MOSMOS: Multi-organ segmentation facilitated by medical report supervision

Weiwei Tian^a, Xinyu Huang^b, Junlin Hou^{b,e}, Caiyue Ren^b, Longquan Jiang^{b,*}, Rui-Wei Zhao^a, Gang Jin^d, Yuejie Zhang^{b,c} and Daoying Geng^a

ARTICLE INFO

Keywords:
Medical report supervision
Multi-label recognition
Multi-organ segmentation
Vision-language pre-training
Visual representation learning

ABSTRACT

Owing to a large amount of multi-modal data in modern medical systems, such as medical images and reports, Medical Vision-Language Pre-training (Med-VLP) has demonstrated incredible achievements in coarse-grained downstream tasks (i.e., medical classification, retrieval, and visual question answering). However, the problem of transferring knowledge learned from Med-VLP to fine-grained multi-organ segmentation tasks has barely been investigated. Multi-organ segmentation is challenging mainly due to the lack of large-scale fully annotated datasets and the wide variation in the shape and size of the same organ between individuals with different diseases. In this paper, we propose a novel pre-training & fine-tuning framework for Multi-Organ Segmentation by harnessing Medical repOrt Supervision (MOSMOS). Specifically, we first introduce global contrastive learning to maximally align the medical image-report pairs in the pre-training stage. To remedy the granularity discrepancy, we further leverage multi-label recognition to implicitly learn the semantic correspondence between image pixels and organ tags. More importantly, our pre-trained models can be transferred to any segmentation model by introducing the pixel-tag attention maps. Different network settings, i.e., 2D U-Net and 3D UNETR, are utilized to validate the generalization. We have extensively evaluated our approach using different diseases and modalities on BTCV, AMOS, MMWHS, and BRATS datasets. Experimental results in various settings demonstrate the effectiveness of our framework. This framework can serve as the foundation to facilitate future research on automatic annotation tasks under the supervision of medical reports.

1. Introduction

Assigning an organ tag to each pixel in a medical image, also known as multi-organ segmentation, is a crucial task in medical image analysis, as it contributes to various computer-aided diagnosis and treatment tasks, including volume measurement [42], 3D reconstruction [51], and treatment planning [29]. To achieve these clinical applications, it is necessary to segment multiple organs in medical images accurately and robustly. However, compared to one particular organ segmentation, manually annotating multiple organs by radiologists is not only time-consuming and laborious but also heavily dependent on their experience. Since automatic multi-organ segmentation is efficient, it becomes an essential issue to address the growing clinical needs [49].

With the development of Fully Convolutional Networks (FCN) [32] and Vision Transformers (ViT) [10], impressive segmentation performance has been achieved. However, existing works on multi-organ segmentation are usually based on the supervised learning paradigm [57, 56, 2, 14], which is dramatically limited by high-quality and

wwtian20@fudan.edu.cn (Tian); xinyuhuang20@fudan.edu.cn (Huang); csejlhou@ust.hk (Hou); rencaiyue@163.com (Ren); lqjiang@fudan.edu.cn (Jiang); rwzhao@fudan.edu.cn (Zhao); jingang@smmu.edu.cn (Jin); yjzhang@fudan.edu.cn (Zhang); gengdy@163.com (Geng)

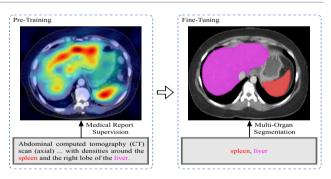


Figure 1: An example of our multi-organ segmentation result when using medical report supervision for pre-training. Left: The attention map locating the organ tags in the radiology image extracted from the corresponding medical report. Right: Our segmentation result for the corresponding organ tags.

high-cost annotations. To tackle this issue, pre-training on large-scale datasets and then fine-tuning on smaller target datasets has become a widely adopted mode. For instance, Swin UNETR [46] leveraged self-supervised pre-training with tailored proxy tasks to alleviate the lack of annotations. Nevertheless, it only learns transferable visual representations from five Computer Tomography (CT) datasets. It is unsuitable for segmentation tasks in other diseases or modalities, such as Magnetic Resonance Imaging (MRI).

^aAcademy for Engineering and Technology, Fudan University, Shanghai, 200433, China

^bSchool of Computer Science, Shanghai Key Laboratory of Intelligent Information Processing, Fudan University, Shanghai, 200433, China

^cShanghai Collaborative Innovation Center of Intelligent Visual Computing, China

^dDepartment of Hepatobiliary Pancreatic Surgery, Changhai Hospital, Second Military Medical University (Naval Medical University), Shanghai, 200433, China

^eDepartment of Computer Science and Engineering, The Hong Kong University of Science and Technology, China

^{*}Corresponding author

In summary, critical difficulties exist in two aspects with multi-organ segmentation: (i) There is a lack of large-scale fully annotated, multi-disease, or multi-modal datasets. (ii) The shape and size of the same organ vary significantly between patients with different diseases, making it difficult for the network to learn representative features. To address the first limitation, we argue that medical reports reflect radiologists' perceptions of multi-disease and multi-modal medical images, which can serve as weakly supervised information to help optimize the multi-organ segmentation network even with fewer annotations (see Fig. 1). Moreover, taking into consideration that radiologists prepare medical reports accompanied with radiology images as part of their daily routine, large-scale medical image-report pairs are easy to access without extra cost, in contrast to pixel-level finegrained annotations. To address the second challenge, we simultaneously introduce global image-report aligning and local pixel-tag aligning to identify discriminative representations for the same organ with different diseases in the pretraining stage. Furthermore, we design pixel-tag attention maps to assist multi-organ segmentation tasks in the finetuning stage.

Concretely, we propose a novel pre-training & finetuning framework named MOSMOS for multi-organ segmentation based on medical report supervision. In the pretraining phase, image-report contrastive learning is used to align the global features of medical images and corresponding reports. In addition, we apply a more fine-grained pretraining task called multi-label recognition. It can locate image regions with the organ tags in the corresponding reports, which has the following advantages: (i) The tags are the organ classification labels extracted from the reports without additional manual annotations. (ii) The tags are encoded into query embeddings and then fed into a Transformer decoder [47, 30, 58] to perform multi-modal interaction, which guarantees the generalizability of the pretrained model transferred to multi-disease, multi-modal, and multi-organ segmentation tasks. (iii) By implicitly optimizing the attention maps in the Transformer decoder, the organ tags can be associated with fine-grained and interpretable location information. They are capable of assisting multiorgan segmentation tasks to be better optimized since attention maps can also be regarded as segmentation results with low resolution. In the fine-tuning phase, we combine the segmentation loss and the pixel-tag aligning loss to supervise the training process.

Our pre-trained model can be fine-tuned on any down-stream segmentation framework to boost performance. A series of comprehensive experiments have proved the effectiveness of our method. In the aspect of the downstream segmentation frameworks, we verify on the representative segmentation models (U-Net [40] & UNETR [14]) with two mainstream visual backbones (ResNet [16] & ViT [10]), respectively. As for the downstream segmentation datasets, we evaluate on four publicly available multi-disease and multiorgan datasets (BTCV [26] & AMOS [22] & MMWHS [64] & BRATS [44]) with different modalities (CT & MRI).

The main contributions of this work are summarized as follows:

- We establish MOSMOS, a novel pre-training & finetuning framework to fully leverage the intrinsic medical report supervision within the paired images and reports to learn medical visual representation instead of purely exploiting radiology images. To the best of our knowledge, this is the first work that the medical vision-language pre-training is applied to downstream tasks of multi-organ segmentation.
- We design global image-report aligning and local pixel-tag aligning in the pre-training stage, which is more suitable for fine-grained segmentation tasks in the downstream.
- We verify the effectiveness of the proposed method on the representative segmentation frameworks and four widely used multi-disease and multi-organ datasets of different modalities with 2D and 3D medical images. Our proposed MOSMOS significantly improves the multi-organ segmentation performance by a substantial margin.

2. Related work

Before introducing the proposed method, we mainly review previous works that inspired the design of our multi-organ segmentation scheme in this section. The two essential parts are (i) multi-organ segmentation; (ii) language supervision, in order to leverage cross-modal information to guide the multi-organ segmentation.

2.1. Multi-organ segmentation

Many attempts have been made to implement multiorgan segmentation more efficiently. According to the backbone, these approaches can be divided into three categories. (i) FCN-based: To leverage the partially labeled datasets, the multi-head strategy [5, 13, 43] was used for segmentation, which consists of a task-shared encoder and multiple decoders (layers) with specific tasks, leading to poor scalability. To improve the flexibility, DoDNet [56] built a dynamic on-demand framework that introduced a dynamic segmentation head to the shared encoder-decoder structure. (ii) ViT-based: Swin-Unet [2] first utilized hierarchical Swin Transformer [31] with shifted window operation to capture global and long-term semantic information. (iii) FCN and ViT combined: Taking advantage of the locality of convolution and the globality of self-attention in Transformer, recent works [3, 53, 14, 61] adopted the hybrid architecture. Based on U-Net [40] architecture, TransUNet [3] and TransDoDNet [53] introduced Transformer as a bottleneck feature extractor for modeling long-range organ-wise dependencies, which is conducive to multi-organ segmentation. UNETR [14] used Transformer as the encoder and delivered the encoded representations to the FCN-based decoder by skip connections. NnFormer [61] applied interleaved convolutional layers and Transformer blocks to play both advantages sufficiently. However, the performance of these supervised learning methods learning from scratch is limited by the quantity and quality of annotations, or that transferring pre-trained weights from ImageNet [9] is suboptimal due to the drastic difference between natural and medical images. Performance improvements have been achieved through supervised learning methods that transferred pre-trained weights from large-scale, partially labeled medical datasets. Nonetheless, these methods also need intensive labor and expertise costs.

