dataset. †: **Note that** CT-CLIP [102] is based on image-text pre-training.

Method	**Voca. Classi.** (AUC%)	**Retrie.** (Recall%)
From Scratch		
3D UNet [129]	53.67	14.69
Swin-B [126]	67.86	18.12
Swin-L [126]	70.89	21.67
Swin-H [126]	70.33	20.28
With Pre-training		
CT-CLIP† [102]	**74.70**	23.46
SwinUNETR [17]	60.23	12.51
SuPreM [26]	65.18	20.23
VoCo (Swin-B)	71.28	23.57
VoCo (Swin-L)	72.61	23.79
VoCo (Swin-H)	73.69	**24.12**

4.2.2 Medical Image Classification

The medical image classification results on CC-CCII [118] and LUNA16 [80] are shown in Table 7. Given the near-optimal accuracy of lung nodule detection on LUNA16 [80], the benefits of pre-training are not obvious. For Covid classification on CC-CCII [118], VoCo outperforms the baseline by **2.76**% and SuPreM [26] by **1.97**%. Notably, SuPrem [26] conducted supervised segmentation pre-training on only abdomen datasets, potentially limiting its transferability to chest classification tasks.

4.2.3 Medical Image Registration

The medical image registration results on IXI [123] and OASIS [124] datasets are shown in Table 8. We adopt TransMorph [131] as the baseline. **Note that** in this paper we focus on evaluating the effectiveness of pre-training, thus we did not propose new registration algorithms. Thus, our registration analyses emphasize backbone comparisons (scratch versus pre-trained). We find that previous pre-training methods [17], [26] did not perform well on registration. While on brain MRI registration dataset OASIS [124], VoCo brings **2.6**% DSC improvements, which is a non-trivial improvement in registration.

4.2.4 Vision-Language Analysis

As shown in Table 3, this study pioneers the assessment of medical image pre-training efficacy in Vision-Language (VL) tasks. Specifically, we evaluate the report generation task on CTRG-Chest [139] and extend the evaluation to vocabulary classification and report-volume retrieval on the CT-RATE [102] dataset. The results are shown in Tables 9 and 10. **Note that** in this paper we focus on medical image pre-training, thus we verify the effectiveness via replacing the vision encoders. For the language models, we maintain the original settings from M2KT [139] and CT-CLIP [102] for CTRG-Chest and CT-RATE, respectively.

VoCo attains superior performances compared to previous medical image pre-training methods [8], [17], [26], [54]. Specifically, for report generation in Table 9, VoCo (Swin-H) achieves the highest score BLEU-4 (37.91%). In Table 10, VoCo (Swin-H) achieves 73.69% AUC in vocabulary classification and 24.12% recall in report-volume retrieval. Although SuPreM [26] performs well in abdomen segmentation datasets, it falls short in enhancing chest VL tasks. The results from VoCo underscore the significance of a robust vision encoder in VL tasks, which can provide more precise visual information for language models.

4.2.5 Discussion

Overall improvements. As shown in Fig. 6, with the same backbone, VoCo outperforms the baseline (from scratch) by a clear margin. SuPreM [26] emerged as the top performer among previous pre-training methods [8], [9], [17], [27], [47], [53], [54], [65], [67], [68]. Specifically, VoCo surpasses SuPreM [26] by an average of **2.93%**, **3.72%**, **2.57%**, **2.18%**, **3.52%**, and **2.72%** on 24 organ segmentation datasets, 14 tumor segmentation datasets, 15 chest analysis datasets, 28 unseen datasets, 13 cross-modal datasets, and

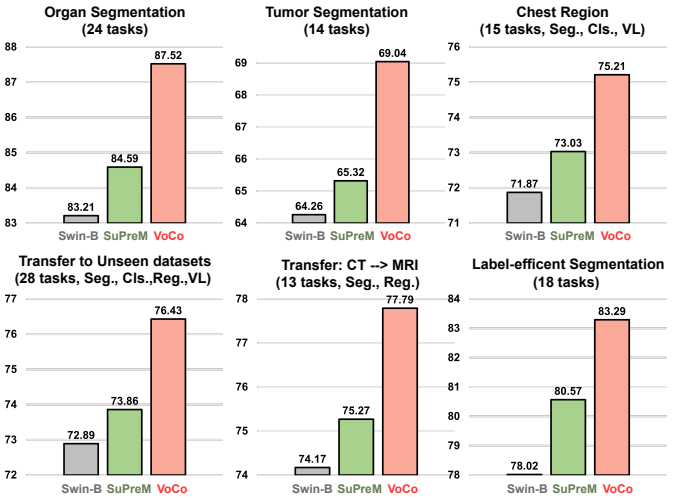

Fig. 6: **Overall comparisons**. Swin-B denotes using the randomly initialized SwinUNETR [126] as the backbone. Both SuPreM [26] and VoCo use Swin-B [126] as backbones for pre-training. Given the significant representation of chest datasets within our benchmark, we present the enhancement outcomes across 15 chest analysis tasks.

Fig. 7: **Efficient finetuning**. Analysis on BTCV [78], SegThor [110], and TotalSegmentator [85], where SegThor [110] is unseen in pre-training. Compared with the randomly initialized backbone Swin-B [126], VoCo achieves higher accuracy within fewer training epochs.

18 label-efficient segmentation datasets, respectively. Consistent improvements on 48 downstream datasets provide strong evidence of the effectiveness of VoCo.

Transferability to unseen datasets. As shown in Table 2, our evaluation benchmark encompasses 28 datasets unseen in pre-training. As shown in Fig. 6, VoCo demonstrates an average improvement of **3.53%** over the baseline Swin-B [126] when evaluated across these 28 unseen datasets.

