ZePT: Zero-Shot Pan-Tumor Segmentation via Query-Disentangling and Self-Prompting

Yankai Jiang[1*], Zhongzhen Huang[1,2*], Rongzhao Zhang[1],
Xiaofan Zhang[1,2✉], Shaoting Zhang[1,3✉]
[1]Shanghai AI Laboratory [2]Shanghai Jiao Tong University [3]SenseTime Research
jiangyankai@pjlab.org.cn, huangzhongzhen@sjtu.edu.cn, zhangrongzhao@pjlab.org.cn,
xiaofan.zhang@sjtu.edu.cn, zhangshaoting@pjlab.org.cn

Abstract

*The long-tailed distribution problem in medical image analysis reflects a high prevalence of common conditions and a low prevalence of rare ones, which poses a significant challenge in developing a unified model capable of identifying rare or novel tumor categories not encountered during training. In this paper, we propose a new **Ze**ro-shot **P**an-**T**umor segmentation framework (ZePT) based on query-disentangling and self-prompting to segment unseen tumor categories beyond the training set. ZePT disentangles the object queries into two subsets and trains them in two stages. Initially, it learns a set of fundamental queries for organ segmentation through an object-aware feature grouping strategy, which gathers organ-level visual features. Subsequently, it refines the other set of advanced queries that focus on the auto-generated visual prompts for unseen tumor segmentation. Moreover, we introduce query-knowledge alignment at the feature level to enhance each query's discriminative representation and generalizability. Extensive experiments on various tumor segmentation tasks demonstrate the performance superiority of ZePT, which surpasses the previous counterparts and evidences the promising ability for zero-shot tumor segmentation in real-world settings.*

1. Introduction

A key challenge in medical image analysis stems from the long-tailed distribution problem, characterized by heavily imbalanced datasets where a few common cases coexist with many rare diseases [73] (Fig. 1 (a)). Most existing methods trained on specific-purpose datasets solely focus on a narrow scope of organs or tumors [14, 15, 21, 28, 49, 54, 74]. Recently, some studies [8, 36] attempted to design general-purposed methods that can handle various organs and tumors with a unified model. However, these models require large amounts of labeled training data and still have difficulty in identifying rare or new lesion categories that are clinically relevant. Obtaining gold-standard annotation for every tumor category from clinical experts can be highly expensive due to labor-intensive manual efforts, complex annotation processes [22], and may incur privacy concerns. In such a scenario, a zero-shot segmentation approach is highly desired, where the model can automatically segment unseen diseases without prior exposure to annotated cases during training. Therefore, we aim to explore the potential of zero-shot segmentation in developing a general-purpose medical image segmentor, as illustrated in Fig. 1 (b).

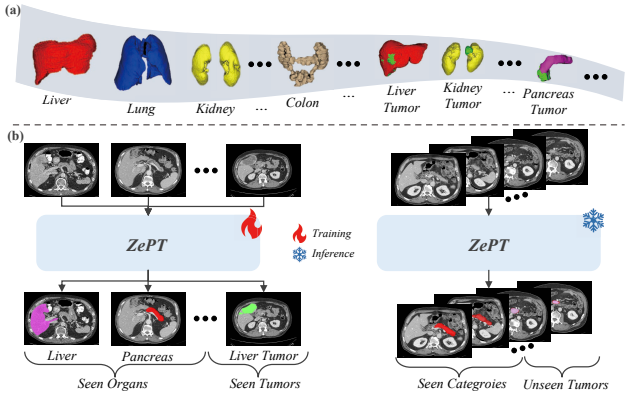


Figure 1. (a) The long-tailed distribution issue in medical image analysis. (b) ZePT is trained on datasets containing multiple organs and tumors. During inference, ZePT can segment both seen categories (*i.e.* organs and tumors) and unseen tumors.

The zero-shot segmentation [6] paradigm has been widely studied in the general image processing field [3, 18, 64, 71], which replaces the learnable weights of the classifier with fixed class semantic embeddings [12, 30, 43, 46] to transfer knowledge from seen (base) categories to unseen

*Equal contribution. ✉Corresponding authors. This work was supported by Shanghai Artificial Intelligence Laboratory. Codes are available at https://github.com/Yankai96/ZePT.

