
Generative Text-Guided 3D Vision-Language Pretraining for Unified Medical Image Segmentation

Yinda Chen^{1,2,*}, Che Liu^{3,*}, Wei Huang¹, Sibo Cheng³, Rossella Arcucci³, Zhiwei Xiong^{1,2}

¹University of Science and Technology of China

²Institute of Artificial Intelligence, Hefei Comprehensive National Science Center

³Imperial College London

{cyd0806, weih5270}@mail.ustc.edu.cn

{lche.liu21, sibo.cheng, r.arcucci}@imperial.ac.uk,

zwxiong@ustc.edu.cn

Abstract

Vision-Language Pretraining (VLP) has demonstrated remarkable capabilities in learning visual representations from textual descriptions of images without annotations. Yet, effective VLP demands large-scale image-text pairs, a resource that suffers scarcity in the medical domain. Moreover, conventional VLP is limited to 2D images while medical images encompass diverse modalities, often in 3D, making the learning process more challenging. To address these challenges, we present Generative Text-Guided 3D Vision-Language Pretraining for Unified Medical Image Segmentation (GTGM), a framework that extends of VLP to 3D medical images without relying on paired textual descriptions. Specifically, GTGM utilizes large language models (LLM) to generate medical-style text from 3D medical images. This synthetic text is then used to supervise 3D visual representation learning. Furthermore, a negative-free contrastive learning objective strategy is introduced to cultivate consistent visual representations between augmented 3D medical image patches, which effectively mitigates the biases associated with strict positive-negative sample pairings. We evaluate GTGM on three imaging modalities - Computed Tomography (CT), Magnetic Resonance Imaging (MRI), and electron microscopy (EM) over 13 datasets. GTGM’s superior performance across various medical image segmentation tasks underscores its effectiveness and versatility, by enabling VLP extension into 3D medical imagery while bypassing the need for paired text.

1 Introduction

Vision-Language Pretraining (VLP) has achieved significant progress in Radford et al. [2021], Li et al. [2022b], Xue et al. [2022], Alayrac et al. [2022], owing to its capabilities in learning visual representations from textual descriptions of images without annotations. While VLP has been introduced to 2D medical image analysis recently, existing medical VLP works rely heavily on textual descriptions written by experienced experts, and the domain of 3D medical VLP remains largely unexplored Zhang et al. [2020], Huang et al. [2021], Boecking et al. [2022], Wang et al. [2022], Zhou et al. [2023a]. Despite the fact that 3D medical images, such as Computed Tomography (CT), Magnetic Resonance Imaging (MRI), and electron microscopy (EM), typically contain more valuable and clinically relevant information compared to 2D images, their utilization is hindered primarily due to the lack of associated 3D medical image-text datasets. Additionally, certain modalities of medical imaging, like EM, often do not have corresponding textual descriptions in real-world applications.

* Equal Contribution.

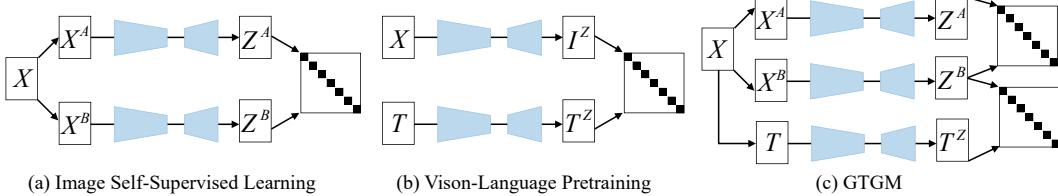


Figure 1: Comparison between our proposed approach and mainstream self-supervised learning (SSL) methods, where X and T represent images and text, respectively. (a) Image-only SSL with augmented views. (b) Pretraining with paired images and corresponding text. (c) Our GTGM framework pretrained with synthetic text-guided VLP and augmented-guided SSL.

To address the above challenges, we propose a framework called **Generative Text-Guided 3D Vision-Language Pretraining for Unified Medical Image Segmentation (GTGM)**. GTGM leverages the power of VLP in 3D medical image analysis by employing large language models (LLM) to generate medically relevant textual descriptions for 3D medical images. The main goal of GTGM is to learn general and robust 3D representations from these synthetic textual descriptions rather than specific organs and modalities. The differences between GTGM and existing image self-supervised algorithms can be summarized in Figure 1. GTGM integrates two learning objectives: acquiring visual-textual invariants from 3D medical images and synthetic text and extracting visual invariants from augmented 3D medical images. To learn general 3D visual representations, we introduce a negative-free contrastive learning strategy. This strategy aims to disentangle the variables in the latent space, rather than following traditional contrastive learning, which may carry biases due to the stringent assumption of one-to-one positive sample pairings.

We evaluate GTGM across various medical image segmentation tasks, covering commonly used modalities like CT and MRI over 10 datasets. We also extend our evaluation to the challenging modality of EM, especially the neuron segmentation task, over three datasets and multiple species. It is noteworthy that the inherent challenges of EM tasks, such as complex neuronal structures, inconsistent image quality, dense packing, and structural heterogeneity, make them significantly more difficult than other modalities Liu et al. [2022b], Huang et al. [2022]. Impressively, our GTGM attains state-of-the-art (SOTA) results on all EM tasks and nearly all CT and MRI tasks, despite not utilizing real textual descriptions during VLP. This accomplishment underscores the efficacy of synthetic text in guiding 3D medical VLP, which indicates the adaptability and potential of GTGM across a broad spectrum of 3D medical VLP applications. The contributions of this paper are as follows:

- Our GTGM framework is the first to showcase the effectiveness of 3D medical VLP, capable of learning 3D visual representations independent of specific organs or modalities. This fills the void in 3D medical visual learning within the scope of VLP. Furthermore, GTGM’s ability to pretrain without the need for expert-generated real text significantly alleviates one of the major challenges in medical VLP: the lack of large-scale image-text pairs.
- GTGM demonstrates superior performance and versatility across various medical image segmentation tasks, supporting different modalities like MRI, CT, and EM, and is adaptable to different species such as human and *Drosophila*. Moreover, GTGM excels in segmenting extremely small and densely packed structures in EM neuron images, expanding its applicability beyond organ and lesion segmentation in CT and MRI images.
- GTGM’s performance across diverse modalities, organs, and species, as well as its ability to handle varying densities and sizes of segmented objects, indicates its proficiency in learning a comprehensive and robust 3D medical image representation. In other words, GTGM provides an opportunity to generalize novel tasks through text-driven zero-shot medical image segmentation.

2 Related Work

Image Self-supervised Learning Self-supervised learning (SSL) has made significant advancements in computer vision by leveraging pretraining tasks without the need for annotations, as demonstrated by various pretext tasks Doersch et al. [2015], Gidaris et al. [2018], Noroozi and Favaro

[2016], Zhang et al. [2016]. Recently, contrastive learning has emerged as the standard method in SSL Grill et al. [2020], Zbontar et al. [2021], Misra and Maaten [2020], Chen and He [2021], Bardes et al. [2022]. To address the limitations of traditional contrastive learning, such as the requirement for large batch sizes and strong augmentations He et al. [2020], Chen et al. [2020], BYOL and BarlowTwins Grill et al. [2020], Zbontar et al. [2021] employ a dual-branch structure to align the embeddings of two augmented images, eliminating the need for negative samples in contrastive learning. SimSiam Chen and He [2021] demonstrates the importance of the stop-gradient mechanism on the dual-encoder, introducing a model without negative samples. In the context of SSL tailored for medical images, PCRLv2 Zhou et al. [2023b] combines contrastive learning with reconstruction pretasks. However, PCRLv2 has limitations in generalization across modalities, particularly in the case of extremely dense and small structures in EM images.

Medical Vision-Language Pretraining Medical VLP Zhang et al. [2020] has been introduced to integrate textual information into medical image SSL. However, the exploration of medical SSL VLP is primarily limited to 2D images, mainly due to the intricacy of medical reports and the scarcity of large-scale medical image-text datasets. Nonetheless, in the medical domain, 3D medical images (such as MRI, CT, and EM) assume a vital role and offer richer and more valuable information compared to their 2D medical images. Studies such as Zhang et al. [2020], Huang et al. [2021], Wang et al. [2022], Tiu et al. [2022] concentrate on the chest X-ray (CXR) domain; however, their applicability to other medical image modalities, including MRI, CT, and various 3D medical images, is yet to be established. In their work Liu et al. [2023], the authors develop a CT segmentation method that incorporates manually generated text describing the organs present in the image, based on corresponding annotations. However, their approach is limited by full supervision and the scale of annotations. Furthermore, their method can only process CT images. In recent works Butoi et al. [2023], Ye et al. [2023], methods are proposed that can process 3D medical segmentation tasks with different modalities. However, these approaches require large-scale well-annotated 3D medical images for supervised pretraining. Moreover, UniSeg Ye et al. [2023] lacks the ability to learn rich textual information as their manually designed prompts only indicate the type of task without describing the images. Despite its importance, the generalizability of VLP to a wider range of medical applications is limited by the absence of publicly available datasets containing 3D medical image-text pairs, as well as the inability of experienced experts to describe certain modalities such as EM images.