Transferability to unseen modality. We conduct pre-training on CT datasets and subsequently transfer the learned models to another 3D medical imaging modality, *i.e.*, MRI. As shown in Table 2, our benchmark encompasses 13 MRI datasets spanning various tasks such as segmentation and registration. As shown in Fig. 6, VoCo yields an average improvement of **3.52%** across these 13 datasets, underscoring its efficacy in facilitating cross-modal transferability.

Label-efficient solution. In 3D medical image analysis, many datasets suffer from the scarcity of labeled data, primarily due to the substantial costs of annotation. As shown in Table 2, there are 18 segmentation datasets with less than 50 labeled cases for finetuning. As shown in Fig. 6, VoCo emerges as a label-efficient solution tailored for datasets constrained by limited labeled data, consistently delivering superior performances.

Pre-trained backbones. We use both nnUNet [47] and SwinUNETR [126] for pre-training. Although nnUNet [47] emerged as a strong segmentation baseline, it is not a scalable network architecture [63], with only 31M model params. Thus, we primarily focus on investigating the scaling law of SwinUNETR. Our analysis reveals that the pre-trained SwinUNETR [126] exhibits more substantial enhancements compared to the pre-trained nnUNet, *i.e.*, **+3.34%** and **+1.98%** DSC on 34 segmentation datasets (Table 5 and 6). The relatively modest improvements observed in pre-trained nnUNet could potentially stem from variations in pre-processing strategies, given nnUNet's reliance on a distinct data-fingerprint processing technique.

Efficient finetuning. Previous works [12], [20], [26] proved that strong pre-training models can notably expedite training convergence, resulting in improved performance with fewer training epochs. As shown in Fig. 7, VoCo substantially expedites the training convergence speed on BTCV [78], Segthor [110], and TotalSegmentator [85], and this phenomenon is generalized in all 48 tasks. This is a non-trivial contribution to efficient finetuning, particularly beneficial for datasets demanding extensive computational resources [85]. Our pre-trained models are poised to save computation costs in medical image analysis, making a strong step towards efficient learning.

Failure cases. Although consistent improvements (at least 1%) are observed on 48 datasets, marginal improvements persist in a handful of cases. Specifically, VoCo gains less than 1.5% improvements on 5 of 48 datasets. The presence of challenging datasets, *e.g.*, Positron Emission Tomography (PET) dataset AutoPET [119] poses unique obstacles, primarily due to their distinct imaging characteristics compared to our pre-training datasets. These differences result in domain gaps that constrain the effectiveness of our pre-training.

4.3 Scaling Law in Medical Image Analysis

Are larger models always better? In medical tasks, the answer appears to be ***no***. It can be observed from Fig. 8 that for some specific tasks, models with smaller sizes can achieve better performances. In this paper, we delves into factors affecting the model capacity scaling law, including: *number of finetuning cases, data diversity, and task difficulties*.

As shown in Fig. 8, **(a)** TotalSegmentator [85] is a challenging dataset, containing 1.2K cases and 104 classes for segmentation. In this case, the largest model VoCo-H yields the best results. **(b)** BTCV [78] is with only 24 cases for finetuning, potentially leading larger models to **overfit** on limited data, thus hindering validation performance. **(c)** Although CC-CCII [118] encompasses 4.2K cases for training, it is a simple binary classification task (over 90% accuracy), suggesting that excessively large models may not be necessary. **(d)** OASIS [124] is brain MRI datasets with only 0.4K cases for registration and it also lacks significant structural diversity. In this case, the smallest VoCo-B delivers the best results. **(e)** CT-Rate [102] is with 50K cases for 18 classes vocabulary classification. Given large-scale data for training, larger models demonstrate higher performances.

Tailor different model sizes to various medical tasks. Drawing from experimental insights discussed above, we

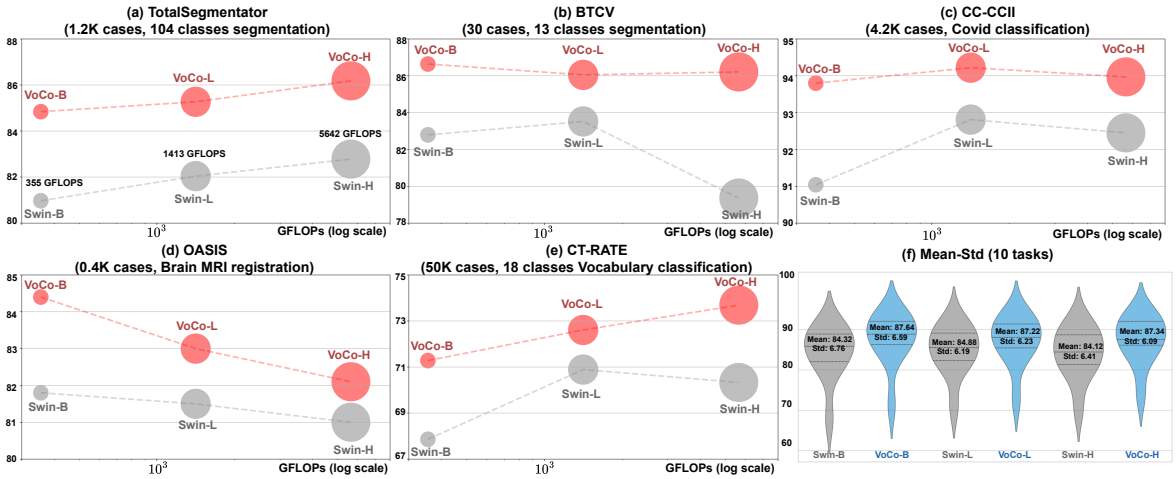

Fig. 8: ***Are larger models always better?*** The answer appears to be ***no***. We present the scaling results of TotalSegmentator [85], BTCV [78], CC-CCII [118], OASIS [124], and CT-RATE [102] in (a)-(e), respectively, covering various downstream tasks. We compared our pre-trainined models with the randomly initialized models [126], taking into account both accuracy and computation costs (GFLOPs computed for a 96×96×96 size of volume, shown in (a)). (f) presents the mean and standard deviation (STD) values across 10 downstream tasks [29], [78], [81], [85], [91], [92], [102], [118], [124].