Recent advances in self-supervised pre-training [4, 45, 63, 46, 52] provided the promise of leveraging unlabeled medical images. Specifically, Swin UNETR [46] first designed three tailored proxy tasks, that is, masked volume inpainting, rotation prediction, and contrastive learning, to pre-train the Swin Transformer encoder. The pre-trained encoder was transferred to downstream segmentation tasks and achieved observable improvements. Despite its success, a gap exists between the upstream self-supervised task and the downstream segmentation tasks. Consequently, ReFs [52] proposed an extra supervised reference task as a bridge to minimize the gap. Unlike these approaches, our pretraining framework introduces the cross-modal supervisory information in paired medical images and reports at no extra cost to facilitate multi-disease, multi-modal, and multiorgan segmentation tasks. Meanwhile, we employ multilabel recognition to align image pixels with organ tags automatically extracted from medical reports, bridging the gap between upstream and downstream tasks.

2.2. Language supervision

Towards the goal of utilizing unlabeled images more efficiently, several follow-ups [23, 55, 39, 20, 54, 21, 19] based on Contrastive Language-Image Pre-training (CLIP) [38] have achieved promising results in learning visual representation with language supervision using plenty of imagetext pairs in the general domain. Inspired by these pioneering works, [59, 12, 60, 41, 6, 50, 28] applied modified CLIP to medical classification, retrieval, and visual question answering tasks. For more fine-grained dense prediction tasks, LViT [27] introduced medical text annotations to lead the generation of pseudo labels in semi-supervised learning. In addition to using global contrastive learning to align medical images and reports, GLoRIA [18] and LoVT [35] proposed utilizing local contrastive learning to align image sub-regions and words or sentences in the paired reports, and BioViL [1] adopted masked language modeling to leverage text semantics sufficiently. Furthermore, MGCA [48] explored the abundant semantic correspondences between radiology images and reports with multiple granularities: disease-level, instance-level, and token-level. Despite achieving exceptional performance, these segmentation or detection approaches that utilize language supervision are confined to the localization of pulmonary lesions or cell nuclei in 2D images. Augmenting the above-mentioned medical segmentation methods, we extend to broader multiorgan segmentation scenarios of different modalities with



Figure 2: Illustration of the 20 organ categories in the tag list. The tag size is proportional to the tag frequency in the training set of the ROCO dataset.

2D and 3D medical images by introducing global imagereport aligning and local pixel-tag aligning using multi-label recognition in the pre-training stage.

3. Material and methods

In this section, we first present the datasets and the overview of our MOSMOS framework. Next, we introduce the pre-training method of MOSMOS, including global image-report aligning and local pixel-tag aligning. Then we illustrate the fine-tuning approach, which utilizes weakly supervised positioning to facilitate multi-organ segmentation.

3.1. Datasets

3.1.1. Dataset for pre-training

The Radiology Objects in COntext (ROCO) dataset [37] contains over 81,000 2D radiology images, split into 73,594 and 8,176 images for training and validation sets, respectively. ROCO does not concentrate on a specific disease or anatomical structure but addresses multi-modal radiology images, including Angiography, CT, Fluoroscopy, MRI, Mammography, Positron Emission Tomography (PET), PET-CT, Ultrasound, and X-Ray.

All images in ROCO have corresponding medical reports and a set of organ tags obtained from the reports. Each report describes the visual element in its semantic context. To acquire the organ tags for radiology images, we first define K=20 common organ categories, including abbreviations and synonyms. Then the list is double-checked by radiologists. After substituting abbreviations and synonyms with the unified forms, we extract the organ tags from the medical reports by matching. An aggregation step is further executed to merge the multiple mentioned tags in a report. A detailed overview of this tag list is shown in Fig. 2.

3.1.2. Datasets for fine-tuning

We extensively evaluate our multi-organ segmentation approach on two modalities of datasets from different human body regions, that is, BTCV [26] for abdominal multi-organ segmentation using CT, AMOS [22] for abdominal multi-organ segmentation using CT and MRI, MMWHS [64] for cardiac substructure segmentation using MRI, and

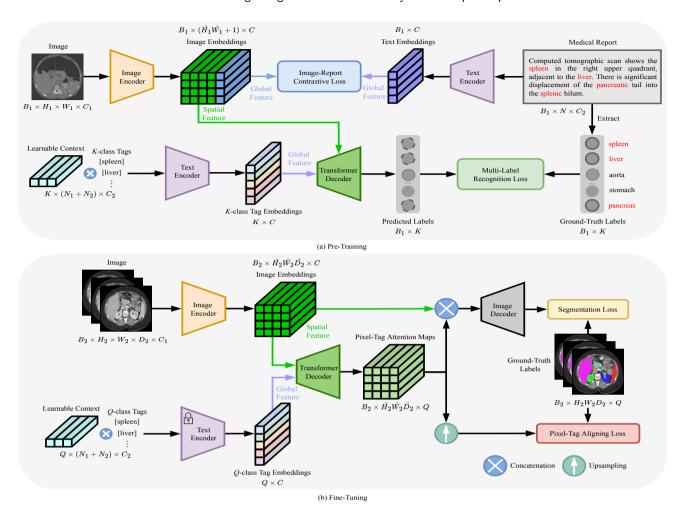


Figure 3: Illustration of our proposed MOSMOS framework in both pre-training and fine-tuning stages. (a) In the pre-training stage, MOSMOS applies image-report contrastive learning to align the global features of radiology images with those of corresponding medical reports. To further learn fine-grained visual representation from medical report supervision, the visual spatial features and the embeddings of the constructed K-class tags are sent to the Transformer decoder for multi-label recognition. Note that the ground-truth tags are extracted from the medical reports with no manual annotation. Note: B_1 : batch size, H_1 : height of the image, W_1 : width of the image, W_1 : width of the image embedding, W_1 : width of the image embedding, W_2 : width of the image embedding, W_3 : token length of the medical report, W_3 : token length of the learnable textual context, W_3 : token length of the tag, W_3 : dimension of the medical report and tag. (b) In the fine-tuning stage, the pixel-tag attention maps calculated by the Transformer decoder are fed into the image decoder. The segmentation loss and the pixel-tag aligning loss are combined to supervise the training process. Note that the learnable textual context is shared across all tags and is continuously updated in both stages. Note: W_3 : batch size, W_3 : height of the image, W_3 : width of the image embedding, W_3 : depth of the image embedding, W_3 : number of the organ tags. For 2D images, W_3 are omitted.

BRATS [44] for brain tumor segmentation using MRI. These datasets adopted for fine-tuning do not need to provide medical reports but human-assisted annotations. Following the split ratios of [11], the percentages for training, validation, and test sets on BTCV, MMWHS, and BRATS are 70%, 10%, and 20%, respectively. The AMOS dataset is divided into training and validation sets at a ratio of 2:1.

• **BTCV** provides annotations of Q=13 abdominal organs (that is, spleen, right kidney, left kidney, gallbladder, esophagus, liver, stomach, aorta, inferior vena cava, portal vein and splenic vein, pancreas, right adrenal gland, and left adrenal gland). There are 30

- abdomen CT scans from colorectal cancer or ventral hernia patients acquired during the portal venous contrast phase. All images are manually annotated and further verified by experienced radiologists from Vanderbilt University Medical Center.
- AMOS consists of 300 CT and 60 MRI scans, collected from multi-center, multi-vendor, multi-phase, multi-disease patients. It provides voxel-level annotations for Q=15 abdominal organs, namely spleen, right kidney, left kidney, gallbladder, esophagus, liver, stomach, aorta, inferior vena cava, pancreas, right adrenal gland, left adrenal gland, duodenum, bladder, and prostate or uterus. Notably, the duodenum,

bladder, and prostate or uterus are considered openset organ categories, for which tags are not extracted during the pre-training stage. Additionally, MRI scans lack annotations for the bladder and prostate or uterus organs.

- MMWHS is a dataset for whole heart segmentation of Q=7 cardiac substructures (that is, myocardium, left atrium, left ventricle, right atrium, right ventricle, ascending aorta, pulmonary artery), containing 20 cardiac MRI images from patients with cardiovascular diseases. These data are obtained using 3D balanced steady-state free precession (b-SSFP) sequences.
- **BRATS** is specifically designed for brain tumor segmentation, which comprises 484 multi-modal MRI scans (including FLAIR, T1w, T1gd, T2w modalities) from patients diagnosed with gliomas. All Q=3 segmentation targets (that is, tumor core, whole tumor, enhancing tumor) are categorized as open-set.

3.2. MOSMOS

As shown in Fig. 3, MOSMOS is a two-stage framework for multi-organ segmentation based on medical report supervision. Given a batch of image-report pairs in the first pre-training stage, we first split the visual representations into global and spatial features through the image encoder, and the textual representations into report-level and tag-level features through the shared text encoder. Then we perform two tasks: global image-report aligning and local pixeltag aligning. The first task adopts the contrastive learning strategy, which reinforces the matching degree between the visual global representations of the radiology images and the textual global representations of the corresponding reports. The second task leverages multi-label recognition to align the image regions and the organ tags in the original medical reports. For this purpose, we apply a Transformer decoder based on cross-attention to fully leverage visual spatial features to recognize tags located in the images. In the second fine-tuning stage, the pixel-tag attention maps generated by the Transformer decoder are concatenated with the visual spatial features and then fed into the image decoder concurrently. Besides the segmentation loss, the pixel-tag aligning loss is also applied to supervise the training process.