3 Method

3.1 Overview

Our GTGM model is designed to learn general representations of unannotated 3D medical images from synthetic textual descriptions. Like other VLP models, GTGM incorporates both a visual encoder $f_I(\cdot)$ and a text encoder $f_T(\cdot)$ to extract representations from images and text respectively. However, GTGM uniquely leverages synthetic textual descriptions rather than the real paired text of 3D medical images, given the lack of public 3D medical image-text datasets. The framework is depicted in Figure 2.

3.2 Generating Textual Descriptions

In the process of generating textual descriptions for medical images, we designate a generator $G(\Theta)$, initialize with BLIP Li et al. [2022a] pretrained weight. This generator is subsequently finetuned on the MedICAT Subramanian et al. [2020] dataset, endowing the synthetic textual descriptions with biomedical style. It is crucial to underscore that MedICAT Subramanian et al. [2020] solely comprises 2D images drawn from biomedical literature, with the associated captions serving as textual descriptions rather than the real description from clinical expertise. Consequently, the generation phase is not limited to any real radiology datasets. During the finetuning phase, we consider a 2D medical image I and its corresponding textual description $T = \{t_1, t_2, \dots, t_n\}$ from MedICAT Subramanian et al. [2020], where n is indicative of the textual description's length. The primary objective here is to amplify the conditional probability of the text given one image I :

$$P(T | I) = \prod_{i=1}^n P(t_i | I, t_{<i}), \quad (1)$$

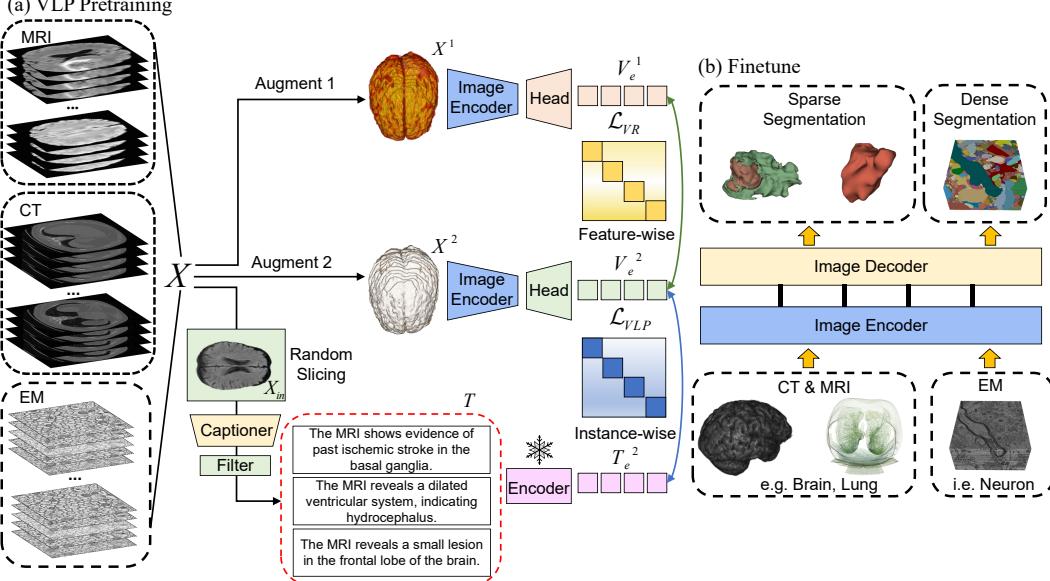


Figure 2: (a) The pipeline of GTGM, where we perform random cropping on medical images and utilize a pretrained generator to generate corresponding text. GTGM parallel learns visual invariants (feature-wise) and visual-textual invariants (instance-wise). (b) The finetuning process, where we employ the image encoder weights obtained during pretraining, along with a limited amount of labeled data, to perform downstream medical image segmentation tasks, including CT, MRI, and EM modalities. The \ddagger represents frozen weights.

where $t_{<i} = \{t_1, t_2, \dots, t_{i-1}\}$ denotes the generated text tokens. The conditional probability of each token t_i can be computed as:

$$P(t_i | I, t_{<i}) = \text{softmax}(W_o h_i + b_o), \quad (2)$$

where W_o and b_o represent the weights and biases of the output layer of $G(\Theta)$, respectively, and h_i is the hidden state of the $G(\Theta)$ at the i^{th} time step, which incorporates information from both the image I and the generated text tokens $t_{<i}$.

The learning objective in the finetuning generator stage is:

$$\mathcal{L}_{Cap} = - \sum_{i=1}^n \log P(t_i | I, t_{<i}). \quad (3)$$

In the generation phase, we take a set of N 3D medical images, $X = \{x_1, x_2, \dots, x_N\}$. For each 3D volume x_i , we randomly sample a 2D slice as an input for the finetuned generator $G(\Theta)$, which then generates the textual description T_i of 3D volume x_i . This can be mathematically formulated as $T_i = G(x_i|\Theta)$. After the generation phase, we filter out duplicate descriptions and remove certain fixed vocabulary that lacks information using regular expressions. To enhance the accuracy and distinctiveness of the textual descriptions, we prepend the name of the dataset to the beginning of each description. Each 3D medical image is then paired with a synthesized textual description, forming the image-text pairs $D = \{(x_1, T_1), (x_2, T_2), \dots, (x_N, T_N)\}$ for subsequent representation learning. Examples of the generated text can be found in the appendix.

3.3 3D Visual-Textual Representation Learning

Given the dataset \mathcal{D} of 3D medical images paired with synthetic text, we aim to learn the visual-textual representation via the image encoder be $f_I(\cdot)$ and the text encoder be $f_T(\cdot)$. The text encoder $f_T(\cdot)$, initialized with the weights from BioBERT Lee et al. [2020], is frozen during pretraining to maximize computational efficiency during the extraction of text embeddings.

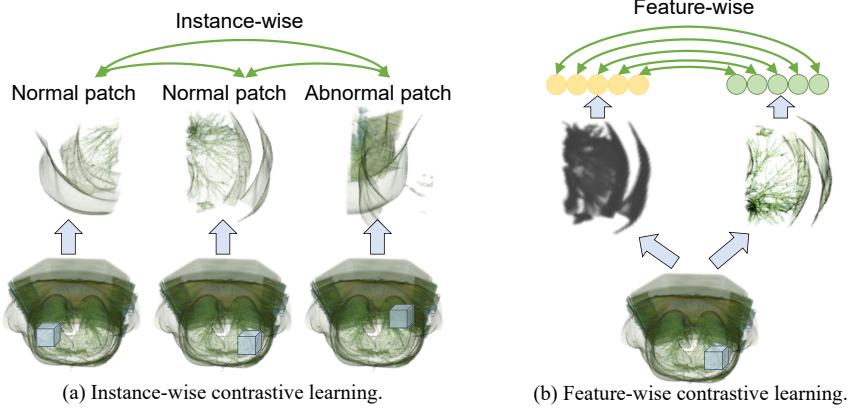


Figure 3: Graphically explanation of the bias introduced by instance-wise 3D image SSL and the effectiveness of our novel feature-wise 3D image SSL in mitigating bias arising from positive and negative sample pairs.