Self		Semi	Total	BTCV	CCII	OAS.	CTRG
intra	inter						
✗	✗	✗	80.97	82.79	91.04	81.79	58.90
✓	✗	✗	81.38	84.51	92.85	82.34	59.37
✓	✓	✗	82.07	85.42	93.64	82.49	60.23
✗	✗	✓	84.02	85.37	91.98	82.12	59.13
✓	✓	✓	**84.84**	**86.64**	**93.80**	**84.43**	**60.45**

TABLE 11: Evaluation of self- and semi-supervised learning. We report the downstream results of VoCo-B on Total. [85], BTCV [78], CCII [118], OASIS [124], and CTRG [139].

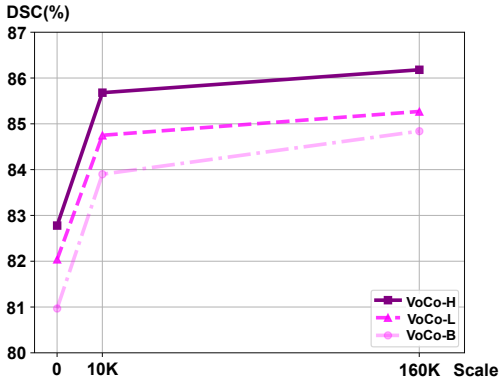

Fig. 9: **Data scaling law**. We scale up the data from 10K to 160K and report the DSC (%) of TotalSegmentator [85].

empirically propose simple and reasonable guidelines for tailoring various medical tasks: **(1)** Tasks with extensive labeled data for fine-tuning potentially benefit from larger models. **(2)** Tasks spanning diverse anatomical regions potentially benefit from larger models. **(3)** Tasks requiring recognition across a higher number of classes (more challenging) are better addressed with larger models.

Although these guidelines have been assessed on our comprehensive benchmark, they may not universally apply to all medical tasks given the substantial diversity within the medical domain. Thus, we release pre-trained models of varying sizes to aid researchers in selecting the most appropriate models for their specific needs.

4.4 Ablation Studies

Our preliminary investigation VoCo-v1 [1] has provided fundamental ablation studies, focusing on exploring various loss functions and hyperparameter configurations. Compared with VoCo-v1 [1], we further evaluate the effectiveness of volume contrast, omni-supervised learning, and data scaling from 10K to 160K. We use Swin-B [126] as the backbone and present the results on diverse datasets, including TotalSegmentator [85], BTCV [78], CC-CCII [118], OASIS [124], and CTRG [139], across segmentation, classification, registration, and vision-language tasks.

Volume contrast. As shown in Table 11, inter-volume contrast consistently enhances performance across five datasets. The combination of intra- and inter-volume contrast can yield higher improvements compared with the randomly initialized backbone [126].

Omni-supervised pre-training. As shown in Table 11, semi-supervised learning can effectively improve the performances. Specifically, for TotalSegmentator [85], it leads to substantial DSC improvements from **82.07**% to **84.84**%, which is a non-trivial boost in this challenging segmentation dataset. It is worth noting that the pure semi-supervised pre-training can achieve competitive results on segmentation tasks [78], [85], but it does not improve significantly in classification, registration, and VL tasks. Combined with self- and semi-supervised learning, the omni-supervised pre-training can achieve the best performances.

Data scaling law in medical image pre-training. We scale up the pre-training data (Table 1) from 10K to 160K and present the findings on the TotalSegmentator [85] dataset in Fig. 9, showcasing the impact of expanding the pre-training dataset. Notably, the enhancements from 10K to 160K appear marginal. This phenomenon could be attributed to factors like data quality and diversity, network scalability, or nearing the upper limit of improvement.

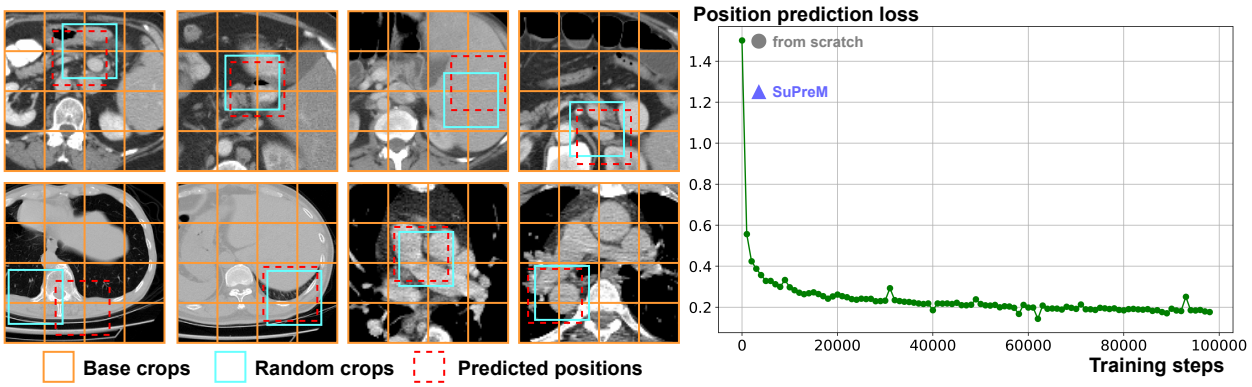

Fig. 10: **Case study for contextual position prediction**. **(1)** The left part shows the contextual position prediction results. Specifically, we set thresholds for the prediction logits to output the most probable positions. The predictions closely match the original positions of random crops. The bottom left is a failure case where two regions share similar structures. **(2)** As shown in the right part, the position prediction loss converges steadily during pre-training. We further verify the position prediction results of the model from scratch and the pre-trained SuPreM [26] model. Notably, through supervised segmentation pre-training, SuPreM [26] also enhances the contextual position prediction capability, implicitly indicating a positive correlation between segmentation performance and our proposed contextual position prediction.