3.2.1. Pre-training

3.2.1.1. Global image-report aligning

In the routine clinical workflow, medical reports paired with radiology images are generated naturally by experienced radiologists. Assume each image-report pair is unique. We utilize global image-report contrastive learning to align image-report representations. For a mini-batch of B_1 image-report pairs (I,R) sampled from training dataset, we use (I_i,R_i) to represent the i-th pair. We embed the 2D image $I_i \in \mathbb{R}^{H_1 \times W_1 \times C_1}$ with resolution (H_1,W_1) and C_1 input dimension via an image encoder e^I and a linear projection layer p^I into a global feature $f_i^G \in \mathbb{R}^C$ and a spatial feature $f_i^S \in \mathbb{R}^{\hat{H}_1 \hat{W}_1 \times C}$, where (\hat{H}_1,\hat{W}_1) and C denote the resolution and dimension of the feature map, respectively:

$$f_i^G, f_i^S = p^I \left(e^I \left(I_i \right) \right). \tag{1}$$

Following the similar processing pipeline, $R_i \in \mathbb{R}^{N \times C_2}$ with token length N and C_2 dimension is converted into a C-dimension global representation f_i^R by a text encoder e^T and a linear projection function p^T :

$$f_i^R = p^T \left(e^T \left(R_i \right) \right). \tag{2}$$

Note that our model is agnostic to the specific option of image and text encoders. Following previous work [38], we apply ResNet [16] and ViT [10] as the image encoders e^I . The main difference between them is that ResNet performs a global attention pooling on the spatial feature f_i^S to obtain the global feature f_i^G , while f_i^G of ViT is the corresponding output of [class] token. As for the text encoder e^T , we follow the encoder part of the Transformer [47] architecture. The projectors p^I and p^T map the representations of images and medical reports into the same space of C dimension so that contrastive learning can be applied. Based on the bidirectional image-to-report and report-to-image InfoNCE losses [36], the global image-report contrastive loss for each training mini-batch can be formulated as:

$$\mathcal{L}_{i2r} = -\log \frac{e^{\cos(f_i^G, f_i^R)/\tau}}{\sum_{j=1}^{B_1} e^{\cos(f_i^G, f_j^R)/\tau}},$$
(3)

$$\mathcal{L}_{\text{r2i}} = -\log \frac{e^{\cos(f_i^R, f_i^G)/\tau}}{\sum_{j=1}^{B_1} e^{\cos(f_i^R, f_j^G)/\tau}},$$
 (4)

$$\mathcal{L}_{\text{irc}} = \frac{1}{2B_1} \sum_{i=1}^{B_1} \left(\mathcal{L}_{i2r} + \mathcal{L}_{r2i} \right), \tag{5}$$

where $\cos(\cdot, \cdot)$ denotes the cosine similarity, $\cos\left(f_i^G, f_i^R\right) = \left(f_i^G\right)^\mathsf{T} f_i^R / \|f_i^G\| \|f_i^R\|$, T represents the transpose operation, $\|\cdot\|$ denotes the L2 normalization, and τ is the learnable temperature parameter and initializes to 0.07 following [38].

Furthermore, compared with natural image-text pairs [38, 23], the publicly available medical multimodal datasets [37, 8, 24] are relatively small to train a generalizable model. Thus we employ CLIP model parameters for initialization. A multitude of medical image-report pairs are subsequently employed for the purpose of fine-tuning CLIP within the medical domain.

3.2.1.2. Local pixel-tag aligning

In the pre-training stage, we expect to gain more language supervision to learn medical visual representations. The global image-report contrastive learning, however, mainly considers coarse-grained representations of both images and medical reports, while downstream tasks of multi-organ segmentation are pixel-level. To narrow the substantial gap

between these stages, we introduce multi-label recognition to implicitly align the image pixels and the organ tags to obtain more fine-grained information.

Multi-label recognition predicts whether each organ tag exists in the radiology images. Unlike the original Query2Label [30] that directly used learnable label embeddings as the input queries, we introduce K-class tags as the input, which can be transferred to downstream segmentation tasks based on medical report supervision better. The details of constructing the tag list can be found in Sec 3.1.1. Motivated by CoOp [62], we apply the learnable textual context to mitigate the domain gap between tags and medical reports. Then the input of the shared text encoder e^T becomes:

$$T_k = \langle p_k, t_k \rangle, \quad 1 \le k \le K,$$
 (6)

where $p_k \in \mathbb{R}^{N_1 \times C_2}$ is the learnable textual context, shared in K-class tags. $t_k \in \mathbb{R}^{N_2 \times C_2}$ is the embedding of k-th organ tag, $\langle \cdot, \cdot \rangle$ denotes the concatenation, and N_1 and N_2 are the token lengths of the learnable textual context and the tag, respectively. Similar to the procedure of medical reports, we get the global representation $f_k^T \in \mathbb{R}^C$ of the tag:

$$f_k^T = p^T \left(e^T \left(T_k \right) \right). \tag{7}$$

On the basis of the spatial feature f_i^S of the input radiology image obtained in Sec 3.2.1.1, we treat $f^T \in \mathbb{R}^{K \times C}$ as queries and leverage the cross-attention mechanism in Transformer decoder [47] to progressively integrate category-related contextualized information from the input image into the query embeddings:

$$f_{i}^{TS} = \text{TransDecoder}\left(f^{T}, f_{i}^{S}, f_{i}^{S}\right), \tag{8}$$

where $f_i^{TS} \in \mathbb{R}^{K \times C}$ are the updated queries. To perform multi-label recognition, we regard predicting each label as a binary classification task and map the feature $f_{i,k}^{TS} \in \mathbb{R}^C$ for k-th category of i-th sample into a logit value applying a linear projection layer p^{TS} followed by a sigmoid function:

$$y_{i,k} = \text{Sigmoid}\left(p^{TS}\left(f_{i,k}^{TS}\right)\right), \tag{9}$$

where $y_{i,k} \in [0,1]$ is the predicted probability for k-th category of i-th sample. We denote the ground-truth labels of input image I_i as $x_i = \left[x_{i,1}, \cdots, x_{i,K}\right]$ where $x_{i,k} \in \{0,1\}$ is a discrete binary label. $x_{i,k} = 1$ if the k-th organ tag presents in the corresponding medical report R_i , otherwise $x_{i,k} = 0$. Medical reports usually only describe organs that appear abnormal on radiology images, so there may be plenty of false negative labels. To address this issue, we adopt a simple and effective loss, that is, weak assume negative loss [7], which introduces a weight parameter $\gamma \in [0,1]$ based on binary cross-entropy loss to reduce the effect of false negatives. For a training mini-batch, the multi-label recognition loss is defined as:

$$\mathcal{L}_{\text{mlr}} = -\frac{1}{B_1 K} \sum_{i=1}^{B_1} \sum_{k=1}^{K} \begin{cases} \log(y_{i,k}), & x_{i,k} = 1, \\ \gamma \log(1 - y_{i,k}), & x_{i,k} = 0, \end{cases}$$
(10)

where $\gamma = 1/(K-1)$ ensures that the approximate single positive label has the same impact on the loss as the K-1 assumed negatives.

Formally, we minimize the total loss function of pretraining tasks of MOSMOS as:

$$\mathcal{L}_{\text{total up}} = \mathcal{L}_{\text{irc}} + \mathcal{L}_{\text{mlr}}.$$
 (11)

3.2.2. Fine-tuning

Since MOSMOS learns visual representations from medical report supervision in the pre-training stage, we would like to explore the effect of transferring the pre-trained model to multi-organ segmentation tasks. Note that our framework is model-agnostic. For our investigation, we consider two main medical segmentation methods, 2D U-Net [40] and 3D UNETR [14], that adopt ResNet [16] and ViT [10] as their image encoders, respectively. To evaluate the contribution of MOSMOS, we substitute the image encoders with the pre-trained ones and introduce the pre-trained language supervision for multi-label recognition without the classifier.

3.2.2.1. Weakly supervised positioning

Thanks to the cross-attention mechanism in Transformer decoder [47], the generated pixel-tag attention maps can provide weakly supervised information. Specifically, the attention maps incorporate language supervision into medical visual representations and roughly locate the spatial distribution of the organ tags in the medical images. Take a B_2 mini-batch of 3D radiology images $\bar{I} \in \mathbb{R}^{B_2 \times H_2 \times W_2 \times D_2 \times C_1}$ and corresponding O-class organ tag embeddings $\bar{T} \in$ $\mathbb{R}^{Q \times (N_1 + N_2) \times C_2}$, for example. We obtain the spatial features of images $\bar{f}^S \in \mathbb{R}^{B_2 \times \hat{H}_2 \hat{W}_2 \hat{D}_2 \times C}$ and the global features of tags $\bar{f}^T \in \mathbb{R}^{Q \times C}$ using the pre-trained image and text encoders followed by corresponding projectors, respectively, where $H_2 \times W_2 \times D_2$ and $\hat{H}_2 \hat{W}_2 \hat{D}_2$ represent the height, width, and depth of the input images and feature maps, respectively. Regarding \bar{f}^T as queries and \bar{f}^S as keys and values, we pass these features to the Transformer decoder and gain the pixel-tag attention maps $\bar{f}^M \in \mathbb{R}^{B_2 \times \hat{H}_2 \hat{W}_2 \hat{D}_2 \times Q}$:

$$\bar{f}^M = \text{TransDecoder}(\bar{f}^T, \bar{f}^S, \bar{f}^S).$$
 (12)

The attention maps represent the degree of pixel-tag aligning, which play a significant role in our framework. Firstly, the attention maps can be concatenated with the visual spatial features to integrate medical language prior to guide the segmentation, that is, $\bar{f}^{SM} = \left\langle \bar{f}^S, \bar{f}^M \right\rangle \in \mathbb{R}^{B_2 \times \hat{H}_2 \hat{W}_2 \hat{D}_2 \times (C+Q)}$, and then fed into the image decoder. We obtain the predicted output $Y_{\text{seg}} \in \mathbb{R}^{B_2 \times H_2 W_2 D_2 \times Q}$. Secondly, we can regard the attention maps as the segmentation results with lower resolution, and thus upsample them to the original resolution by linear interpolation LI to calculate a pixel-tag aligning loss:

$$Y_{\text{pta}} = \text{LI}\left(\bar{f}^M/\epsilon\right),$$
 (13)

where $Y_{\text{pta}} \in \mathbb{R}^{B_2 \times H_2 W_2 D_2 \times Q}$ is the pixel-tag aligning output, and ε denotes a learnable temperature coefficient and initializes to 0.07 following [15].