For a sample batch of image-text pairs (X_i, T_i) , we first compute their respective feature representations: $v_{e,i} = f_I(X_i)$ and $t_{e,i} = f_T(T_i)$.

we employ a contrastive learning objective to predict the matched pair $(v_{e,i}, t_{e,i})$ among $N \times N$ potential image-text pairs, while concurrently ensuring that $N^2 - N$ negative pairs are distinctly separated. Concretely, we utilize two non-linear visual and text projectors, \mathcal{F}_l and \mathcal{F}_v , to transform $\mathbf{v}_{e,i}$ and $\mathbf{t}_{e,i}$ into the same dimensional space d , where $\hat{\mathbf{v}}_{e,i} = \mathcal{F}_I(\mathbf{v}_{e,i})$, $\hat{\mathbf{t}}_{e,i} = \mathcal{F}_T(\mathbf{t}_{e,i})$, and $\{\hat{\mathbf{v}}_{e,i}, \hat{\mathbf{t}}_{e,i}\} \in \mathbb{R}^d$. Subsequently, we generate image vectors $[\hat{\mathbf{V}}_{e,i}]_{i=1}^N$ and text vectors $[\hat{\mathbf{T}}_{e,i}]_{i=1}^N$ within a training batch to compute cosine similarities:

$$\mathcal{L}_v^{v2t} = -\log \frac{\exp(s_{i,i}^{v2t}/\sigma_1)}{\sum_{j=1}^K \exp(s_{i,j}^{v2t}/\sigma_1)}, \quad \mathcal{L}_t^{t2v} = -\log \frac{\exp(s_{i,i}^{t2v}/\sigma_1)}{\sum_{j=1}^K \exp(s_{i,j}^{t2v}/\sigma_1)}, \quad (4)$$

where \mathcal{L}_v^{v2t} and \mathcal{L}_t^{t2v} are image-text and text-image InfoNCE Oord et al. [2018] contrastive loss, respectively. $s_{i,i}^{v2t} = \hat{\mathbf{v}}_{e,i}^\top \hat{\mathbf{t}}_{e,i}$ and $s_{i,i}^{t2v} = \hat{\mathbf{t}}_{e,i}^\top \hat{\mathbf{v}}_{e,i}$ represent image-text and text-image similarities. K is the batch size of each step. σ_1 is the temperature hyper-parameter set to 0.07 in our experiments.

The loss function can be articulated as:

$$\mathcal{L}_{VLP} = \frac{1}{2N} \sum_{i=1}^N (\mathcal{L}_v^{v2t} + \mathcal{L}_t^{t2v}). \quad (5)$$

Through overall loss \mathcal{L}_{VLP} , the model learns maximal mutual information between the matched multi-modal pairs containing cross-view attributes within a batch.

3.4 3D Visual Representation Learning

Image contrastive learning, commonly deployed to learn visual invariants, typically involves defining one positive sample (such as an augmented view) and treating the remainder of the batch’s samples as negatives. Nevertheless, this rigid 1-to-n positive-negative pairing tends to introduce substantial bias when learning 3D visual representation, particularly because in 3D medical imaging, each sample represents a patch of the original volume. Consequently, as shown in Figure 3, two slices could both represent normal organ semantics, even if their source volumes contain abnormal organs. Moreover, 1-to-n contrastive learning requires a large batch size Grill et al. [2020], Zbontar et al. [2021], which is not feasible in 3D visual learning tasks Liu et al. [2021].

To address these challenges, we introduce a negative-free learning objective instead of the rigid positive-negative based loss. This objective aims to disentangle the latent space feature-wisely and maximize the information in each feature dimension Zbontar et al. [2021].

We first generate two distinct views X^1 and X^2 of the medical volume X through random data augmentation. We initiate by normalizing the augmented embedding pairs $\{\mathbf{V}_e^1, \mathbf{V}_e^2\} \in \mathbb{R}^{N \times d}$ along the batch K dimension. This normalization ensures each feature dimension has a zero-mean and $1/\sqrt{K}$ standard deviation distribution, resulting in $\tilde{\mathbf{V}}_e$. Subsequently, we compute their cross-correlation $\tilde{\mathbf{V}}_e^{\text{corr}} = \tilde{\mathbf{V}}_e^{1\text{T}} \tilde{\mathbf{V}}_e^2$. The following defines the feature-dimension decorrelation formulas:

where N represents the batch size. Our objective is to minimize the off-diagonal elements of the cross-correlation matrix \hat{V}_e^{corr} and maximize the diagonal elements. The loss function can be formulated as:

$$\mathcal{L}_{VR} = \frac{1}{D'} \left\{ \underbrace{\sum_{j=1}^{D'} \left(1 - \sum_i^K \tilde{\mathbf{V}}_{e,i}^{1,j\text{T}} \tilde{\mathbf{V}}_{e,i}^{2,j} \right)^2}_{\text{cross-view invariants}} + \underbrace{\lambda_1 \sum_{j=1}^{D'} \sum_{i \neq j}^K \tilde{\mathbf{V}}_{e,i}^{1,j\text{T}} \tilde{\mathbf{V}}_{e,i}^{2,j}}_{\text{cross-view superfluity reduction}} \right\}, \quad \tilde{\mathbf{V}}_e = \frac{\mathbf{V}_e - \mu_K(\mathbf{V}_e)}{\sigma(\mathbf{V}_e)\sqrt{K}}. \quad (6)$$

Here, λ_1 is a non-negative hyperparameter used to adjust the trade-off between learning invariants and reducing superfluity in Equation 6. We set the value of λ_1 according to the default setting used in Zbontar et al. [2021]. The first term is crafted to learn a visual-invariant representation by optimizing the diagonal elements of the cross-correlation matrix \hat{V}_e^{corr} to be close to one. The second term is designed to lessen the correlation between distinct latent variables, thereby encouraging maximal information in each latent dimension by minimizing the off-diagonal elements in \hat{V}_e^{corr} . Finally, the loss is normalized along the feature dimension d .

The overall loss function can be articulated as:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{VLP} + \lambda_2 \mathcal{L}_{VR}, \quad (7)$$

The coefficients λ_1 and λ_2 are used to control the weights, and we set λ_1 and λ_2 to be 1 and 0.01.

4 Experiments

We conduct extensive experiments on a large number of cross-modal, cross-species, and cross-organ medical images. Generally speaking, image pretraining often encounters information bottlenecks, and the performance of downstream tasks does not improve with increasing amounts of data. However, due to the introduction of generated textual information, GTGM achieves good performance across a wide range of downstream tasks. In this section, we provide a detailed description of the pretraining and downstream task datasets, parameter settings, and report our experimental results.

4.1 Dataset and Metrics

Dataset. Our dataset encompasses imaging data from three modalities: CT, MRI, and EM. The primary sources of CT and MRI data are the Medical Segmentation Decathlon (MSD) Antonelli et al. [2022] competition, which includes 3D or 4D imaging of 10 different organs. During pretraining, we utilize all ImageTr and ImageTs, intentionally excluding labels. For the downstream segmentation tasks, we divide the ImageTr and corresponding ImageTs data into an 80% training set and a 20% test set. We conduct experiments using 1%, 10%, and 100% of the training set data (excluding the 1% setting when the data is insufficient to form a complete file). This allows us to evaluate the algorithm's performance under conditions of both label scarcity and abundance.

EM data primarily originate from large-scale EM datasets, namely the Full Adult Fly Brain (FAFB) Schlegel et al. [2021], MitoEM Wei et al. [2020], FIB-25 Takemura et al. [2017], and Kasthuri15 Kasthuri et al. [2015]. These datasets contain images from diverse organisms, including *Drosophila*, mice, rats, and humans. For the downstream segmentation tasks, we evaluate the algorithm's performance using three datasets from the CREMI competition Funke et al. [2016]. The CREMI dataset consists of three subsets: A, B, and C, each containing 125 images. We select the last 50 images from each subset for testing, while training is conducted using either the first 75 images or the first 10 images from each subset.

Table 1: Mean Dice scores (%) of CT image segmentation results. **Red** and **blue** entries denote the best and second-best results, respectively.

Method	Liver			Pancreas			Lung	
	1 %	10%	100%	1 %	10%	100%	10%	100%
Random	45.21	51.24	61.34	37.07	56.21	63.96	57.36	73.49
BYOL Grill et al. [2020]	45.11	52.33	61.67	39.83	56.8	64.51	59.84	76.41
SimSiam Chen and He [2021]	48.22	51.29	62.39	40.03	54.82	64.69	59.71	79.43
BarlowTwins Zbontar et al. [2021]	50.13	55.85	64.93	39.67	57.01	63.59	55.22	71.37
PCRLv2 Zhou et al. [2023b]	51.69	56.63	65.19	39.80	56.05	63.38	55.30	74.19
GTGM	52.46	58.67	65.61	40.55	59.96	65.61	61.3	80.19

Metrics. Our primary evaluation of the algorithm’s performance is conducted through downstream segmentation tasks. Among these, tasks involving CT and MRI scans fall into the category of semantic segmentation. Given the relatively small proportion of the entire volume occupied by the organs in these cases, we employ the Dice coefficient as a performance metric. In contrast, EM tasks are considered instance segmentation tasks. Here, there is no background in the volume and the neurons to be segmented are densely packed. Consequently, we use two metrics, Variation of Information (VOI) Nunez-Iglesias et al. [2013] and Adjusted Rand Index (ARAND) Arganda-Carreras et al. [2015], to evaluate the segmentation performance.

4.2 Implementation Details

Our pretraining is conducted on 8 NVIDIA A100 GPUs, including both text-image and image-only pretraining. The batch size is set to 16 per GPU, with an initial learning rate of 2e-5 and a learning rate decay of 5e-2. All pretraining tasks are iterated for 100k iterations. For all downstream tasks, we use 2 NVIDIA RTX 3090 GPUs or 1 NVIDIA A100 GPU. For CT and MRI tasks, we train for 40k iterations, while for EM tasks we train for 100k iterations. We utilize the AdamW optimizer with beta coefficients set to 0.9 and 0.999 for all tasks.