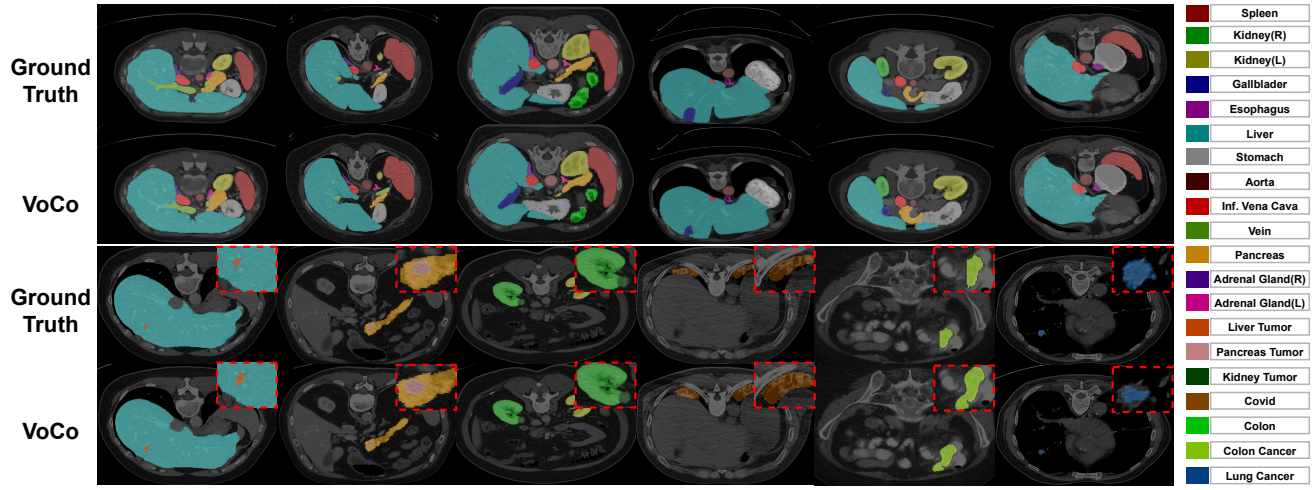

Fig. 11: Qualitative segmentation results on [78], [81], [86], [87], [114]. Tumor regions are zoomed in for better visualization.

4.5 Qualitative Visualization Results

Contextual Position Prediction. As shown in Fig. 10, we present some visualization results of contextual position prediction. The loss for contextual position prediction converges steadily during pre-training. The position predictions generated by VoCo closely align with the ground truth, underscoring the efficacy of our proposed pretext task.

Qualitative segmentation results. We present some visualization results in Fig. 11, which covers different anatomical regions. The visualization results demonstrate that our method can broadly apply to various downstream tasks.

5 Limitations and Future Directions

Although our pre-training method has demonstrated promising results across various medical tasks, there are still several limitations that can be further explored in the future:

- **Data engines for improving data quality**. The improvements become marginal when scaling the data from 10K to 160K. Although we have curated and pre-processed the pre-training dataset, the PreCT-160K still inevitably includes numerous low-quality cases. Data quality plays a pivotal role in pre-training to fully leverage the potential of large-scale datasets [20], [22], [24], [140]. In the future, we will focus on constructing data engines to improve the quality of datasets.
- **Data diversity to encompass distinctive image characteristics**. As discussed in Sec. 4.2.5, VoCo shows marginal enhancements in a few downstream tasks characterized by unique imaging features. Given the extensive diversity of medical datasets, we will further enhance the diversity of our pre-training dataset in future endeavors.
- **Multi-modal pre-training**. In this study, we exclusively utilize CT data for 3D medical image pre-training. In the future, we will also build a large-scale MRI pre-training dataset and combine with CT to facilitate multi-modal 3D medical image pre-training.
- **Scalable network architectures**. This study does not center on crafting novel network architectures for 3D medical image analysis. The current backbones [47],

[126] we employ may not exhibit scalability for large-scale pre-training. In the future, we will delve into the development of advanced network architectures for scalable pre-training.

- **Advance omni-supervised learning strategies**. As in Sec. 4.4, the omni-supervised pre-training yields better performances than the pure SSL method. In the future, our focus will shift toward omni-supervised learning techniques to effectively leverage both labeled and unlabeled data.

6 Conclusion

In this paper, we proposed a simple-yet-effective **Vo**lume **Co**ntrast (**VoCo**) framework for large-scale 3D medical image pre-training. Inspired by the consistent geometric relation between different organs, we propose to leverage the geometric context priors to learn consistent semantic representations for SSL. VoCo can also be seamlessly integrated into a semi-supervised learning framework for omni-supervised pre-training. To facilitate the study of large-scale 3D medical image pre-training, we curated the existing largest medical image pre-training dataset PreCT-160K, which encompasses 160K CT volumes (42M slices) covering diverse anatomical structures. We further delve into the scaling law of model capacity and propose the guidelines for tailoring different model sizes to various medical tasks. To evaluate the effectiveness of pre-training, we establish a comprehensive evaluation benchmark encompassing 48 downstream datasets across various tasks. Extensive experiments highlighted the superior performances of VoCo compared with previous methods.