(novel) ones. Nevertheless, the performance of these methods is bottlenecked due to the absence of necessary knowledge about the novel classes [61, 76]. In more recent developments, open-vocabulary semantic segmentation (OVSS) techniques [13, 17, 35, 48, 67] utilize vision-language models (VLMs), *e.g.*, CLIP [50], to significantly enhance the accuracy of zero-shot segmentation. Some of them further utilize object queries from MaskFormer [9, 10] trained on base categories to produce class-agnostic mask proposals and then classify proposals with VLMs, demonstrating strong and robust zero-shot segmentation capabilities.

Although OVSS methods have achieved success in segmenting novel categories, their performance heavily depends on the quality of the generated proposals. Our pilot experiments in Tab. 1 indicate a primary challenge in applying the conventional OVSS strategy to tumor segmentation tasks: the vision clues of the semantics of tumors are usually more subtle and ambiguous than most common-life objects in natural images, which, as a consequence, makes it difficult to generate high-quality proposals for the unseen tumor categories. Therefore, the commonly adopted assumption in most OVSS methods [13, 17, 35, 48, 67] that the generated proposals cover almost all the potential object-of-interest no longer holds in the scenario of tumor segmentation, where a considerable number of tumors of unseen categories may be poorly covered.

Driven by the aforementioned limitations, we present a novel framework named ZePT for zero-shot tumor segmentation. ZePT adopts a query-disentangling scheme that partitions the object queries into two distinct subsets: fundamental queries and advanced queries. Then we decouple the learning process into two stages, allowing the model to first understand comprehensive organ anatomies and then focus on tumor segmentation, analogous to the learning process of human radiologists. In Stage-I, we pretrain the fundamental queries via object-aware feature grouping to acquire organ-level semantics for precise organ segmentation. In Stage-II, we train advanced queries, guided by self-generated visual prompts emerging from the fundamental queries, to concentrate on the subtle visual cues associated with tumors. Through the query-disentangling and self-prompting, ZePT captures fine-grained visual features associated with pathological changes and generates high-quality proposals that precisely cover unseen tumors. At last, we introduce cross-modal alignment between automatically sourced medical domain knowledge and query embeddings to provide weak supervision and augment the model with additional high-level semantic information, further enhancing the model's generalizability to unseen tumors.

In our experiments, we train ZePT using an assembly of 10 public benchmarks. We measure the tumor segmentation performance on MSD dataset [2] and a curated real-world dataset in a zero-shot manner. ZePT shows robust segmentation performance across four unseen tumor categories, significantly outperforming the previous leading methods by an average improvement of 15.85% in DSC, 17.43% in AUROC, and 23.27% in FPR_{95}. Meanwhile, ZePT also improves the segmentation performance of seen organs and tumors by at least absolute 4.83% in DSC on BTCV [34], 4.51% in DSC per case score on LiTS [5], 2.21% in DSC on KiTS [20], compared with the strong baseline nnUNet [23] and Swin UNETR [56].

Our main contributions can be summarized as follows:

- We propose ZePT, a novel two-stage framework with a query-disentangling scheme tailored for zero-shot tumor segmentation.
- We formulate tumor segmentation as a unique self-prompting process to localize unseen tumors.
- ZePT performs feature-level alignment between object queries and medical domain knowledge, further enhancing its generalizability to unseen tumors.
- ZePT consistently outperforms SOTA counterpart methods on multiple segmentation tasks, showing its effectiveness and robustness.

2. Related Work

Multi-Organ and Tumor Segmentation. The advancement of innovative model architectures [16, 23, 69, 72] and learning strategies [26, 29, 55, 56, 65, 75] has significantly propelled the field of automatic multi-organ segmentation, allowing it to achieve expert-level performance. Despite this significant progress, pan-tumor segmentation persistently presents a challenge. Existing efforts are usually specialized for single tumors [19, 21, 27, 28, 49, 68]. Some latest attempts are dedicated to training a universal model for segmenting various organs and tumors [8, 36]. In addition, a growing trend is emerging in efforts [7, 11, 39, 59, 62] to transfer the capabilities of SAM [33] to segment the abdominal organs and specific tumors. Differently, ZePT takes one step further by investigating a model that is capable of segmenting tumors from multiple organs in a zero-shot manner.