3.2.2.2. Multi-organ segmentation

In addition to the segmentation loss \mathcal{L}_{seg} , we propose a pixeltag aligning loss \mathcal{L}_{pta} to make better use of the pixel-tag attention maps and help dense segmentation tasks converge faster. Both losses are a combination of cross-entropy loss and dice loss [34]:

$$\mathcal{L}(X,Y) = \frac{1}{B_2} \sum_{b=1}^{B_2} \left(1 - \frac{1}{V} \sum_{v=1}^{V} \sum_{q=1}^{Q} X_{b,v,q} \log Y_{b,v,q} - \frac{2}{Q} \sum_{q=1}^{Q} \frac{\sum_{v=1}^{V} X_{b,v,q} Y_{b,v,q}}{\sum_{v=1}^{V} X_{b,v,q}^2 + \sum_{v=1}^{V} Y_{b,v,q}^2} \right),$$
(14)

$$\mathcal{L}_{\text{seg}} = \mathcal{L}\left(X, Y_{\text{seg}}\right),\tag{15}$$

$$\mathcal{L}_{\text{pta}} = \mathcal{L}\left(X, Y_{\text{pta}}\right),\tag{16}$$

where $X \in \mathbb{R}^{B_2 \times H_2 W_2 D_2 \times Q}$ and $Y \in \mathbb{R}^{B_2 \times H_2 W_2 D_2 \times Q}$ denote the one-hot encoded ground truth and the predicted output, respectively, and V is the number of pixels.

The final loss function is a linear combination of the above two parts:

$$\mathcal{L}_{\text{total down}} = \mathcal{L}_{\text{seg}} + \lambda \mathcal{L}_{\text{pta}}, \tag{17}$$

where λ is the hyper-parameter to balance the two-part losses.

4. Results

In this section, we present the experimental details and analyze the results to demonstrate the flexibility and generalization of our proposed multi-organ segmentation algorithm that is facilitated by medical report supervision.

4.1. Implementation details

We implement MOSMOS in PyTorch on a single NVIDIA V100 GPU. Two segmentation baselines are considered, that is, U-Net [40] with ResNet-50 [16] and UN-ETR [14] with ViT-B/16 [10] visual backbones. The textual backbone is the same text encoder as in the CLIP [39]. For the sake of a comprehensive analysis, we compare our method with the following seven methods. To ensure a fair comparison, we implement the other methods using the same backbone and hyper-parameter settings as those applied in MOSMOS. The detailed hyper-parameters are listed in Table 1.

- **Random Init.**: The visual backbone of the baseline is initialized using default random initialization.
- ImageNet [9] Init.: The visual backbone of the baseline is initialized with weights pre-trained on ImageNet.
- Inpainting+Contrast+Rotation [46]: The visual backbone of the baseline is pre-trained through the utilization of three self-supervised proxy tasks on ROCO images, specifically, mask volume inpainting, contrastive learning, and rotation prediction.

 Table 1

 Hyper-parameters applied in our MOSMOS.

Symbo	Hyper-Parameter	Value
τ	the temperature parameter of InfoNCE loss	0.07
ε	the temperature parameter of linear interpolation	0.07
λ	the weight of pixel-tag aligning loss	8.0

- **CLIP** [38]: The visual backbone of the baseline is initialized with weights pre-trained on CLIP.
- CLIP+DenseCLIP [39]: The visual and textual backbones of the DenseCLIP are initialized with weights pre-trained on CLIP.
- PubMedCLIP [12]: The visual backbone of the baseline is initialized with weights pre-trained on PubMed-CLIP.
- PubMedCLIP+DenseCLIP [39]: The visual and textual backbones of the DenseCLIP are initialized with weights pre-trained on PubMedCLIP.

In the pre-training stage, we resize all 2D images to $H_1 \times W_1 = 224 \times 224$ as the input resolution and set the token lengths of the medical reports, learnable textual context, and tags to N=77, $N_1=16$, and $N_2=10$, respectively. The feature dimensions of input images, input texts, and outputs are $C_1=768$, $C_2=512$, and C=512, respectively. The network is trained for 50 epochs with a fixed batch size B_1 of 64, and the optimizer is Adam [25] with the learning rate of 10^{-5} . We compute the validation loss after every epoch and save the checkpoint with the lowest validation loss.

During the fine-tuning stage, all images are preprocessed following the procedures in [14]. For training, we randomly crop 3D images into a resolution of $H_2 \times W_2 \times D_2 =$ $96 \times 96 \times 96$. For 2D images, the D_2 is omitted. We train the whole network using AdamW optimizer [33] and set the initial learning rate of 10^{-4} for 5,000 epochs. After 50 epochs, the learning rate is decayed according to the cosine attenuation approach [17]. Given the memory constraints, we set the batch size B_2 to 96 for ResNet-50-based and 2 for ViT-B/16-based methods. The text encoder is fixed to retain more medical language supervision learned from the largescale image-report pre-training. For inference, we apply the sliding window method with an overlap ratio of 0.5 and keep the same resolution as the training sets. We calculate the evaluation metrics every 100 epochs and select the models with the best values to perform the test.

4.2. Evaluation metrics

To objectively evaluate the segmentation performance, we apply the Dice similarity coefficient and Hausdorff Distance 95% (HD95) as the evaluation metrics. For a given organ category, let X_v and Y_v represent the ground truth and prediction for pixel v, and X' and Y' denote the ground truth and predicted surface point sets. The Dice and HD metrics are defined as:

Table 2

Quantitative comparisons of segmentation performance using Dice (%) metric on BTCV test set. The best results are bolded. The results of our approach are marked with a gray background color. We calculate the p-value between the average performance of our MOSMOS and PubMedCLIP+DenseCLIP [39] in Dice metric. Note: Spl: spleen, RKid: right kidney, LKid: left kidney, Gal: gallbladder, Eso: esophagus, Liv: liver, Sto: stomach, Aor: aorta, IVC: inferior vena cava, Vein: portal vein and splenic vein, Pan: pancreas, RAG: right adrenal gland, LAG: left adrenal gland.

Baseline	Method	Spl	RKid	LKid	Gal	Eso	Liv	Sto	Aor	IVC	Vein	Pan	RAG	LAG	Avg.↑
	Random [40]	91.07	91.66	91.73	41.28	71.76	95.07	73.08	87.36	79.54	65.87	66.34	64.44	54.04	74.86
	ImageNet [9]	92.87	91.16	91.87	60.00	69.83	96.19	71.08	89.80	81.84	67.59	73.99	58.03	53.20	76.73
	Inpainting+Contrast+Rotation [46]	91.62	89.85	89.30	48.94	71.81	95.35	71.56	90.32	82.70	68.49	70.77	65.46	57.25	76.42
U-Net [40]	CLIP [38]	95.01	93.20	93.14	50.91	71.11	96.48	76.04	89.08	82.47	71.09	65.54	64.38	46.70	76.55
(ResNet-50)	CLIP+DenseCLIP [39]	91.88	89.26	90.97	63.10	73.54	95.97	80.40	90.59	82.96	68.53	72.27	61.33	50.02	77.76
(Nesivet-30)	PubMedCLIP [12]	89.71	91.43	91.64	47.04	68.18	95.03	78.64	89.07	78.00	65.72	66.29	65.15	57.29	75.63
	PubMedCLIP+DenseCLIP [39]	93.09	90.04	90.89	62.67	69.94	95.48	81.25	90.77	79.95	67.27	69.35	63.01	50.47	77.24
	MOSMOS (Ours)	94.36	93.56	93.81	73.01	72.24	96.11	81.85	89.04	83.72	60.74	76.16	64.95	61.68	80.10
	P-value	2.7e-2 (Dice)													
	Random [14]	93.67	93.56	93.46	66.69	70.28	96.31	77.87	88.18	82.09	65.05	67.94	65.99	59.83	78.53
	ImageNet [9]	91.39	93.78	94.00	64.13	70.52	96.48	77.82	89.77	82.28	68.85	73.65	61.77	64.91	79.18
	Inpainting+Contrast+Rotation [46]	93.17	93.16	92.44	61.97	72.16	95.63	75.38	87.28	80.87	65.16	67.90	67.13	61.62	77.99
UNETR [14]	CLIP [38]	93.08	93.79	93.71	62.95	70.85	96.48	78.21	88.16	83.01	68.30	73.66	66.58	61.58	79.26
(ViT-B/16)	CLIP+DenseCLIP [39]	87.97	92.25	93.00	70.64	72.58	96.31	76.13	88.43	82.38	69.88	70.20	67.14	60.76	79.05
(VII-D/10)	PubMedCLIP [12]	92.75	93.27	93.08	64.96	70.84	96.10	78.27	87.97	82.18	68.18	73.55	65.57	60.46	79.01
	PubMedCLIP+DenseCLIP [39]	93.19	93.74	93.28	61.32	71.39	96.50	79.28	88.97	82.61	69.56	74.41	66.52	58.97	79.21
	MOSMOS (Ours)	93.21	93.68	93.39	72.71	72.96	96.46	80.46	88.84	82.83	70.39	75.00	66.90	57.87	80.36
	P-value							8.0e-2	(Dice)						

$$Dice = \frac{2\sum_{v=1}^{V} X_{v} Y_{v}}{\sum_{v=1}^{V} X_{v} + \sum_{v=1}^{V} Y_{v}},$$
(18)