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] L. Wu, J. Zhuang, and H. Chen, "Voco: A simple-yet-effective volume contrastive learning framework for 3d medical image analysis," in *CVPR*, 2024.
[2] J. Ma *et al.*, "Abdomenct-1k: Is abdominal organ segmentation a solved problem?" *TPAMI*, vol. 44, no. 10, pp. 6695–6714, 2021.
[3] Z. Qiu *et al.*, "Rethinking dual-stream super-resolution semantic learning in medical image segmentation," *TPAMI*, 2023.
[4] D. O. Medley *et al.*, "Cycoseg: A cyclic collaborative framework for automated medical image segmentation," *TPAMI*, vol. 44, no. 11, pp. 8167–8182, 2021.
[5] R. Azad *et al.*, "Medical image segmentation review: The success of u-net," *TPAMI*, 2024.
[6] Y. Xie *et al.*, "Learning from partially labeled data for multi-organ and tumor segmentation," *TPAMI*, 2023.
[7] F. Wu and X. Zhuang, "Minimizing estimated risks on unlabeled data: A new formulation for semi-supervised medical image segmentation," *TPAMI*, vol. 45, no. 5, pp. 6021–6036, 2022.
[8] H.-Y. Zhou *et al.*, "A unified visual information preservation framework for self-supervised pre-training in medical image analysis," *TPAMI*, 2023.
[9] Y. Xie *et al.*, "Unimiss+: Universal medical self-supervised learning from cross-dimensional unpaired data," *TPAMI*, 2024.
[10] A. Taleb *et al.*, "3d self-supervised methods for medical imaging," *NeurIPS*, vol. 33, pp. 18 158–18 172, 2020.
[11] K. Grünberg *et al.*, "Annotating medical image data," *MedIA*, pp. 45–67, 2017.
[12] M. Caron *et al.*, "Emerging properties in self-supervised vision transformers," in *ICCV*, 2021.
[13] J. Zhou *et al.*, "ibot: Image bert pre-training with online tokenizer," *arXiv preprint arXiv:2111.07832*, 2021.
[14] T. Chen *et al.*, "A simple framework for contrastive learning of visual representations," in *ICML*, 2020, pp. 1597–1607.
[15] X. Chen *et al.*, "An empirical study of training self-supervised vision transformers," *arXiv preprint arXiv:2104.02057*, 2021.
[16] J. Zhuang *et al.*, "Mim: Mask in mask self-supervised pre-training for 3d medical image analysis," *arXiv preprint arXiv:2404.15580*, 2024.
[17] Y. Tang *et al.*, "Self-supervised pre-training of swin transformers for 3d medical image analysis," in *CVPR*, 2022, pp. 20 730–20 740.
[18] Y. Jiang *et al.*, "Anatomical invariance modeling and semantic alignment for self-supervised learning in 3d medical image analysis," in *ICCV*, 2023, pp. 15 859–15 869.
[19] Y. Ye *et al.*, "Continual self-supervised learning: Towards universal multi-modal medical data representation learning," in *CVPR*, 2024, pp. 11 114–11 124.
[20] M. Oquab *et al.*, "Dinov2: Learning robust visual features without supervision," *TMLR*, 2023.
[21] K. He *et al.*, "Masked autoencoders are scalable vision learners," in *CVPR*, 2022, pp. 16 000–16 009.
[22] A. Kirillov *et al.*, "Segment anything," in *ICCV*, 2023, pp. 4015–4026.
[23] Z. Chen *et al.*, "Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks," in *CVPR*, 2024, pp. 24 185–24 198.
[24] L. Yang *et al.*, "Depth anything: Unleashing the power of large-scale unlabeled data," in *CVPR*, 2024.
[25] H.-Y. Zhou *et al.*, "Preservational learning improves self-supervised medical image models by reconstructing diverse contexts," in *ICCV*, 2021, pp. 3499–3509.
[26] W. Li, A. Yuille, and Z. Zhou, "How well do supervised models transfer to 3d image segmentation?" in *ICLR*, 2024.
[27] Y. Xie *et al.*, "Unimiss: Universal medical self-supervised learning via breaking dimensionality barrier," in *ECCV*, 2022, pp. 558–575.
[28] Y. He *et al.*, "Geometric visual similarity learning in 3d medical image self-supervised pre-training," in *CVPR*, 2023, pp. 9538–9547.
[29] C. Qu *et al.*, "Abdomenatlas-8k: Annotating 8,000 ct volumes for multi-organ segmentation in three weeks," *NeurIPS*, vol. 36, 2024.
[30] X. Zhuang *et al.*, "Self-supervised feature learning for 3d medical images by playing a rubik's cube," in *MICCAI*, 2019, pp. 420–428.
[31] Z. Chen *et al.*, "Masked image modeling advances 3d medical image analysis," in *WACV*, 2023, pp. 1970–1980.
[32] L. Yang *et al.*, "Depth anything v2," *NeurIPS*, 2024.
[33] J. Deng *et al.*, "Imagenet: A large-scale hierarchical image database," in *CVPR*, 2009, pp. 248–255.
[34] M. o. Caron, "Unsupervised learning of visual features by contrasting cluster assignments," *NeurIPS*, vol. 33, pp. 9912–9924, 2020.
[35] G. Larsson *et al.*, "Colorization as a proxy task for visual understanding," in *CVPR*, 2017, pp. 6874–6883.
[36] D. Pathak *et al.*, "Context encoders: Feature learning by inpainting," in *CVPR*, 2016, pp. 2536–2544.
[37] M. Chen *et al.*, "Generative pretraining from pixels," in *ICML*, 2020, pp. 1691–1703.
[38] K. He *et al.*, "Momentum contrast for unsupervised visual representation learning," in *CVPR*, 2020, pp. 9729–9738.
[39] J.-B. Grill *et al.*, "Bootstrap your own latent-a new approach to self-supervised learning," *NeurIPS*, vol. 33, pp. 21 271–21 284, 2020.
[40] Y. Gao *et al.*, "Disco: Remedying self-supervised learning on lightweight models with distilled contrastive learning," in *ECCV*, 2022, pp. 237–253.
[41] L. Wu *et al.*, "Sparsely annotated semantic segmentation with adaptive gaussian mixtures," in *CVPR*, 2023, pp. 15 454–15 464.
[42] X. Chen and K. He, "Exploring simple siamese representation learning," in *CVPR*, 2021, pp. 15 750–15 758.
[43] A. Dosovitskiy, L. Beyer *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *ICLR*, 2020.
[44] R. J. Chen *et al.*, "Towards a general-purpose foundation model for computational pathology," *Nature Medicine*, vol. 30, no. 3, pp. 850–862, 2024.