Open-Vocabulary Semantic Segmentation. The emerging concept of OVSS defines a generalized zero-shot semantic segmentation paradigm that allows a model to be trained on conventional vision datasets with close-set labels while possessing the ability to segment an image into arbitrary semantic regions according to text descriptions [13, 17, 35, 48, 67]. Nevertheless, as pointed out in [35], the mask proposals in OVSS methods are not truly "class-agnostic". They tend to overfit to seen categories and fail to cover previously unseen obscure objects. This issue hinders the transfer of the power of OVSS to zero-shot tumor segmentation on medical images. Differently, ZePT disentangles the object queries into two sets and adopts a self-prompting strategy to guide the model to explicitly learn semantics related to unseen (novel) tumor categories.

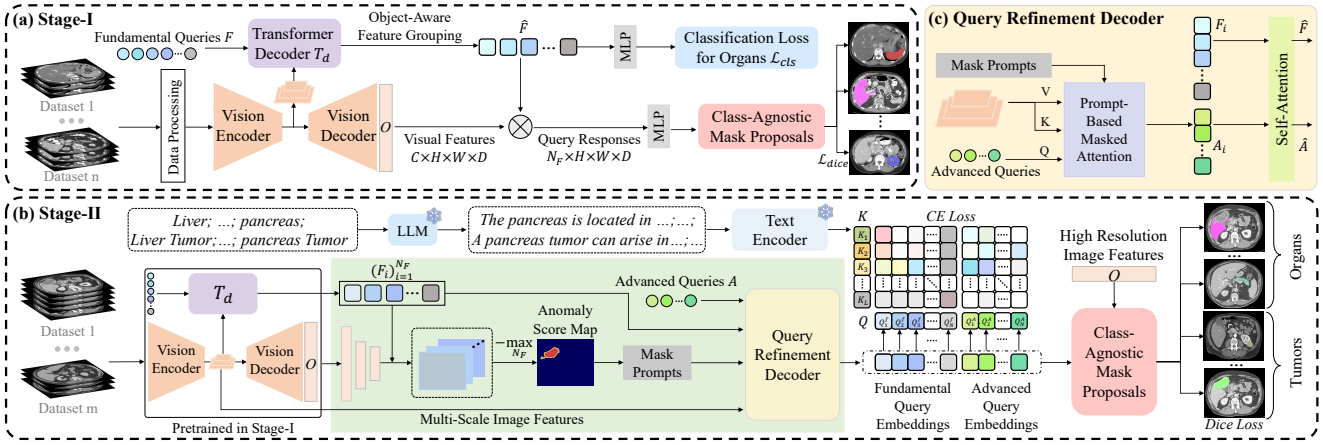


Figure 2. Overall pipeline. **Stage-I**: Based on MaskFormer [9, 10], we propose an object-aware feature grouping strategy to train a set of fundamental queries for multi-organ segmentation. **Stage-II**: A set of advanced queries for tumor segmentation attend to visual prompts derived from the affinity between fundamental query embeddings and visual features which indicates the presence of unseen abnormalities. Finally, we incorporate medical domain knowledge to better align text embeddings with query embeddings for cross-modal reasoning.

Unseen Lesion Detection and Segmentation. Some research efforts [47, 53, 57, 70, 77] also explored innovative approaches to detect or segment unseen lesions/tumors, which formulated the task as out-of-distribution (OOD) detection. Although unseen tumors can be regarded as a kind of OOD sample, the zero-shot segmentation task in this paper is largely different from OOD detection from the perspective that our main objective is to segment and classify multi-class tumors in a zero-shot manner with text descriptions, while OOD detection only recognizes unseen tumors as one single outlier class.

3. Method

As illustrated of Fig. 2, ZePT differs significantly from existing OVSS approaches [13, 17, 35, 48, 67] in that it disentangles the learning process of seen organs and unseen tumors into two stages and partitions object queries into two distinct subsets: fundamental queries and advanced queries. Specifically, Stage-I aims at pretraining fundamental queries on multiple datasets containing only organ labels to attain high-quality organ segmentation capability. In Stage-II, we formulate tumor segmentation as a self-prompting process, where advanced queries attend to the visual prompts derived from the affinity between embeddings of fundamental queries and visual features for capturing critical fine-grained context information and pathological changes related to tumors. We elaborate on the details of our designs in the following.