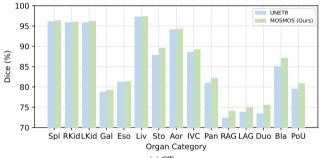
$$HD = \max \left\{ \max_{x' \in X'} \min_{y' \in Y'} \|x' - y'\|, \\ \max_{y' \in Y'} \min_{x' \in X'} \|y' - x'\| \right\},$$
(19)

where Dice measures the overlaps of ground truth and predicted values of V pixels, and HD95 calculates the 95th percentile of the surface distances between ground truth and predicted point sets.

4.3. Quantitative segmentation results

4.3.1. Abdominal multi-organ segmentation on BTCV

As shown in Table 2, we report the abdominal multiorgan segmentation results of our MOSMOS and other approaches with two different baselines on BTCV. We see that the pre-training methods generally perform better than training from scratch. Compared with other pretraining methods, MOSMOS consistently attains the highest Dice scores, both on average and across the majority of organ categories. This noteworthy accomplishment is attributed to the incorporation of two key components during the pre-training phase: global image-report alignment and local pixel-tag alignment. Specifically, our MOSMOS is 3.37% and 1.18% Dice higher than the ImageNet-based pretraining [9] on ResNet-50 and ViT-B/16 visual backbones, respectively. MOSMOS also surpasses the state-of-the-art self-supervised pre-trained baselines (denoted by Inpainting+Contrast+Rotation [46]) by 3.68% and 2.37% on average of 13 organs. Besides, MOSMOS consistently maintains



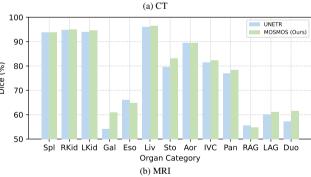


Figure 4: The indication of Dice gap between UNETR (Blue) and MOSMOS (Green) on AMOS validation sets for CT (a) and MRI (b). Notably, Duo, Bla, and PoU belong to open-set organ categories. Note: Spl: spleen, RKid: right kidney, LKid: left kidney, Gal: gallbladder, Eso: esophagus, Liv: liver, Sto: stomach, Aor: aorta, IVC: inferior vena cava, Pan: pancreas, RAG: right adrenal gland, LAG: left adrenal gland, Duo: duodenum, Bla: bladder, PoU: prostate or uterus.

advantages of at least 1.10% with respect to these contrastive language-image pre-training models [38, 39, 12]. Although MOSMOS with ViT-B/16 visual backbone does not improve

Table 3

Quantitative comparisons of segmentation performance using Dice (%) and HD95 (mm) metrics on MMWHS test set. The best results are bolded. The results of our approach are marked with a gray background color. The p-values are computed based on the average performance of our MOSMOS and PubMedCLIP+DenseCLIP [39] in both Dice and HD95 metrics. Note: Myo: myocardium, LA: left atrium, LV: left ventricle, RA: right atrium, RV: right ventricle, AA: ascending aorta, PA: pulmonary artery.

Baseline	Method	N	1yo	LA		- 1	LV		RA	RV		AA		PA		Avg.	
Daseiine	Method	Dice↑	HD95↓	Dice†	HD95↓	Dice†	HD95↓	Dice†	HD95↓	Dice†	HD95↓	Dice†	HD95↓	Dice†	HD95↓	Dice†	HD95↓
	Random [40]	79.92	2.89	85.00	3.47	90.57	2.99	83.35	7.16	82.92	4.74	74.43	11.48	71.29	6.12	81.07	5.55
	ImageNet [9]	82.76	2.46	87.19	3.05	93.39	2.39	85.91	6.51	90.02	3.16	72.88	12.04	71.10	6.03	83.32	5.09
	Inpainting+Contrast+Rotation [46]	82.91	2.08	85.47	3.61	93.69	1.93	86.15	9.18	89.24	3.77	73.74	11.84	64.37	8.33	82.22	5.82
U-Net [40]	CLIP [38]	81.87	2.49	84.66	3.46	92.79	3.23	81.17	7.65	82.64	5.72	71.18	11.55	70.31	4.90	80.66	5.57
(ResNet-50)	CLIP+DenseCLIP [39]	82.30	2.46	86.34	3.18	93.74	1.78	78.77	5.76	87.44	3.45	73.02	12.22	71.07	5.15	81.81	4.86
(Resivet-50)	PubMedCLIP [12]	79.74	2.55	81.55	4.06	93.50	2.19	81.00	7.99	87.05	5.15	71.09	13.13	65.26	6.90	79.88	5.99
	PubMedCLIP+DenseCLIP [39]	81.39	2.35	84.95	3.24	93.48	2.16	85.80	9.32	88.65	4.01	69.13	15.59	65.10	10.38	81.21	6.72
	MOSMOS (Ours)	83.74	2.00	84.99	3.46	93.78	1.70	86.51	6.76	90.89	3.25	74.87	13.22	70.24	6.81	83.57	5.31
	P-values	1.6e-2 (Dice), 3.1e-2 (HD95)															
	Random [14]	84.21	1.83	83.96	4.02	93.85	1.93	86.20	7.08	91.72	2.88	82.27	6.01	76.46	8.25	85.52	4.57
	ImageNet [9]	83.96	1.91	83.38	4.58	93.35	2.23	84.39	6.84	89.96	3.56	82.82	8.19	77.62	7.61	85.07	4.99
	Inpainting+Contrast+Rotation [46]	85.03	1.77	87.16	3.47	94.06	1.83	87.18	6.31	92.42	2.48	82.91	5.22	77.87	5.69	86.66	3.82
UNETR [14]	CLIP [38]	84.72	1.83	86.88	3.55	94.03	1.93	87.19	4.85	92.31	2.63	83.54	5.37	78.74	6.86	86.77	3.86
(ViT-B/16)	CLIP+DenseCLIP [39]	84.74	1.77	86.54	3.44	93.83	1.83	87.85	5.35	92.01	3.11	84.80	5.08	78.35	5.47	86.87	3.72
	PubMedCLIP [12]	84.67	1.83	85.65	4.00	93.91	1.85	87.58	4.36	92.93	2.76	83.53	5.75	78.84	5.33	86.73	3.70
	PubMedCLIP+DenseCLIP [39]	82.69	2.24	79.35	5.11	92.86	2.58	82.60	8.05	88.75	10.47	79.88	7.69	69.36	8.90	82.21	6.44
	MOSMOS (Ours)	84.65	1.77	87.62	3.48	94.02	1.85	88.61	4.62	93.22	2.30	84.08	4.60	79.49	5.60	87.38	3.46
	P-values		1.6e-2 (Dice), 1.6e-2 (HD95)														

as much as with ResNet-50, it outperforms using ResNet-50, so ViT is more suitable for multi-organ segmentation tasks. As for why MOSMOS does not perform best in some organs, we consider that the feature extraction differences in visual backbones affect the positioning capability of attention maps.

4.3.2. Abdominal multi-organ segmentation on AMOS

A performance comparison of multi-organ segmentation tasks on the AMOS dataset for both CT and MRI modalities using MOSMOS versus the baseline UNETR [14] is presented in Fig. 4. As depicted in Fig. 4a, MOSMOS consistently outperforms UNETR across all CT segmentation tasks on AMOS, with an average Dice score improvement from 85.37% to 86.29%. Significant improvements can be observed in the closed-set tasks of the stomach, pancreas, right adrenal gland, left adrenal gland, and the open-set tasks of the duodenum, bladder, prostate or uterus, with Dice scores advancing from 87.89% to 89.63%, 80.99% to 82.24%, 72.38% to 74.07%, 73.88% to 74.98%, 73.47% to 75.50%, 85.06% to 87.11%, and 79.48% to 80.87%, respectively. In Fig. 4b, for all MRI tasks on AMOS, the average Dice score increases from 76.84% to 78.17%. Distinct improvements are evident in the closed-set stomach category and the open-set duodenum category, with Dice scores improving from 79.58% to 83.05% and 57.19% to 61.48%, respectively. The gallbladder category in the closedset displays the most substantial improvement, with a Dice score of 60.99% compared to 54.08%.