[45] E. Vorontsov *et al.*, "A foundation model for clinical-grade computational pathology and rare cancers detection," *Nature Medicine*, pp. 1–12, 2024.

[46] A. Vaidya *et al.*, "Demographic bias in misdiagnosis by computational pathology models," *Nature Medicine*, vol. 30, no. 4, pp. 1174–1190, 2024.

[47] F. Isensee *et al.*, "nnu-net: a self-configuring method for deep learning-based biomedical image segmentation," *Nature Methods*, vol. 18, no. 2, pp. 203–211, 2021.

[48] J.-X. Zhuang *et al.*, "Advancing volumetric medical image segmentation via global-local masked autoencoder," *arXiv preprint arXiv:2306.08913*, 2023.

[49] Y. Wang *et al.*, "Swinmm: masked multi-view with swin transformers for 3d medical image segmentation," in *MICCAI*, 2023.

[50] G. Wang *et al.*, "Mis-fm: 3d medical image segmentation using foundation models pretrained on a large-scale unannotated dataset," *arXiv preprint arXiv:2306.16925*, 2023.

[51] Y. He *et al.*, "Foundation model for advancing healthcare: Challenges, opportunities, and future directions," *arXiv preprint arXiv:2404.03264*, 2024.

[52] S. He *et al.*, "Meddr: Diagnosis-guided bootstrapping for large-scale medical vision-language learning," *arXiv preprint arXiv:2404.15127*, 2024.

[53] Z. Zhou *et al.*, "Models genesis," *MedIA*, vol. 67, p. 101840, 2021.

[54] F. Haghighi *et al.*, "Transferable visual words: Exploiting the semantics of anatomical patterns for self-supervised learning," *TMI*, vol. 40, no. 10, pp. 2857–2868, 2021.

[55] X. 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*, 2017, pp. 2097–2106.

[56] J. Irvin *et al.*, "Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison," in *AAAI*, vol. 33, no. 01, 2019, pp. 590–597.

[57] Y. Ye, "Meduniseg: 2d and 3d medical image segmentation via a prompt-driven universal model," *arXiv preprint arXiv:2410.05905*, 2024.

[58] T. M. Buzug, "Computed tomography," in *Springer handbook of medical technology*, 2011, pp. 311–342.

[59] P. J. Withers *et al.*, "X-ray computed tomography," *Nature Reviews Methods Primers*, vol. 1, no. 1, p. 18, 2021.

[60] L. Wu *et al.*, "Freetumor: Advance tumor segmentation via large-scale tumor synthesis," 2024.

[61] Y.-C. Chou, "Embracing massive medical data," in *MICCAI*. Springer, 2024, pp. 24–35.

[62] Y. Fang *et al.*, "Eva: Exploring the limits of masked visual representation learning at scale," in *CVPR*, 2023, pp. 19 358–19 369.

[63] Z. Huang *et al.*, "Stu-net: Scalable and transferable medical image segmentation models empowered by large-scale supervised pre-training," 2023.

[64] X. He *et al.*, "Intra-and inter-slice contrastive learning for point supervised oct fluid segmentation," *TIP*, vol. 31, pp. 1870–1881, 2022.

[65] X. Tao *et al.*, "Revisiting rubik's cube: self-supervised learning with volume-wise transformation for 3d medical image segmentation," in *MICCAI*, 2020, pp. 238–248.

[66] F. Haghighi *et al.*, "Dira: Discriminative, restorative, and adversarial learning for self-supervised medical image analysis," in *CVPR*, 2022, pp. 20 824–20 834.

[67] J. Zhang *et al.*, "Dodnet: Learning to segment multi-organ and tumors from multiple partially labeled datasets," in *CVPR*, 2021, pp. 1195–1204.

[68] J. Liu *et al.*, "Clip-driven universal model for organ segmentation and tumor detection," in *ICCV*, 2023, pp. 21 152–21 164.

[69] Y. o. Shu, "Omni-training: bridging pre-training and meta-training for few-shot learning," *TPAMI*, 2023.

[70] X. Tan *et al.*, "Positive-negative receptive field reasoning for omni-supervised 3d segmentation," *TPAMI*, 2023.

[71] L. Wu *et al.*, "Modeling the label distributions for weakly-supervised semantic segmentation," *arXiv preprint arXiv:2403.13225*, 2024.

[72] L. Yang *et al.*, "St++: Make self-training work better for semi-supervised semantic segmentation," in *CVPR*, 2022, pp. 4268–4277.

[73] L. Wu *et al.*, "Querying labeled for unlabeled: Cross-image semantic consistency guided semi-supervised semantic segmentation," *TPAMI*, vol. 45, no. 7, pp. 8827–8844, Jul. 2023.

[74] L. Yang *et al.*, "Revisiting weak-to-strong consistency in semi-supervised semantic segmentation," in *CVPR*, 2023, pp. 7236–7246.

[75] A. v. d. Oord *et al.*, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.

[76] P. Khosla *et al.*, "Supervised contrastive learning," *NeurIPS*, vol. 33, pp. 18 661–18 673, 2020.

[77] N. Srivastava *et al.*, "Dropout: a simple way to prevent neural networks from overfitting," *JMLR*, vol. 15, no. 1, pp. 1929–1958, 2014.