4.3 Experiment Results

We conduct extensive experiments on downstream applications involving organ-wise, modality-wise, and species-wise segmentation tasks. During the pretraining phase, we train the vision encoder using the same dataset as described earlier. In the finetuning phase, we concurrently update the parameters of the pretrained vision encoder and a randomly initialized decoder, using various label ratios. Our framework is compared with state-of-the-art self-supervised algorithms, including BYOL Grill et al. [2020], BarlowTwins Zbontar et al. [2021], and SimSiam Chen and He [2021] for natural images, as well as the latest SOTA algorithm specifically designed for medical images, PCRLv2 Zhou et al. [2023b]. Due to the fact that these baselines either do not provide a 3D vision encoder Grill et al. [2020], Zbontar et al. [2021], Chen and He [2021] or are only pretrained on limited 3D datasets Zhou et al. [2023b], we replicate these algorithms on our pretraining dataset and evaluate their performance on downstream tasks using the same experimental setup. To ensure compatibility with 3D medical images and enable a fair comparison, we adopt the 3D ResNet-50 He et al. [2016] as the vision encoder in both the pretraining and fine-tuning stages for all experiments. For pixel-level instance segmentation in downstream tasks, we employ a U-Net-style decoder. The results obtained using the Swin-Transformer-base vision encoder Liu et al. [2021] can be found in the appendix.

Experimental Results for CT Segmentation. CT imaging plays a crucial role in the medical field, particularly in lesion segmentation. However, the limited contrast differences between different tissues in CT images can lead to blurry boundaries between lesions and surrounding normal tissues. The segmentation results of six datasets of CT images are presented in Table 1. We achieve state-of-the-art (SOTA) performance on all datasets except for Hepatic Vessel. This discrepancy may be attributed to the fact that in the pretraining dataset, the Hepatic Vessel often coexists with the liver, and textual descriptions tend to focus more on the larger scale of the liver itself, resulting in a decline in the effectiveness of pretraining.

Method	HepaticVessel			Colon			Spleen	
	1 %	10%	100%	1%	10%	100%	10%	100%
Random	49.84	51.53	64.56	50.29	50.8	50.6	73.85	84.92
BYOL Grill et al. [2020]	49.67	58.85	65.57	50.1	50.29	50.22	77.64	85.98
SimSiam Chen and He [2021]	50.07	52.34	63.78	50.27	51.18	53.8	81.93	83.49
BarlowTwins Zbontar et al. [2021]	51.08	59.21	64.77	50.62	51.46	51.61	86.43	87.91
PCRLv2 Zhou et al. [2023b]	50.12	58.82	64.97	50.08	51.43	53.13	84.32	85.12
GTGM	50.39	59.74	65.13	51.12	51.74	53.88	86.95	89.64

Table 2: Mean Dice scores (%) of MRI image segmentation results. Red and blue entries denote the best and second-best results, respectively.

Method	BrainTumour			Heart		Hippocampus			Prostate	
	1%	10%	100%	10%	100%	1%	10%	100%	10%	100%
Random	30.33	32.37	40.45	82.37	94.82	48.32	78.46	84.18	33.53	39.19
BYOL Grill et al. [2020]	31.89	31.95	44.94	83.31	94.97	50.96	79.82	84.47	40.73	43.72
SimSiam Chen and He [2021]	31.27	32.67	37.45	86.06	93.51	52.24	78.99	83.35	42.48	41.64
BarlowTwins Zbontar et al. [2021]	32.13	32.21	45.83	83.37	94.68	50.84	78.75	83.96	33.37	40.15
PCRLv2 Zhou et al. [2023b]	32.83	34.87	43.14	85.89	90.77	52.29	77.38	81.24	32.73	40.54
GTGM	33.19	34.12	45.23	86.33	94.71	53.41	79.01	84.92	40.93	44.24

Experimental Results for MRI Segmentation. In comparison to CT imaging, MRI imaging typically involves four dimensions and exhibits lower imaging resolution, as well as more artifacts and noise in the images. Consequently, lesion segmentation in MRI images presents greater challenges. In our pretraining approach, tailored to 3D imaging, we extract the last three dimensions from the MRI images as input to the network. Despite these challenges, our approach has delivered promising experimental results, as depicted in Table 2. Consistently, our approach achieves either optimal or near-optimal solutions on the MRI dataset. We observe that numerous image-based self-supervised approaches yield degraded outcomes (with Dice scores lower than random initialization) due to variations in image dimensions within the MRI dataset. However, our approach exhibits robustness and effectively mitigates the degradation of the pretrained network by incorporating text as guidance.

Experimental Results for EM Neuron Segmentation. Electron Microscopy (EM) is an imaging technique with a resolution approximately a thousand times greater than CT and MRI, permitting the examination of structures at the nano and sub-nanometer levels. The ultra-high resolution makes neuron segmentation tasks in EM particularly challenging due to the densely packed structures. The typical methodology for EM neuron segmentation involves neural network-based affinity prediction, followed by post-processing with methods like WaterZ Funke et al. [2018] for instance segmentation. Due to the complexity of neurons, commonly used Vision-Language Pretraining (VLP) methods are not applicable. However, our proposed GTGM overcomes this limitation and demonstrates the effectiveness of generated text as a form of self-supervision training guidance. GTGM achieves state-of-the-art results on three neuron datasets in two settings. Please refer to Table 3 for specific numerical results.

Visual Results. GTGM demonstrates significant improvements in medical instance segmentation, particularly in terms of the connectivity of segmented surfaces, as shown in Figure 4. Our approach effectively segments liver tumors, and surface geometric structures of the left atrium, and exhibits the strongest integrity in neuronal segmentation.

5 Further Analysis

5.1 Ablation Study of Component Design

Table 4 presents the results of the ablation study for our proposed three components. We utilize 3D ResNet-50 He et al. [2016] as the vision backbone and conduct instance segmentation tests on three representative datasets (Liver, Prostate, CREMI C). As shown in Table 4, the impact of utilizing the synthetic text-guided VLP is clearly evident in the significant improvements observed in downstream

Table 3: Neuron segmentation results of three EM datasets. **Red** and **blue** entries denote the best and second-best results, respectively (The performance is better when the values of VOI and Arand are smaller).

Method	CREMI A 10		CREMI A 75		CREMI B 10		CREMI B 75		CREMI C 10		CREMI C 75	
	VOI	Arand										
Random	1.051	0.184	0.744	0.104	2.181	0.234	1.560	0.261	1.987	0.145	1.424	0.140
BYOL Grill et al. [2020]	0.961	0.206	0.764	0.119	1.581	0.155	1.441	0.142	1.672	0.196	1.326	0.124
SimSiam Chen and He [2021]	0.985	0.171	0.770	0.100	1.511	0.125	1.332	0.150	1.578	0.178	1.364	0.130
BarlowTwins Zbontar et al. [2021]	0.987	0.200	0.743	0.101	1.584	0.185	1.291	0.187	1.483	0.147	1.303	0.129
PCRLv2 Zhou et al. [2023b]	0.921	0.189	0.738	0.100	1.568	0.158	1.374	0.157	1.596	0.178	1.326	0.127
GTGM	0.902	0.166	0.728	0.092	1.525	0.117	1.279	0.106	1.422	0.137	1.280	0.118

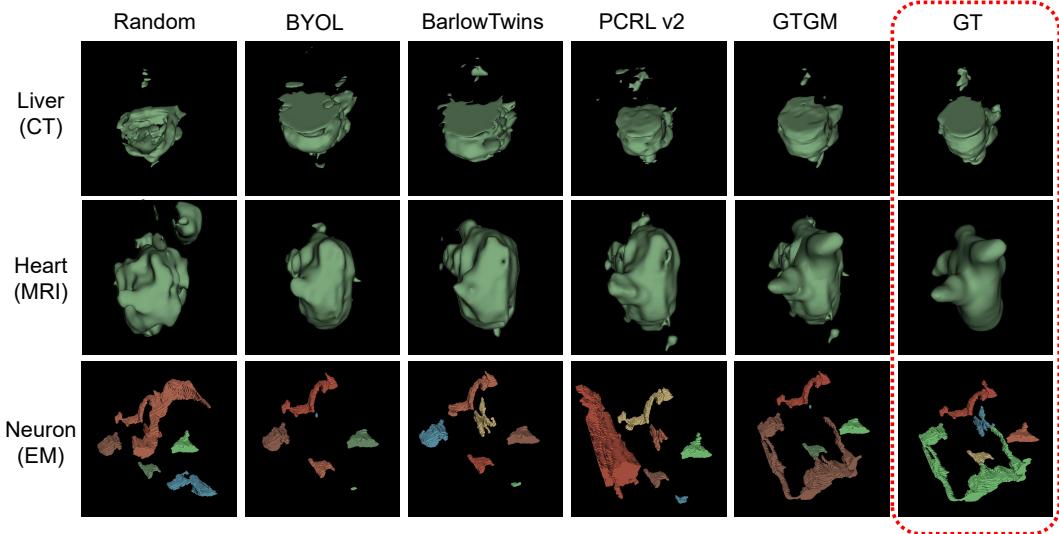


Figure 4: Visualization Results of 3D Instance Segmentation. ‘GT’ indicates the ground truth (Neurons are distinguished by their geometric shapes rather than color labels for the same instance).