4.3.3. Cardiac substructure segmentation on MMWHS

Table 3 presents the class-specific results in both Dice and HD95 metrics on cardiac substructure segmentation using the MMWHS dataset. Compared to previous approaches, the proposed MOSMOS displays more strength in Dice than

in HD95. Specifically speaking, our MOSMOS outperforms other methods in 9 out of 14 categories in Dice under two different baselines, while 5 out of 14 categories in HD95. From the perspective of average performance, we notice that MOSMOS often achieves better capability. For instance, we gain the state-of-the-art Dice of 83.57% and 87.38% when adopting U-Net [40] and UNETR [14] as the segmentation baselines, respectively, while keeping the lowest HD95 in UNETR baseline. Moreover, MOSMOS surpasses DenseCLIP [39]—the best-performing contrastive language-image pre-training approach—by over 0.51% in Dice.

4.3.4. Statistical significance

In Table 2 and Table 3, we employ Wilcoxon signed rank test to calculate p-values between the average performance of our MOSMOS and PubMedCLIP+DenseCLIP [39] in both Dice and HD95 metrics. As we can see, MOSMOS demonstrates statistically significant performance, yielding p-values below 5e-2 across both Dice and HD95 metrics on two distinct baselines (U-Net & UNETR) and two public datasets (BTCV & MMWHS). The sole exception is observed with UNETR baseline on the BTCV dataset. These findings indicate that, in general, MOSMOS has significant advantages over PubMedCLIP+DenseCLIP.

4.4. Analytical ablation studies 4.4.1. Different modules in MOSMOS

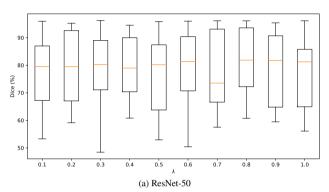
We provide a thorough empirical study on MOSMOS by removing the individual modules in Table 4. For simplicity, we conduct experiments on the BTCV dataset and choose UNETR as the default baseline.

First, we investigate the effect of medical image-report contrastive learning in the pre-training stage. By comparing row 1 with row 0, we observe that dropping the cross-modal

Table 4

Exploration of the effects of different modules used in MOSMOS. MLR and IRC denote the multi-label recognition loss and the image-report contrastive loss applied in the pre-training stage, respectively. Prompt represents the learnable textual context. CLIP refers to the introduction of the pre-trained CLIP parameters. We highlight the best results in bold front.

Row	MLR	IRC	Prompt	CLIP	Avg. Dice↑	Avg. HD95↓
0	1	/	/	1	80.36	6.39
1	✓		✓	✓	79.82	7.98
2	✓	1		✓	79.06	8.03
3	✓	1	✓		78.77	9.94
4			✓		75.01	42.95



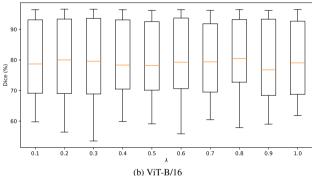


Figure 5: Dice box plots of our approach based on ResNet-50 (a) and ViT-B/16 (b) visual backbones for BTCV, when varying the weight λ of pixel-tag aligning loss from 0.1 to 1.0.

contrastive task would adversely affect the overall performance by 0.54% and 1.59mm in average Dice and HD95, respectively. We argue the reason behind this is that the global image-report aligning is a prerequisite for the local pixel-tag aligning and thus benefits downstream segmentation tasks. Next, we remove the learnable textual context so that each organ tag is embedded alone by the text encoder (row 2). Such an operation causes Dice to drop by 1.30% and HD95 to rise by 1.64mm (compared with row 0). This phenomenon demonstrates the helpfulness of mitigating the gaps between organ tags and reports. In addition, we consider not adopting CLIP parameters for initialization (row 3), where we can see a 1.59% decrease in Dice and a 3.55mm increase in HD95. This result verifies the advantage of large-scale cross-modal pre-training. Last but not least, we evaluate the significance of the entire pre-training process. A comparison between

Table 5

Segmentation performance comparisons of the ViT-B backbone with different patch resolutions using Dice (%) metric. We highlight the best results in bold front.

Patch	ВТО	CV	MMWHS					
Resolution	UNETR [14]	MOSMOS	UNETR [14]	MOSMOS				
32	77.13	78.24	77.86	80.79				
16	78.53	80.36	85.52	87.38				

Row 4 and Row 0 reveals that the integration of supervision derived from medical reports can markedly improve overall performance. Such an improvement is attributed to the introduction of comprehensive medical prior knowledge without any additional manpower expense.

4.4.2. Different weights of the pixel-tag aligning loss

We vary the weight of pixel-tag aligning loss to explore the sensitivity of results to the trade-off parameter λ in Eq. (17). To be specific, we range $\lambda \in [0.1, 1.0]$ at a step of 0.1 and analyze the organ-wise segmentation performance on the BTCV dataset. As shown in Fig. 5a, the box plot displays the average Dice across each organ of our method based on the ResNet-50 visual backbone. Our MOSMOS achieves the best performance on the BTCV test set when the λ is set to 0.8. The performance fluctuations are very small, except when λ is 0.7. In comparison, our MOSMOS is capable of generally outperforming the baseline (e.g., 74.86% for U-Net shown in Table 2). In Fig. 5b, we compare the performance of MOSMOS based on the ViT-B/16 visual backbone with different λ . Although MOSMOS attains the highest Dice score when λ equals 0.8, it exhibits low sensitivity to λ . Considering the performance across both visual backbones, we empirically set λ to 0.8.

4.4.3. Different patch resolutions of the ViT-B backbone

In Table 5, we compare the average performance of the ViT-B visual backbone with different input patch resolutions. It shows that the performance significantly improves when decreasing the patch resolution. Specifically, dropping the resolution from 32 to 16 boosts the Dice of our MOSMOS by 2.12% and 6.59% on BTCV and MMWHS datasets, respectively. We can also observe that the proposed MOSMOS consistently maintains a significant advantage over the baseline UNETR in different resolutions and datasets. However, a lower patch resolution leads to a longer sequence and, therefore, higher memory cost. Considering the trade-off between segmentation performance and memory consumption, we empirically set the input patch resolution of ViT-B to 16.

4.4.4. Different label ratios in the fine-tuning stage

Fig. 6 displays the performance comparison of various approaches under different label ratios on BTCV test dataset. Using only 25% of labeled data, our MOSMOS achieves a 7% improvement in performance compared to training a model from scratch. When utilizing the full set of labeled

data, MOSMOS outperforms models trained from scratch or those using other pre-training methods by an average Dice score increase of 3.37%. Notably, MOSMOS only needs 75% of the annotated training data to match the performance comparable with those of other methods under a 100% labeled ratio. This highlights MOSMOS's efficiency in reducing annotation efforts by approximately 25% for the multi-organ segmentation task on BTCV.

4.5. Visualization for qualitative segmentation results

Fig. 7 illustrates the segmentation and weakly supervised positioning results for qualitative evaluation. We mainly visualize the segmentation maps of our MOSMOS and other pre-training approaches on two public datasets. Compared to training from scratch and self-supervised or imagereport contrastive pre-training, our MOSMOS displays visual improvements in capturing the shape of inferior vena cava (IVC, row 1 on BTCV) and pancreas (Pan, row 1 on BTCV), right atrium (RA, row 1 on MMWHS), and ascending aorta (AA, row 1 and row 2 on MMWHS). In addition, MOSMOS can reduce the prediction of false positives. One representative example is the second case on BTCV. Other methods predict the wrong liver (Liv) pixels near the stomach (Sto). Furthermore, the weakly supervised positioning results show that MOSMOS can distinguish between left and right organs through tailored pre-training tasks, demonstrating the superiority of our approach.

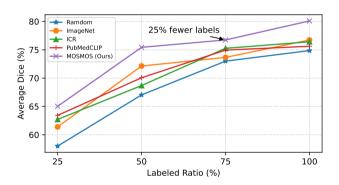


Figure 6: Performance of different approaches under various label ratios on BTCV test dataset. We highlight the percentage of annotated training data in the fine-tuning stage that our MOSMOS needs to achieve results comparable with those obtained by training from scratch or using other pre-training methods. Note that all five methods utilize the same ResNet-50 visual backbone. Note: ICR: Inpainting+Contrast+Rotation.

5. Discussion

5.1. Strengths

We verify the effectiveness and generalization of our MOSMOS on multi-disease, multi-modal, and multi-organ datasets. Considering the cost of collection and annotation, BTCV, AMOS, and MMWHS datasets consist of a small number of annotated images, which cannot train effective models using randomly initialized weights from scratch.

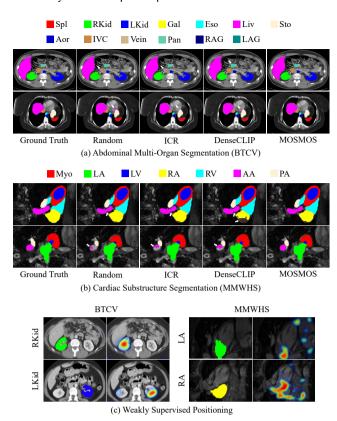


Figure 7: Visualization of segmentation and weakly supervised positioning results. We mainly compare different pre-training strategies for the UNETR baseline on two widely used datasets. Note: ICR: Inpainting+Contrast+Rotation, DenseCLIP: Pub-MedCLIP+DenseCLIP, Spl: spleen, RKid: right kidney, LKid: left kidney, Gal: gallbladder, Eso: esophagus, Liv: liver, Sto: stomach, Aor: aorta, IVC: inferior vena cava, Vein: portal vein and splenic vein, Pan: pancreas, RAG: right adrenal gland, LAG: left adrenal gland, Myo: myocardium, LA: left atrium, LV: left ventricle, RA: right atrium, RV: right ventricle, AA: ascending aorta, PA: pulmonary artery. Error predictions are pointed by white arrows (best views in color).