[78] B. Landman *et al.*, "Miccai multi-atlas labeling beyond the cranial vault–workshop and challenge," in *MICCAI workshop*, vol. 5, 2015, p. 12.

[79] K. Clark *et al.*, "The cancer imaging archive (tcia): maintaining and operating a public information repository," *Jour. of Dig. Imag.*, vol. 26, pp. 1045–1057, 2013.

[80] A. Setio *et al.*, "Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: the luna16 challenge," *MedIA*, vol. 42, pp. 1–13, 2017.

[81] J. Ma *et al.*, "Fast and low-gpu-memory abdomen ct organ segmentation: the flare challenge," *MedIA*, vol. 82, p. 102616, 2022.

[82] A. Grossberg *et al.*, "Md anderson cancer center head and neck quantitative imaging working group," *The Cancer Imaging Archive*, 2020. [Online]. Available: https://doi.org/10.7937/k9/tcia.2020.a8sh-7363

[83] M.-P. Revel *et al.*, "Study of thoracic ct in covid-19: the stoic project," *Radiology*, vol. 301, no. 1, pp. E361–E370, 2021.

[84] S. G. Armato III *et al.*, "The lung image database consortium (lidc) and image database resource initiative (idri): a completed reference database of lung nodules on ct scans," *Medical physics*, vol. 38, no. 2, pp. 915–931, 2011.

[85] J. Wasserthal *et al.*, "Totalsegmentator: Robust segmentation of 104 anatomic structures in ct images," *Radiology: Artificial Intelligence*, vol. 5, no. 5, 2023.

[86] M. Antonelli *et al.*, "The medical segmentation decathlon," *Nature Commun.*, vol. 13, no. 1, p. 4128, 2022.

[87] N. Heller *et al.*, "The kits21 challenge: Automatic segmentation of kidneys, renal tumors, and renal cysts in corticomedullary-phase ct," 2023.

[88] A. E. Kavur *et al.*, "Chaos challenge-combined (ct-mr) healthy abdominal organ segmentation," *MedIA*, vol. 69, p. 101950, 2021.

[89] H. Roth *et al.*, "Data from pancreas-ct," *The Cancer Imaging Archive*, 2016.

[90] Q. Chen *et al.*, "Towards generalizable tumor synthesis," in *CVPR*, 2024.

[91] X. Luo *et al.*, "WORD: A large scale dataset, benchmark and clinical applicable study for abdominal organ segmentation from ct image," *MedIA*, vol. 82, p. 102642, 2022.

[92] Y. Ji *et al.*, "Amos: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation," *NeurIPS*, vol. 35, pp. 36 722–36 732, 2022.

[93] M. De Grauw *et al.*, "The uls23 challenge public training dataset," Oct. 2023. [Online]. Available: https://doi.org/10.5281/zenodo.10035161

[94] N. Alves *et al.*, "The PANORAMA Study Protocol: Pancreatic Cancer Diagnosis-Radiologists Meet AI," Feb. 2024. [Online]. Available: https://doi.org/10.5281/zenodo.10599559

[95] M. L. Welch *et al.*, "Computed tomography images from large head and neck cohort," *The Cancer Imaging Archive*, 2023. [Online]. Available: https://doi.org/10.7937/J47W-NM11

[96] M. Vallières *et al.*, "Data from head-neck-pet-ct," *The Cancer Imaging Archive*, 2017. [Online]. Available: https://doi.org/10.7937/K9/TCIA.2017.8oje5q00

[97] R. R. Beichel *et al.*, "Data from qin-headneck," *The Cancer Imaging Archive*, 2015. [Online]. Available: https://doi.org/10.7937/K9/TCIA.2015.K0F5CGLI

[98] M. L. Zuley *et al.*, "The cancer genome atlas head-neck squamous cell carcinoma collection (tcga-hnsc)," *The Cancer Imaging Archive*, 2016. [Online]. Available: https://doi.org/10.7937/K9/TCIA.2016.LXKQ47MS

[99] S. K *et al.*, "Data from ct-colonography," *The Cancer Imaging Archive*, 2015. [Online]. Available: https://doi.org/10.7937/K9/TCIA.2015.NWTESAY1

[100] S. Song *et al.*, "MELA Dataset: A Benchmark for Mediastinal Lesion Analysis," 2022. [Online]. Available: https://doi.org/10.5281/zenodo.6575197

[101] J. Saltz *et al.*, "Stony brook university covid-19 positive cases," *The Cancer Imaging Archive*, 2021. [Online]. Available: https://doi.org/10.7937/TCIA.BBAG-2923

[102] I. E. Hamamci *et al.*, "A foundation model utilizing chest ct volumes and radiology reports for supervised-level zero-shot detection of abnormalities," *arXiv preprint arXiv:2403.17834*, 2024.

[103] N. L. S. T. R. Team, "Data from the national lung screening trial (nlst)," *The Cancer Imaging Archive*, 2013. [Online]. Available: https://doi.org/10.7937/TCIA.HMQ8-J677

[104] H. Wang and X. Li, "Towards generic semi-supervised framework for volumetric medical image segmentation," *NeurIPS*, vol. 36, 2024.

[105] X. Zhuang, "Multivariate mixture model for myocardial segmentation combining multi-source images," *TPAMI*, vol. 41, no. 12, pp. 2933–2946, 2018.

[106] L. Radl *et al.*, "Avt: Multicenter aortic vessel tree cta dataset collection with ground truth segmentation masks," *Data in brief*, vol. 40, p. 107801, 2022.