tasks. The combination of all components of GTGM achieves the highest performance, clearly demonstrating the efficacy of GTGM. Learning only visual invariants or visual-textual invariants exhibits varying performance across different datasets. Notably, only learning visual invariants during pretraining proves to be more effective for large-scale and dense instance segmentation, while visual-textual invariants excel in guiding small and sparse segmentation. This difference in performance can be attributed to the fact that the former captures more intricate structural information, which is challenging to concisely describe through text alone.

Table 4: Ablation study of our framework is conducted on CT, MRI, and EM datasets, reporting Dice scores for CT and MRI, and VOI results for EM. **Red** and **blue** entries indicate the best and second-best results, respectively.

\mathcal{L}_{Cap}	\mathcal{L}_{VLP}	\mathcal{L}_{VR}	Liver (Dice \uparrow)			Prostate (Dice \uparrow)		CREMI C (VOI \downarrow)	
			1 %	10%	100%	10%	100%	10	75
		✓	50.11	55.91	64.87	33.37	40.15	1.483	1.303
	✓		46.84	51.95	61.21	33.67	39.21	1.871	1.413
✓	✓		51.89	57.63	64.39	38.91	41.39	1.497	1.333
✓	✓	✓	52.46	58.67	65.61	40.93	44.24	1.422	1.280

Table 5: Analysis of the trade-off between pre-training and finetuning.

Iters	Pancreas		
	1 %	10%	100%
3.5 k	38.93	56.7	64.73
7 k	39.26	56.09	64.97
15 k	39.67	56.83	64.95
50 k	40.13	56.96	65.61
Last	40.55	56.87	65.47

Table 6: Error bars of our methods across three modalities.

Dataset	10 %	100%
	Heart	86.33 ± 0.41
Metrics	Spleen	89.64 ± 0.43
	VOI	1.422 ± 0.031
Arand	0.137 ± 0.011	0.118 ± 0.008

5.2 Analysis of the Trade-off between Pretrain and Downstream tasks

The difference in objective functions between the pretraining and finetuning phases can lead to suboptimal performance in downstream tasks, despite the convergence of the loss function during pretraining. This highlights the existence of a trade-off between these two stages. To showcase this trade-off, we conduct experiments on the representative Pancreas dataset, and the results are presented in Table 5. Notably, the bold values in the table indicate the optimal segmentation results obtained under the current settings. These results signify that the number of iterations in our pretraining phase closely aligns with the performance achieved in the downstream task. This observation highlights the strong alignment between our objective function design and the requirements of the downstream task.

5.3 Analysis of Error Bars

Table 6 presents the error bars of our segmentation results on three modalities. We conduct three runs for each task and compute the mean and standard deviation of results. As observed from Table 6, our results demonstrate relatively minor variations, indicating the stability of GTGM’s performance in downstream tasks.

6 Conclusion

This work presents GTGM, a generative text-guided 3D vision-language pretraining framework. It accomplishes both instance-level visual-textual alignment and feature-level visual representation alignment using only 3D medical image inputs. GTGM delivers outstanding performance on 13 diverse medical datasets, tackling a variety of segmentation tasks with different data ratios. This demonstrates the efficiency and effectiveness of GTGM. Our work not only achieves the best performance but also opens up new opportunities to apply VLP to 3D medical images without relying on paired text. The broader impact and limitations are shown in the Appendix.

A Overview

In the supplementary material, we provide detailed explanations of our models and discuss the technical aspects involved. We have also included additional experimental results showcasing the performance of the Transformer backbone. Moreover, we have included numerous visualizations to enhance the understanding of our approach. To facilitate readability, we have also provided pseudo code for the core procedures.

B Latent Representation of Pretrained Model

We implement various pretraining approaches Zhou et al. [2023b], Zbontar et al. [2021], Grill et al. [2020], Chen and He [2021] and GTGM on diverse 3D medical image datasets with annotation. Then we utilize the pretrained visual encoder to extract the latent representation from three distinct medical image modalities, namely CT, MRI, and EM. Subsequently, we subject the representation to dimensionality reduction using the t-SNE algorithm Van der Maaten and Hinton [2008]. The resulting 2D representations obtained from different models are depicted in Figure 5.

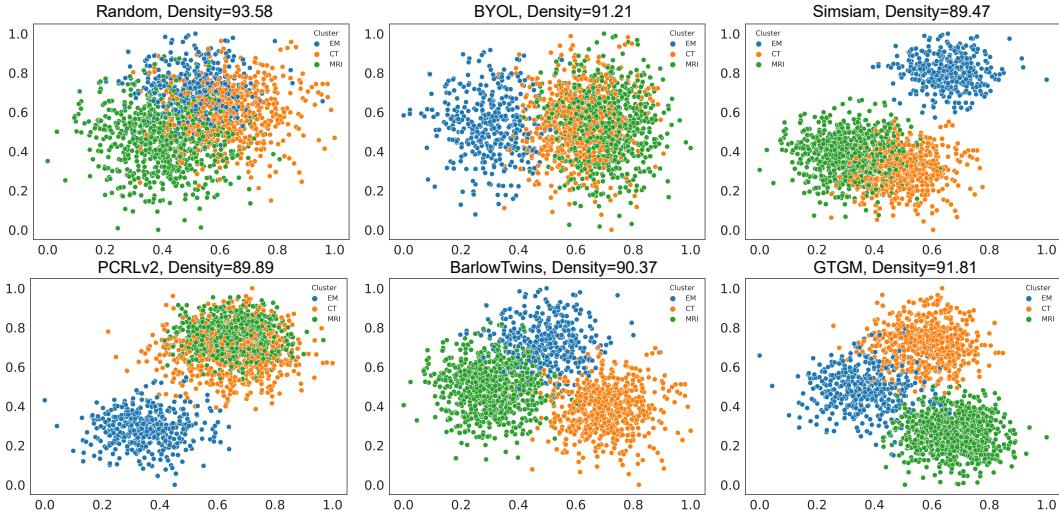


Figure 5: 2d Visualization of volume Representation. All visualizations are rendered using t-SNE.

Based on the visualization results obtained from t-SNE in Figure 5, it is evident that in scenarios with multiple modalities, the representation extracted through pretraining tend to exhibit confusion. Specifically, the CT and MRI modalities, compared to EM, display significantly lower resolution by several orders of magnitude. Although most pretraining models can effectively discriminate EM from the other two modalities, distinguishing between CT and MRI becomes challenging. This observation provides an explanation for the occurrence of model collapse phenomena in our experiments, wherein pretrained weights sometimes perform worse than random initialization.

Among all the models evaluated, only BarlowTwins and GTGM demonstrate effective differentiation among the three modalities. To further investigate this, we compute the density of the 2D representation, represented as the reciprocal of the k-nearest neighbor distances (higher values indicating tighter clustering). Interestingly, our model exhibits superior class separability compared to BarlowTwins, indicating that our text-guided approach captures informative features more effectively. In contrast, BarlowTwins' representation tend to be sparser, leading to potential overlap among modalities. This finding demonstrates the efficacy of our generative text-guided framework in capturing discriminated aspects among the three data modalities.

The additional analysis on density further supports the superiority of our approach in terms of class separability. The density measurements confirm that our model achieves tighter clustering, which is beneficial for class discrimination and avoiding confounding among the modalities. Overall, these findings highlight the effectiveness of our approach in capturing informative features and facilitating

discrimination among the three different data types, surpassing the performance of the Barlow Twins method.

C Discussion

C.1 Limitation

In the context of ablation experiments discussed in the main text, it was observed that for large-scale or higher-dimensional datasets, utilizing text generated through slicing for text-to-image pretraining yielded marginal information gain. Conversely, pretraining based on multi-view features derived from images proved more effective. In the future, it may be worthwhile to consider augmenting textual data appropriately or generating additional alternative texts through multiple slicing techniques, thereby harnessing the information contained within the text more comprehensively.