Thus we make a thorough comparison of the pre-training strategies. We observe that MOSMOS outperforms the previous pre-training approaches on most of the average metrics and organ categories. The main reasons for this are: (i) The ImageNet-based pre-training utilizes nature images, which exit enormous differences from medical images. In addition, this method is supervised and requires extensive annotations. (ii) Without the need for annotation effort, the self-supervised pre-training approach designs proxy tasks to learn solely visual representations, which does not introduce additional potentially exploitable supervisory information and has a gap with downstream tasks. (iii) As for the imagereport contrastive pre-training, it adopts language priors paired with medical images as supervision without extra human effort. However, the visual spatial features transferred downstream are indirectly aligned to the text embeddings via the visual global features. In contrast, MOSMOS directly aligns the visual spatial features and the tag embeddings corresponding to the organ tags by introducing multi-label recognition in the pre-training stage, which can roughly

Table 6
Segmentation performance comparisons of UNETR and MOSMOS using Dice (%) metric on BRATS test set. We highlight the best results in bold front. Note: TC: Tumor Core, WT: Whole Tumor, ET: Enhancing Tumor.

Method	TC	WT	ET	Avg.↑
UNETR [14]	66.70	83.31	69.38	73.13
MOSMOS	65.79	84.43	69.27	73.16

identify the same organ with different shapes and sizes using attention maps in the Transformer decoder. Unlike the traditional multi-label classification, which encodes the multiple labels into a string of numbers as input, we take the embeddings of multiple organ tags as input. In this way, our MOSMOS is scalable and generalized. Meanwhile, MOSMOS is suitable for any segmentation model. The performance improvements on U-Net and UNETR demonstrate the universality of the MOSMOS framework.

5.2. Limitations

Our approach still has some limitations that can be improved in future works.

First, the proposed MOSMOS has only been pre-trained using 2D medical image-report pairs and transferred to 2D and 3D multi-organ segmentation tasks. This is mainly due to the lack of publicly available 3D image-report pairs, which are more consistent with clinical practice in most medical imaging modalities. In future effort, we will extend our framework to 3D image-report pre-training by constructing this dataset.

Second, in the current pre-training stage, we have constructed only 20 limited organ tag categories, primarily focused on abdominal multi-organs and cardiac substructures, which are not sufficient for fine-grained segmentations of the entire complex human body organs. Despite this, the diversity of medical reports in the pre-training stage provides a preliminary basis, as demonstrated in Fig. 4, for MOSMOS to exhibit a degree of open-set segmentation capability for abdominal organs on the AMOS dataset. Therefore, we can further refine our approach by expanding the tag list used in the pre-training stage and developing more advanced algorithms for open-set multi-organ segmentation.

Third, we mainly focus on organ segmentation, but ignore the descriptions of lesion morphology, size, location, and number in the reports, which can guide more significant tasks of fine-grained lesion segmentation. To further explore the generalization of our MOSMOS model, we extend its application to a slightly out-of-domain task—brain tumor segmentation on the BRATS dataset. Table 6 shows the performance comparison in Dice score between MOSMOS and the baseline UNETR. MOSMOS surpasses UNETR by 1.12% Dice in whole tumor segmentation, yet exhibits comparable or inferior performance in the more granular tumor core and enhancing tumor segmentation tasks. Due to the significant differences between organs and lesions,

the performance improvement in open-set brain tumor segmentation on BRATS is not substantial. Consequently, we aim to optimize our framework to be more suitable for lesion segmentation by mining the medical reports for more detailed information to further demonstrate its generality.

6. Conclusions

In this paper, we present a novel framework, dubbed MOSMOS, for multi-organ segmentation by leveraging cross-modal pre-training with medical image-report pairs. Based on global image-report aligning, MOSMOS first introduces the proxy task of local pixel-tag aligning. It utilizes a multi-label recognition approach to position the organ tags extracted from reports in the corresponding images, which is more suitable for complex fine-grained segmentation tasks in the downstream. Thus, the proposed framework is capable of being general for multi-disease, multi-modal, and multi-organ segmentation tasks on both 2D and 3D networks.

CRediT authorship contribution statement

Weiwei Tian: Writing - review & editing, Writing original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. Xinyu Huang: Writing - review & editing, Writing original draft, Visualization, Validation, Supervision, Software, Resources, Methodology, Investigation, Data curation, Conceptualization. Junlin Hou: Writing - review & editing, Writing - original draft, Visualization, Validation, Software, Methodology, Formal analysis. Caivue Ren: Writing - review & editing, Writing - original draft, Visualization, Validation, Software, Formal analysis. Longquan Jiang: Writing - review & editing, Writing - original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Investigation, Funding acquisition. Rui-Wei Zhao: Writing – review & editing, Writing – original draft, Visualization, Validation, Funding acquisition, Formal analysis. **Gang Jin:** Writing – review & editing, Writing - original draft, Supervision, Conceptualization. Yuejie Zhang: Writing - review & editing, Writing - original draft, Supervision, Funding acquisition, Formal analysis. Daoying Geng: Writing - review & editing, Writing original draft, Supervision, Investigation, Formal analysis.

Declaration of competing interest

The authors declare that they have no competing interests.

Data availability

Data will be made available on request.

Acknowledgments

This work was supported in part by the Science and Technology Commission of Shanghai Municipality (No.22511106003, No.23511100602) and the Shanghai Research and Innovation Functional Program under Grant 17DZ2260900.

References

- [1] Boecking, B., Usuyama, N., Bannur, S., Castro, D.C., Schwaighofer, A., Hyland, S., Wetscherek, M., Naumann, T., Nori, A., Alvarez-Valle, J., et al., 2022. Making the most of text semantics to improve biomedical vision-language processing, in: European conference on computer vision, Springer. pp. 1–21.
- [2] Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., Wang, M., 2022. Swin-unet: Unet-like pure transformer for medical image segmentation, in: European conference on computer vision, Springer. pp. 205–218.
- [3] Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A.L., Zhou, Y., 2021. Transunet: Transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306.
- [4] Chen, L., Bentley, P., Mori, K., Misawa, K., Fujiwara, M., Rueckert, D., 2019a. Self-supervised learning for medical image analysis using image context restoration. Medical image analysis 58, 101539.
- [5] Chen, S., Ma, K., Zheng, Y., 2019b. Med3d: Transfer learning for 3d medical image analysis. arXiv preprint arXiv:1904.00625.
- [6] Chen, Z., Li, G., Wan, X., 2022. Align, reason and learn: Enhancing medical vision-and-language pre-training with knowledge, in: Proceedings of the 30th ACM International Conference on Multimedia, pp. 5152–5161.
- [7] Cole, E., Mac Aodha, O., Lorieul, T., Perona, P., Morris, D., Jojic, N., 2021. Multi-label learning from single positive labels, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 933–942.
- [8] Demner-Fushman, D., Kohli, M.D., Rosenman, M.B., Shooshan, S.E., Rodriguez, L., Antani, S., Thoma, G.R., McDonald, C.J., 2016. Preparing a collection of radiology examinations for distribution and retrieval. Journal of the American Medical Informatics Association 23, 304–310.
- [9] Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L., 2009. Imagenet: A large-scale hierarchical image database, in: 2009 IEEE conference on computer vision and pattern recognition, Ieee. pp. 248– 255.
- [10] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N., 2021. An image is worth 16x16 words: Transformers for image recognition at scale. ICLR.
- [11] Dou, Q., Liu, Q., Heng, P.A., Glocker, B., 2020. Unpaired multi-modal segmentation via knowledge distillation. IEEE transactions on medical imaging 39, 2415–2425.
- [12] Eslami, S., de Melo, G., Meinel, C., 2021. Does clip benefit visual question answering in the medical domain as much as it does in the general domain? arXiv preprint arXiv:2112.13906.
- [13] Fang, X., Yan, P., 2020. Multi-organ segmentation over partially labeled datasets with multi-scale feature abstraction. IEEE Transactions on Medical Imaging 39, 3619–3629.
- [14] Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Myronenko, A., Landman, B., Roth, H.R., Xu, D., 2022. Unetr: Transformers for 3d medical image segmentation, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 574– 584
- [15] He, K., Fan, H., Wu, Y., Xie, S., Girshick, R., 2020. Momentum contrast for unsupervised visual representation learning, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 9729–9738.
- [16] He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition, in: Proceedings of the IEEE conference on