[107] B. van Ginneken, "Sliver07 [data set]," Mar. 2019. [Online]. Available: https://doi.org/10.5281/zenodo.2597908

[108] L. Soler *et al.*, "3d image reconstruction for comparison of algorithm database," *Data*, 2010. [Online]. Available: https://www.ircad.fr/research/data-sets/liver-segmentation-3d-ircadb-01/

[109] Y. He *et al.*, "Meta grayscale adaptive network for 3d integrated renal structures segmentation," *MedIA*, vol. 71, p. 102055, 2021.

[110] Z. Lambert, C. Petitjean, B. Dubray, and S. Kuan, "Segthor: Segmentation of thoracic organs at risk in ct images," in *Inter. Conf. Image Process. Theory Tools. Appli.*, 2020, pp. 1–6.

[111] B. Wu *et al.*, "Bhsd: A 3d multi-class brain hemorrhage segmentation dataset," in *Inter. Workshop Machine Learn. Medical Imag.*, 2023, pp. 147–156.

[112] J. Shi, "Structseg2019 gtv segmentation," *IEEE Dataport*, 2023. [Online]. Available: https://dx.doi.org/10.21227/h75x-gt46

[113] A. Sekuboyina *et al.*, "Verse: A vertebrae labelling and segmentation benchmark for multi-detector ct images," *MedIA*, vol. 73, p. 102166, 2021.

[114] H. R. Roth *et al.*, "Rapid artificial intelligence solutions in a pandemic—the covid-19-20 lung ct lesion segmentation challenge," *MedIA*, vol. 82, p. 102605, 2022.

[115] M. Masoudi *et al.*, "A new dataset of computed-tomography angiography images for computer-aided detection of pulmonary embolism," *Scientific data*, vol. 5, no. 1, pp. 1–9, 2018.

[116] G. Luo *et al.*, "Efficient automatic segmentation for multi-level pulmonary arteries: The parse challenge," *arXiv preprint arXiv:2304.03708*, 2023.

[117] Y. Nan *et al.*, "Fuzzy attention neural network to tackle discontinuity in airway segmentation," *TNNLS*, 2023.

[118] K. Zhang *et al.*, "Clinically applicable ai system for accurate diagnosis, quantitative measurements, and prognosis of covid-19 pneumonia using computed tomography," *Cell*, vol. 181, no. 6, pp. 1423–1433, 2020.

[119] S. Gatidis *et al.*, "A whole-body fdg-pet/ct dataset with manually annotated tumor lesions," *Scientific Data*, vol. 9, no. 1, p. 601, 2022.

[120] O. Bernard *et al.*, "Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: is the problem solved?" *TMI*, vol. 37, no. 11, pp. 2514–2525, 2018.

[121] F. Quinton *et al.*, "A tumour and liver automatic segmentation (atlas) dataset on contrast-enhanced magnetic resonance imaging for hepatocellular carcinoma," *Data*, vol. 8, no. 5, p. 79, 2023.

[122] U. Baid *et al.*, "The rsna-asnr-miccai brats 2021 benchmark on brain tumor segmentation and radiogenomic classification," *arXiv preprint arXiv:2107.02314*, 2021.

[123] B. Kim *et al.*, "Cyclemorph: cycle consistent unsupervised deformable image registration," *MedIA*, vol. 71, p. 102036, 2021.

[124] D. S. Marcus *et al.*, "Open access series of imaging studies (oasis): cross-sectional mri data in young, middle aged, nondemented, and demented older adults," *Journal of Cognit. Neur.*, vol. 19, no. 9, pp. 1498–1507, 2007.

[125] Y. Tang *et al.*, "Work like a doctor: Unifying scan localizer and dynamic generator for automated computed tomography report generation," *Expert Sys. Appli.*, vol. 237, p. 121442, 2024.

[126] A. Hatamizadeh *et al.*, "Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images," in *MICCAI Workshop*, 2021, pp. 272–284.

[127] M. J. Cardoso *et al.*, "Monai: An open-source framework for deep learning in healthcare," *arXiv preprint arXiv:2211.02701*, 2022.

[128] A. Hatamizadeh *et al.*, "Unetr: Transformers for 3d medical image segmentation," in *WACV*, 2022, pp. 574–584.

[129] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *MICCAI*, 2015, pp. 234–241.

[130] P. Chen *et al.*, "Jigsaw clustering for unsupervised visual representation learning," in *CVPR*, 2021, pp. 11 526–11 535.

[131] J. Chen *et al.*, "Transmorph: Transformer for unsupervised medical image registration," *MedIA*, vol. 82, p. 102615, 2022.

[132] G. Balakrishnan *et al.*, "Voxelmorph: a learning framework for deformable medical image registration," *TMI*, vol. 38, no. 8, pp. 1788–1800, 2019.

[133] H. Siebert *et al.*, "Fast 3d registration with accurate optimisation and little learning for learn2reg 2021," in *MICCAI*, 2021, pp. 174–179.

[134] T. C. Mok and A. C. Chung, "Conditional deformable image registration with convolutional neural network," in *MICCAI*, 2021, pp. 35–45.

[135] K. Papineni *et al.*, "Bleu: a method for automatic evaluation of machine translation," in *ACL*, 2002, pp. 311–318.

[136] M. Cornia *et al.*, "Meshed-memory transformer for image captioning," in *CVPR*, 2020, pp. 10 578–10 587.

[137] X. Zhang *et al.*, "Rstnet: Captioning with adaptive attention on visual and non-visual words," in *CVPR*, June 2021, pp. 15 465–15 474.

[138] S. Yang *et al.*, "Knowledge matters: Chest radiology report generation with general and specific knowledge," *MedIA*, vol. 80, p. 102510, 2022.

[139] ——, "Radiology report generation with a learned knowledge base and multi-modal alignment," *MedIA*, vol. 86, p. 102798, 2023.

[140] L. Yang *et al.*, "Freemask: Synthetic images with dense annotations make stronger segmentation models," *NeurIPS*, vol. 36, 2024.