C.2 Broader Impact

The potential of vision-language multimodal pretraining has been widely recognized. Our proposed generation approach offers the possibility of joint training for datasets lacking textual descriptions. In addition to medical datasets, this approach can be extended to train on challenging datasets such as videos, point clouds, and light fields that are difficult to describe. Furthermore, directly fusing multimodal features at the downstream stage has been shown to significantly improve model performance. For instance, Liu et al. [2023] achieved remarkable performance gains by introducing simple text prompts at the downstream stage, although the use of generated text to directly guide downstream tasks has yet to be validated.

Moreover, our approach provides a means to leverage large language models (LLMs) effectively in computer vision tasks, leveraging existing pretrained weights. In the future, exploring the synchronization of LLM models with techniques like Stable Diffusion can be pursued to achieve zero-cost acquisition of high-quality data through text-guided image generation.

C.3 Future Work

Investigating these methodologies on varied medical data types, such as electrocardiograms linked with clinical monitoring records and multilingual reports, presents a fascinating future trajectory Li et al. [2023], Wan et al. [2023]. Moreover, the alignment of heterogeneous modality data can be perceived as a data fusion task, an issue frequently tackled in the field of physics Cheng et al. [2023, 2022], Liu et al. [2022a] or recommendation system Wan et al. [2022].

D Transformer-based Results

To validate the robustness and stability of our approach, we conducted experiments using a Transformer-based backbone. Taking inspiration from the network architecture of SwinUNETR Tang et al. [2022], we fine-tuned our model on an electron microscopy dataset. The experimental results, as shown in Table 7, demonstrate the effectiveness of our approach.

Based on our experimental results, it can be observed that the performance of using Swin Transformer as a backbone is slightly inferior to that of using ResNet50 as a backbone. Additionally, the segmentation results of Swin Transformer are worse when dealing with a small amount of data. However, in comparison to the ResNet backbone, the gains obtained from pretraining for downstream tasks are more significant. Therefore, when larger datasets are available, Swin Transformer as a backbone holds tremendous potential.

E Visualization

E.1 Visualization of Segmentation Results

We present the visualization results of the segmentation task on the MSD dataset Antonelli et al. [2022], as shown in Figure 6, 7. Our approach demonstrates superior capability in capturing detailed

Table 7: Experimental Results of Swin Transformer as a Backbone for Electron Microscope Neuron Segmentation. **Red** and **blue** entries denote the best and second-best results, respectively. Among these models, SwinUNETR refers to the implementation of the original pretraining approach described in the respective paper, with results reproduced on our pretraining dataset.

Method	CREMI A 10		CREMI A 75		CREMI B 10		CREMI B 75		CREMI C 10		CREMI C 75	
	VOI	Arand										
Random	3.720	0.898	0.967	0.244	4.495	0.646	1.988	0.289	5.082	0.770	1.537	0.175
BYOL Grill et al. [2020]	1.943	0.559	0.894	0.220	4.400	0.670	1.902	0.256	2.570	0.448	1.539	0.210
SimSiam Chen and He [2021]	1.832	0.531	0.927	0.192	3.381	0.442	1.871	0.282	2.419	0.428	1.442	0.166
BarlowTwins Zbontar et al. [2021]	2.104	0.615	0.956	0.194	3.292	0.590	1.851	0.201	2.419	0.434	1.437	0.162
PCRLv2 Zhou et al. [2023b]	2.094	0.640	0.881	0.184	3.174	0.442	1.594	0.241	2.219	0.298	1.474	0.164
SwinUNETR Tang et al. [2022]	1.913	0.579	0.855	0.177	3.813	0.617	1.859	0.208	2.419	0.434	1.423	0.160
GTM	1.693	0.529	0.832	0.180	2.949	0.419	1.423	0.160	2.188	0.304	1.393	0.151

anatomical structures (tumors) compared to other pretrained methods. Specifically, our method exhibits noticeable qualitative improvement, particularly for challenging instance segmentation tasks involving intricate structures such as Hepatic Vessels. The enhanced visual results highlight the efficacy of our approach in accurately delineating the morphological characteristics of organs (tumors) within the medical imaging context.

E.2 Caption Generate

We present the generated medical text descriptions in Figure 8. The descriptions generated by the untuned large-scale language models (e.g., BLIP Li et al. [2022a]) exhibit limited information, often resembling natural image descriptions, and contain numerous errors. For example, despite correctly identifying the CT image, there are instances where lung slices are incorrectly labeled as brain slices, introducing misleading information.

However, with our proposed fine-tuning approach, the generated text contains a substantial amount of useful information. Furthermore, utilizing our introduced filter, redundant and repetitive information in the text is eliminated (as indicated by the red text in the figure), thereby enhancing the information density and descriptive accuracy of the generated text.

F Pseudo Code

The core code for our pretraining process is outlined in Algorithm F.

Algorithm 1 PyTorch pseudo code of GTGM

```

# img1, img2, caption: Two data augmentations for a volume, \
#                               and the generated description
while niter < self.max_iterations:
    epoch_loss = 0
    epoch_loss_BT = 0
    epoch_loss_clip_diag = 0
    # get raw volume and generate descriptions
    img1, img2, caption = train_provider.next()

    # get image
    img1 = img1.to(torch.float32).to(self.device).contiguous()
    img2 = img2.to(torch.float32).to(self.device).contiguous()

    self.optimizer.zero_grad()

    # amp style (might decrease precision)
    with autocast():
        imp_tokenize_output = self.model.module._tokenize(caption)
        input_ids = imp_tokenize_output.input_ids.to(
            self.device).contiguous()
        attention_mask = imp_tokenize_output.attention_mask.to(
            self.device).contiguous()

        output_dict = self.model(img1, img2, input_ids, attention_mask)
        img_emb1, img_emb2 = output_dict['img_emb1'],
                             output_dict['img_emb2']

        proj_img_emb1, proj_img_emb2 = output_dict['proj_img_emb1'],
                                       output_dict['proj_img_emb2']

        proj_text_emb = output_dict['proj_text_emb']

        loss_clip_diag1, acc1_1 = self.clip_loss(x=proj_img_emb1, y=proj_text_emb)
        loss_clip_diag2, acc1_2 = self.clip_loss(x=proj_img_emb2, y=proj_text_emb)

        acc1 = (acc1_1 + acc1_2) / 2

        cov_loss = self.covar_loss(img_emb1, img_emb2) * 0.01

        loss = loss_clip_diag1 + loss_clip_diag2 + cov_loss
        # accumulate loss for logging
        epoch_loss += loss.item()
        epoch_loss_clip_diag += loss_clip_diag1.item() + loss_clip_diag2.item()
        epoch_loss_BT += cov_loss.item()
        scaler.scale(loss).backward()
        scaler.step(self.optimizer)
        scaler.update()