- computer vision and pattern recognition, pp. 770-778.
- [17] He, T., Zhang, Z., Zhang, H., Zhang, Z., Xie, J., Li, M., 2019. Bag of tricks for image classification with convolutional neural networks, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 558–567.
- [18] Huang, S.C., Shen, L., Lungren, M.P., Yeung, S., 2021. Gloria: A multimodal global-local representation learning framework for labelefficient medical image recognition, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 3942–3951.
- [19] Huang, X., Huang, Y.J., Zhang, Y., Tian, W., Feng, R., Zhang, Y., Xie, Y., Li, Y., Zhang, L., 2023a. Open-set image tagging with multigrained text supervision. arXiv e-prints, arXiv-2310.
- [20] Huang, X., Zhang, Y., Cheng, Y., Tian, W., Zhao, R., Feng, R., Zhang, Y., Li, Y., Guo, Y., Zhang, X., 2022. Idea: Increasing text diversity via online multi-label recognition for vision-language pretraining, in: Proceedings of the 30th ACM International Conference on Multimedia, Association for Computing Machinery, New York, NY, USA. p. 4573–4583. URL: https://doi.org/10.1145/3503161.3548108.
- [21] Huang, X., Zhang, Y., Ma, J., Tian, W., Feng, R., Zhang, Y., Li, Y., Guo, Y., Zhang, L., 2023b. Tag2text: Guiding vision-language model via image tagging. arXiv preprint arXiv:2303.05657.
- [22] Ji, Y., Bai, H., Ge, C., Yang, J., Zhu, Y., Zhang, R., Li, Z., Zhanng, L., Ma, W., Wan, X., et al., 2022. Amos: A large-scale abdominal multiorgan benchmark for versatile medical image segmentation. Advances in Neural Information Processing Systems 35, 36722–36732.
- [23] Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q., Sung, Y.H., Li, Z., Duerig, T., 2021. Scaling up visual and vision-language representation learning with noisy text supervision, in: International Conference on Machine Learning, PMLR. pp. 4904–4916.
- [24] Johnson, A.E., Pollard, T.J., Berkowitz, S.J., Greenbaum, N.R., Lungren, M.P., Deng, C.y., Mark, R.G., Horng, S., 2019. Mimic-cxr, a deidentified publicly available database of chest radiographs with free-text reports. Scientific data 6, 1–8.
- [25] Kingma, D.P., Ba, J., 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.
- [26] Landman, B., Xu, Z., Igelsias, J., Styner, M., Langerak, T., Klein, A., 2015. Miccai multi-atlas labeling beyond the cranial vault—workshop and challenge, in: Proc. MICCAI Multi-Atlas Labeling Beyond Cranial Vault—Workshop Challenge, p. 12.
- [27] Li, Z., Li, Y., Li, Q., Wang, P., Guo, D., Lu, L., Jin, D., Zhang, Y., Hong, Q., 2023. Lvit: language meets vision transformer in medical image segmentation. IEEE Transactions on Medical Imaging.
- [28] Liao, R., Moyer, D., Cha, M., Quigley, K., Berkowitz, S., Horng, S., Golland, P., Wells, W.M., 2021. Multimodal representation learning via maximization of local mutual information, in: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part II 24, Springer. pp. 273–283.
- [29] Lin, H., Li, Z., Yang, Z., Wang, Y., 2021. Variance-aware attention u-net for multi-organ segmentation. Medical Physics 48, 7864–7876.
- [30] Liu, S., Zhang, L., Yang, X., Su, H., Zhu, J., 2021a. Query2label: A simple transformer way to multi-label classification. arXiv preprint arXiv:2107.10834.
- [31] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B., 2021b. Swin transformer: Hierarchical vision transformer using shifted windows, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10012–10022.
- [32] Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3431– 3440
- [33] Loshchilov, I., Hutter, F., 2017. Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101.
- [34] Milletari, F., Navab, N., Ahmadi, S.A., 2016. V-net: Fully convolutional neural networks for volumetric medical image segmentation, in: 2016 fourth international conference on 3D vision (3DV), IEEE.

- pp. 565-571.
- [35] Müller, P., Kaissis, G., Zou, C., Rueckert, D., 2022. Joint learning of localized representations from medical images and reports, in: European Conference on Computer Vision, Springer. pp. 685–701.
- [36] Oord, A.v.d., Li, Y., Vinyals, O., 2018. Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748.
- [37] Pelka, O., Koitka, S., Rückert, J., Nensa, F., Friedrich, C.M., 2018. Radiology objects in context (roco): a multimodal image dataset, in: Intravascular Imaging and Computer Assisted Stenting and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis. Springer, pp. 180–189.
- [38] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al., 2021. Learning transferable visual models from natural language supervision, in: International Conference on Machine Learning, PMLR. pp. 8748–8763.
- [39] Rao, Y., Zhao, W., Chen, G., Tang, Y., Zhu, Z., Huang, G., Zhou, J., Lu, J., 2022. Denseclip: Language-guided dense prediction with context-aware prompting, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 18082–18091.
- [40] Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation, in: International Conference on Medical image computing and computer-assisted intervention, Springer. pp. 234–241.
- [41] Seibold, C., Reiß, S., Sarfraz, M.S., Stiefelhagen, R., Kleesiek, J., 2022. Breaking with fixed set pathology recognition through report-guided contrastive training, in: Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part V, Springer, pp. 690–700.
- [42] Sharbatdaran, A., Romano, D., Teichman, K., Dev, H., Raza, S.I., Goel, A., Moghadam, M.C., Blumenfeld, J.D., Chevalier, J.M., Shimonov, D., et al., 2022. Deep learning automation of kidney, liver, and spleen segmentation for organ volume measurements in autosomal dominant polycystic kidney disease. Tomography 8, 1804–1819.
- [43] Shi, G., Xiao, L., Chen, Y., Zhou, S.K., 2021. Marginal loss and exclusion loss for partially supervised multi-organ segmentation. Medical Image Analysis 70, 101979.
- [44] Simpson, A.L., Antonelli, M., Bakas, S., Bilello, M., Farahani, K., Van Ginneken, B., Kopp-Schneider, A., Landman, B.A., Litjens, G., Menze, B., et al., 2019. A large annotated medical image dataset for the development and evaluation of segmentation algorithms. arXiv preprint arXiv:1902.09063.
- [45] Taleb, A., Loetzsch, W., Danz, N., Severin, J., Gaertner, T., Bergner, B., Lippert, C., 2020. 3d self-supervised methods for medical imaging. Advances in Neural Information Processing Systems 33, 18158–18172.
- [46] Tang, Y., Yang, D., Li, W., Roth, H.R., Landman, B., Xu, D., Nath, V., Hatamizadeh, A., 2022. 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, pp. 20730–20740.
- [47] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. Advances in neural information processing systems 30.
- [48] Wang, F., Zhou, Y., Wang, S., Vardhanabhuti, V., Yu, L., 2022a. Multi-granularity cross-modal alignment for generalized medical visual representation learning. Advances in Neural Information Processing Systems 35, 33536–33549.
- [49] Wang, Y., Zhou, Y., Shen, W., Park, S., Fishman, E.K., Yuille, A.L., 2019. Abdominal multi-organ segmentation with organ-attention networks and statistical fusion. Medical image analysis 55, 88–102.
- [50] Wang, Z., Wu, Z., Agarwal, D., Sun, J., 2022b. Medclip: Contrastive learning from unpaired medical images and text. arXiv preprint arXiv:2210.10163.
- [51] Wang, Z., Zhang, C., Jiao, T., Gao, M., Zou, G., 2018. Fully automatic segmentation and three-dimensional reconstruction of the liver in ct images. Journal of Healthcare Engineering 2018.

- [52] Xie, Y., Zhang, J., Liu, L., Wang, H., Ye, Y., Verjans, J., Xia, Y., 2024. Refs: A hybrid pre-training paradigm for 3d medical image segmentation. Medical Image Analysis 91, 103023.
- [53] Xie, Y., Zhang, J., Xia, Y., Shen, C., 2023. Learning from partially labeled data for multi-organ and tumor segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence.
- [54] Xu, J., De Mello, S., Liu, S., Byeon, W., Breuel, T., Kautz, J., Wang, X., 2022. Groupvit: Semantic segmentation emerges from text supervision, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 18134–18144.
- [55] Xu, M., Zhang, Z., Wei, F., Lin, Y., Cao, Y., Hu, H., Bai, X., 2021. A simple baseline for zero-shot semantic segmentation with pre-trained vision-language model. arXiv preprint arXiv:2112.14757.
- [56] Zhang, J., Xie, Y., Xia, Y., Shen, C., 2021. Dodnet: Learning to segment multi-organ and tumors from multiple partially labeled datasets, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1195–1204.
- [57] Zhang, L., Zhang, J., Shen, P., Zhu, G., Li, P., Lu, X., Zhang, H., Shah, S.A., Bennamoun, M., 2020. Block level skip connections across cascaded v-net for multi-organ segmentation. IEEE Transactions on Medical Imaging 39, 2782–2793.
- [58] Zhang, Y., Huang, X., Ma, J., Li, Z., Luo, Z., Xie, Y., Qin, Y., Luo, T., Li, Y., Liu, S., et al., 2023. Recognize anything: A strong image tagging model. arXiv preprint arXiv:2306.03514.
- [59] Zhang, Y., Jiang, H., Miura, Y., Manning, C.D., Langlotz, C.P., 2022. Contrastive learning of medical visual representations from paired images and text, in: Machine Learning for Healthcare Conference, PMLR, pp. 2–25.
- [60] Zhou, H.Y., Chen, X., Zhang, Y., Luo, R., Wang, L., Yu, Y., 2022a. Generalized radiograph representation learning via cross-supervision between images and free-text radiology reports. Nature Machine Intelligence 4, 32–40.
- [61] Zhou, H.Y., Guo, J., Zhang, Y., Yu, L., Wang, L., Yu, Y., 2021. nnformer: Interleaved transformer for volumetric segmentation. arXiv preprint arXiv:2109.03201.
- [62] Zhou, K., Yang, J., Loy, C.C., Liu, Z., 2022b. Learning to prompt for vision-language models. International Journal of Computer Vision 130, 2337–2348.
- [63] Zhu, J., Li, Y., Hu, Y., Ma, K., Zhou, S.K., Zheng, Y., 2020. Rubik's cube+: A self-supervised feature learning framework for 3d medical image analysis. Medical image analysis 64, 101746.
- [64] Zhuang, X., Shen, J., 2016. Multi-scale patch and multi-modality atlases for whole heart segmentation of mri. Medical image analysis 31, 77–87.