```

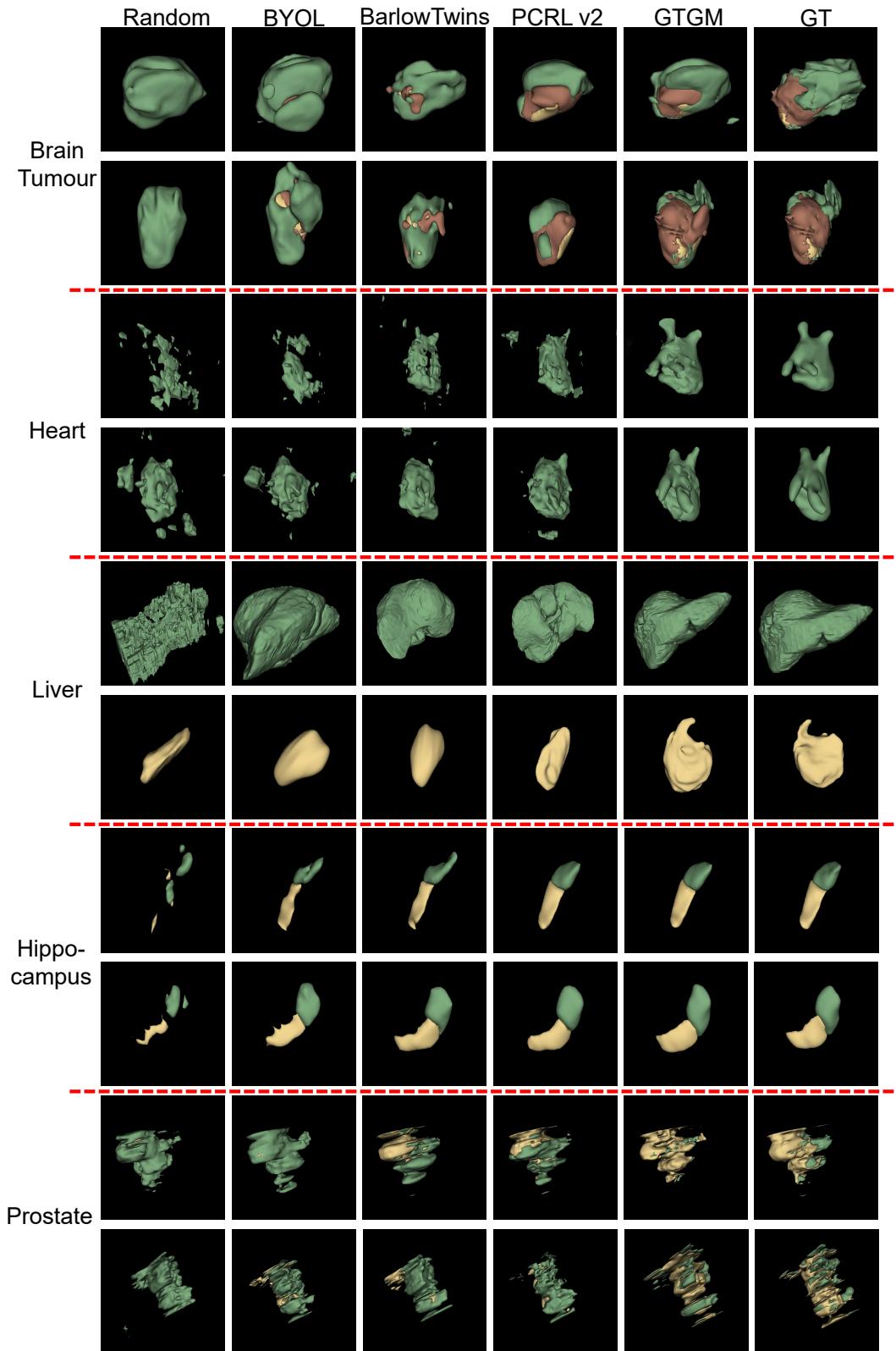


Figure 6: Visualization results of the first 5 tasks of MSD. ‘GT’ indicates ground truth.

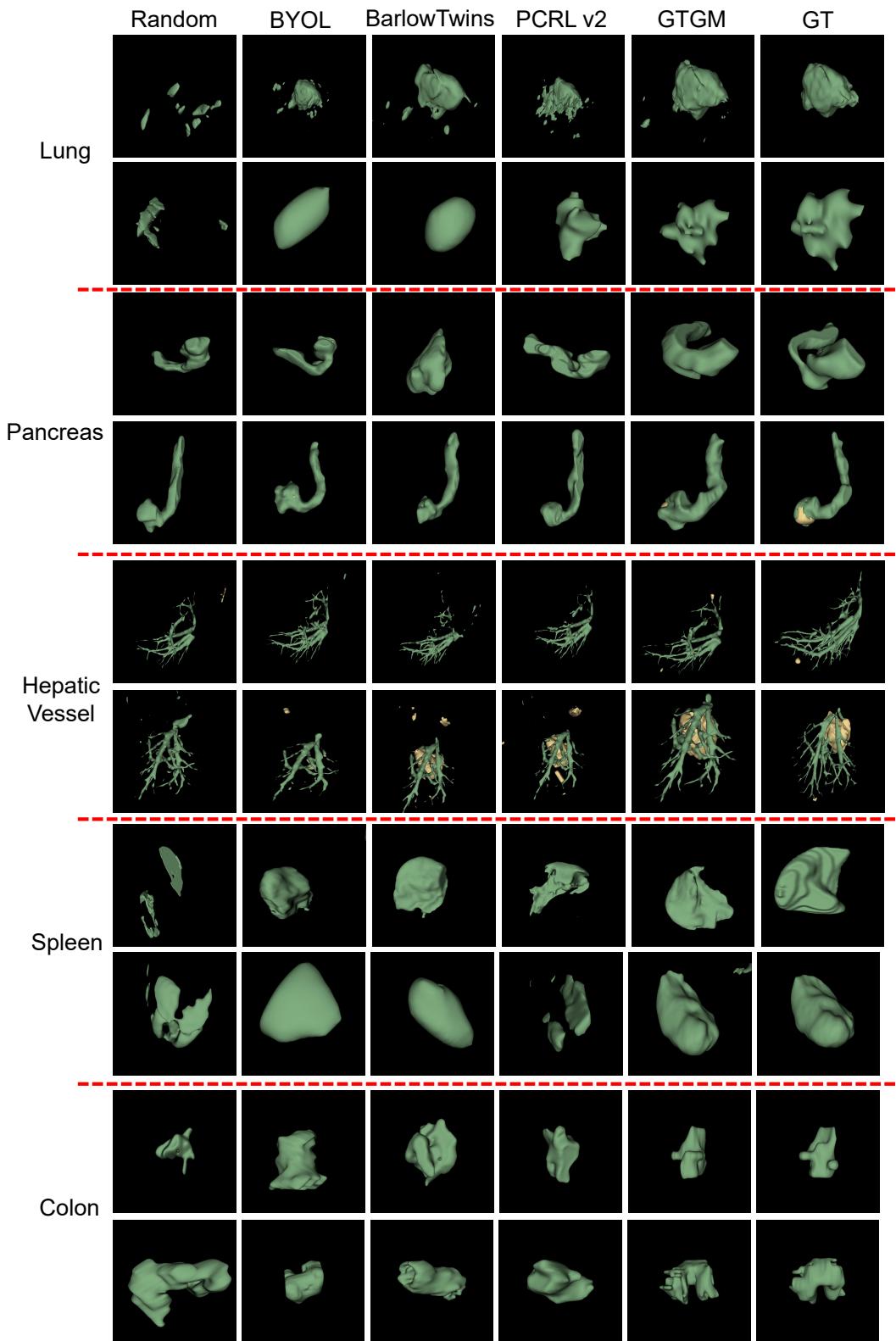
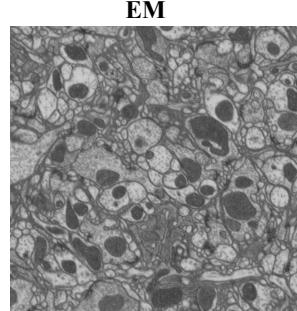


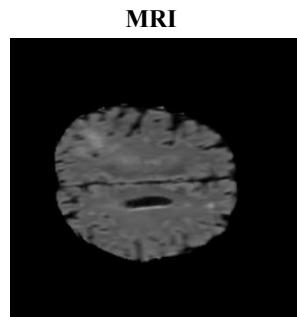
Figure 7: Visualization results of the last 5 tasks of MSD



Without finetuning:
a black and white image of a cell

After finetuning
fig. 2. electron micrograph of a transverse section of a human cerebellum. the cell bodies are arranged in a single layer of filaments. **the cell bodies are arranged in a single layer of filaments.** the cell

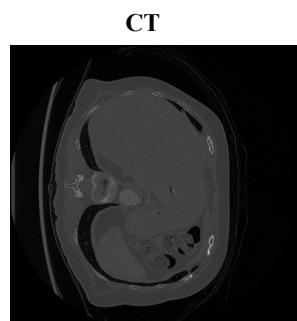
After filter:
electron micrograph of a transverse section of a human cerebellum. the cell bodies are arranged in a single layer of filaments.



Without finetuning:
the image shows a small hole in the surface of the moon

After finetuning
fig. 1. axial diffusion - weighted image of the brain of a patient with a left parietal lobe lesion. the lesion is **hyperintense** on the diffusion - **weighted image**.

After filter:
axial diffusion - weighted image of the brain of a patient with a left parietal lobe lesion. the lesion is hyperintense on the diffusion



Without finetuning:
ct scan of the brain

After finetuning
figure 1. ct scan of the chest showing a large mass in the right lung.

After filter:
ct scan of the chest showing a large mass in the right lung.

Figure 8: Generated medical description

References

- Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022.
- Michela Antonelli, Annika Reinke, Spyridon Bakas, Keyvan Farahani, Annette Kopp-Schneider, Bennett A Landman, Geert Litjens, Bjoern Menze, Olaf Ronneberger, Ronald M Summers, et al. The medical segmentation decathlon. *Nature communications*, 13(1):4128, 2022.
- Ignacio Arganda-Carreras, Srinivas C Turaga, Daniel R Berger, Dan Cireşan, Alessandro Giusti, Luca M Gambardella, Jürgen Schmidhuber, Dmitry Laptev, Sarvesh Dwivedi, Joachim M Buhmann, et al. Crowdsourcing the creation of image segmentation algorithms for connectomics. *Frontiers in neuroanatomy*, page 142, 2015.
- Adrien Bardes, Jean Ponce, and Yann Lecun. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. In *ICLR 2022-10th International Conference on Learning Representations*, 2022.
- Benedikt Boecking, Naoto Usuyama, Shruthi Bannur, Daniel C Castro, Anton Schwaighofer, Stephanie Hyland, Maria Wetscherek, Tristan Naumann, Aditya Nori, Javier Alvarez-Valle, et al. Making the most of text semantics to improve biomedical vision-language processing. *arXiv preprint arXiv:2204.09817*, 2022.
- Victor Ion Butoi, Jose Javier Gonzalez Ortiz, Tianyu Ma, Mert R Sabuncu, John Guttag, and Adrian V Dalca. Universeg: Universal medical image segmentation. *arXiv preprint arXiv:2304.06131*, 2023.
- Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.
- Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15750–15758, 2021.
- Sibo Cheng, I Colin Prentice, Yuhan Huang, Yufang Jin, Yi-Ke Guo, and Rossella Arcucci. Data-driven surrogate model with latent data assimilation: Application to wildfire forecasting. *Journal of Computational Physics*, 464:111302, 2022.
- Sibo Cheng, César Quilodrán-Casas, Said Ouala, Alban Farchi, Che Liu, Pierre Tandeo, Ronan Fablet, Didier Lucor, Bertrand Iooss, Julien Brajard, et al. Machine learning with data assimilation and uncertainty quantification for dynamical systems: a review. *arXiv preprint arXiv:2303.10462*, 2023.
- Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by context prediction. In *Proceedings of the IEEE international conference on computer vision*, pages 1422–1430, 2015.
- J Funke, S Saalfeld, DD Bock, SC Turaga, and E Perlman. Miccai challenge on circuit reconstruction from electron microscopy images, 2016.
- Jan Funke, Fabian Tschopp, William Grisaitis, Arlo Sheridan, Chandan Singh, Stephan Saalfeld, and Srinivas C Turaga. Large scale image segmentation with structured loss based deep learning for connectome reconstruction. *IEEE transactions on pattern analysis and machine intelligence*, 41(7):1669–1680, 2018.
- Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. *arXiv preprint arXiv:1803.07728*, 2018.
- Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020.
- Shih-Cheng Huang, Liyue Shen, Matthew P Lungren, and Serena Yeung. Gloria: A multimodal global-local representation learning framework for label-efficient medical image recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3942–3951, 2021.

- Wei Huang, Chang Chen, Zhiwei Xiong, Yueyi Zhang, Xuejin Chen, Xiaoyan Sun, and Feng Wu. Semi-supervised neuron segmentation via reinforced consistency learning. *IEEE Transactions on Medical Imaging*, 41(11):3016–3028, 2022.
- Narayanan Kasthuri, Kenneth Jeffrey Hayworth, Daniel Raimund Berger, Richard Lee Schalek, José Angel Conchello, Seymour Knowles-Barley, Dongil Lee, Amelio Vázquez-Reina, Verena Kaynig, Thouis Raymond Jones, et al. Saturated reconstruction of a volume of neocortex. *Cell*, 162(3):648–661, 2015.
- Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020.
- Jun Li, Che Liu, Sibo Cheng, Rossella Arcucci, and Shenda Hong. Frozen language model helps ecg zero-shot learning. In *Medical Imaging with Deep Learning*, 2023.
- Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022a.
- Yanghao Li, Haoqi Fan, Ronghang Hu, Christoph Feichtenhofer, and Kaiming He. Scaling language-image pre-training via masking. *arXiv preprint arXiv:2212.00794*, 2022b.
- C Liu, R Fu, D Xiao, R Stefanescu, P Sharma, C Zhu, S Sun, and C Wang. Enkf data-driven reduced order assimilation system. *Engineering Analysis with Boundary Elements*, 139:46–55, 2022a.
- Jie Liu, Yixiao Zhang, Jie-Neng Chen, Junfei Xiao, Yongyi Lu, Bennett A Landman, Yixuan Yuan, Alan Yuille, Yucheng Tang, and Zongwei Zhou. Clip-driven universal model for organ segmentation and tumor detection. *arXiv preprint arXiv:2301.00785*, 2023.
- Xiaoyu Liu, Wei Huang, Yueyi Zhang, and Zhiwei Xiong. Biological instance segmentation with a superpixel-guided graph. In *International Joint Conference on Artificial Intelligence. IJCAI*, 2022b.
- Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.
- Ishan Misra and Laurens van der Maaten. Self-supervised learning of pretext-invariant representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6707–6717, 2020.
- Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European conference on computer vision*, pages 69–84. Springer, 2016.
- Juan Nunez-Iglesias, Ryan Kennedy, Toufiq Parag, Jianbo Shi, and Dmitri B Chklovskii. Machine learning of hierarchical clustering to segment 2d and 3d images. *PloS one*, 8(8):e71715, 2013.
- Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.
- Philipp Schlegel, Alexander S Bates, Tejal Parag, Gregory SXE Jefferis, and Davi D Bock. Automatic detection of synaptic partners in a whole-brain drosophila em dataset. *Nature Methods*, 18(8):877–884, 2021.
- Sanjay Subramanian, Lucy Lu Wang, Sachin Mehta, Ben Bogin, Madeleine van Zuylen, Sravanthi Parasa, Sameer Singh, Matt Gardner, and Hannaneh Hajishirzi. Medicat: A dataset of medical images, captions, and textual references. *arXiv preprint arXiv:2010.06000*, 2020.
- Shin-ya Takemura, Yoshinori Aso, Toshihide Hige, Allan Wong, Zhiyuan Lu, C Shan Xu, Patricia K Rivlin, Harald Hess, Ting Zhao, Toufiq Parag, et al. A connectome of a learning and memory center in the adult drosophila brain. *Elife*, 6:e26975, 2017.
- Yucheng Tang, Dong Yang, Wenqi Li, Holger R Roth, Bennett Landman, Daguang Xu, Vishwesh Nath, and Ali Hatamizadeh. Self-supervised pre-training of swin transformers for 3d medical image analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20730–20740, 2022.

Ekin Tiu, Ellie Talius, Pujan Patel, Curtis P Langlotz, Andrew Y Ng, and Pranav Rajpurkar. Expert-level detection of pathologies from unannotated chest x-ray images via self-supervised learning. *Nature Biomedical Engineering*, pages 1–8, 2022.

Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

Zhongwei Wan, Benyou Wang, Xin Liu, Jiezhong Qiu, Boyu Li, Ting Guo, Guangyong Chen, and Yang Wang. Spatio-temporal contrastive learning enhanced gnns for session-based recommendation. *ArXiv*, abs/2209.11461, 2022.

Zhongwei Wan, Che Liu, Mi Zhang, Jie Fu, Benyou Wang, Sibo Cheng, Lei Ma, César Quilodrán-Casas, and Rossella Arcucci. Med-unic: Unifying cross-lingual medical vision-language pre-training by diminishing bias. *arXiv preprint arXiv:2305.19894*, 2023.

Fuying Wang, Yuyin Zhou, Shujun Wang, Varut Vardhanabuti, and Lequan Yu. Multi-granularity cross-modal alignment for generalized medical visual representation learning. In *Advances in Neural Information Processing Systems*, 2022.

Donglai Wei, Zudi Lin, Daniel Franco-Barranco, Nils Wendt, Xingyu Liu, Wenjie Yin, Xin Huang, Aarush Gupta, Won-Dong Jang, Xueying Wang, et al. Mitoem dataset: Large-scale 3d mitochondria instance segmentation from em images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 66–76. Springer, 2020.

Hongwei Xue, Peng Gao, Hongyang Li, Yu Qiao, Hao Sun, Houqiang Li, and Jiebo Luo. Stare at what you see: Masked image modeling without reconstruction. *arXiv preprint arXiv:2211.08887*, 2022.

Yiwen Ye, Yutong Xie, Jianpeng Zhang, Ziyang Chen, and Yong Xia. Uniseg: A prompt-driven universal segmentation model as well as a strong representation learner. *arXiv preprint arXiv:2304.03493*, 2023.

Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *International Conference on Machine Learning*, pages 12310–12320. PMLR, 2021.

Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *European conference on computer vision*, pages 649–666. Springer, 2016.

Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D Manning, and Curtis P Langlotz. Contrastive learning of medical visual representations from paired images and text. *arXiv preprint arXiv:2010.00747*, 2020.

Hong-Yu Zhou, Chenyu Lian, Liansheng Wang, and Yizhou Yu. Advancing radiograph representation learning with masked record modeling. *arXiv preprint arXiv:2301.13155*, 2023a.

Hong-Yu Zhou, Chixiang Lu, Chaoqi Chen, Sibei Yang, and Yizhou Yu. A unified visual information preservation framework for self-supervised pre-training in medical image analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023b.