3.1. Stage-I: Fundamental Queries for Organs

We build our model upon a MaskFormer [9, 10]. As shown in Fig. 2 (a), the segmentation backbone consists of three components. A vision encoder V_e that extracts multi-scale visual features $V = \{V_i\}_{i=1}^{4}$, $V_i \in \mathbb{R}^{H_i\times W_i\times D_i\times C_i}$ from 3D volumes. Here, H_i, W_i, D_i and C_i denote the height, width, depth and channel dimension of V_i, respectively. A transformer decoder T_d that updates a set of N_F learnable fundamental queries $F \in \mathbb{R}^{N_F\times C}$ by interacting them with multi-scale visual features. A vision decoder V_d that gradually upsamples visual features to high-resolution image embeddings $O \in \mathbb{R}^{H\times W\times D\times C}$.

To capture organ-level information and achieve object-aware cross-modal reasoning, we propose an object-aware feature grouping strategy in T_d to guide each learnable fundamental query to represent and specify a corresponding organ category. This is achieved by grouping visual features into the query embeddings for context-aware reasoning.

Specifically, T_d consists of a series of transformer blocks enabling the queries to interact with multi-scale features. In the i-th transformer block, the fundamental queries $F \in \mathbb{R}^{N_F\times C}$ first exploit global information from 3D image feature maps $V_i \in \mathbb{R}^{H_i\times W_i\times D_i\times C_i}$ via a classical cross-attention as follows:

$$\mathcal{Q} = W^q\delta(F), \mathcal{K} = W^k V_i, \mathcal{V} = W^v V_i \quad (1)$$

$$\hat{F}_i = \mathrm{MLP}(\mathrm{LayerNorm}(\mathrm{Softmax}\left(\frac{\mathcal{Q}\mathcal{K}^T}{\sqrt{d}}\right)\mathcal{V}), \quad (2)$$

where $W^q, W^k, W^v \in \mathbb{R}^{C_i\times C_i}$ are learnable projection matrices. δ is a linear projection.

Subsequently, we explicitly assign the relevant local context information from visual features into the fundamental queries based on affinity in the embedding space to ensure that different queries focus on different visual regions without overlaps. We first calculate an assignment similarity matrix $S_i \in \mathbb{R}^{N_F\times H_iW_iD_i}$ between the

fundamental queries $\hat{F}_i$ and the image features V_i via a Gumbel-Softmax [24, 42] operation:

$$S_i^{\text{gumbel}} = \text{Softmax}\left(\left(\hat{F}_i V_i^T + G\right)/\tau\right). \quad (3)$$

Here $G \in \mathbb{R}^{N_F \times H_i W_i D_i}$ are i.i.d random samples drawn from the $Gumbel(0,1)$ distribution and τ is a learnable coefficient to assist in finding a suitable assignment boundary.

We then group the visual features in V_i and corresponds the groups to the queries $\hat{F}_i$ by taking the one-hot operation of the argmax over S_i^{gumbel}:

$$S_i^{\text{onehot}} = \text{onehot}\left(\text{argmax}_{N_F}\left(S_{\text{gumbel}}\right)\right). \quad (4)$$

Since the straightforward hard assignment (*i.e.*, one-hot) via argmax is not differentiable, we adopt the straight through trick in [58, 66] to compute the assignment similarities $\hat{S}_i$ of one-hot value as follows:

$$\hat{S}_i = \left(S_i^{\text{onehot}}\right)^\top + S_i^{\text{gumbel}} - \text{sg}\left(S_i^{\text{gumbel}}\right), \quad (5)$$

where sg is the stop gradient operator.

With the above operations, the whole transfomrer block is differentiable and end-to-end trainable. $\hat{S}_i$ indicates the assignment of object-level visual features to each query. Finally, query embeddings $\hat{F}_i$ will be updated via being assigned with the most corresponding features in V_i according to $\hat{S}_i$, which can be denoted as follows:

$$\hat{F}_{i+1} = \text{MLP}\left(S_i V_i\right) + \hat{F}_i. \quad (6)$$

The binary mask proposals $BM \in [0,1]^{N_F \times H \times W \times D}$ for fundamental queries are obtained by a multiplication operation between the query embedding and high-resolution image features O followed by a Sigmoid. We adopt Dice Loss to supervise mask proposals with organ labels. We also process the query embeddings through a MLP layer to get class embeddings, which are then supervised using the category information of organs through a Cross-Entropy loss. As later shown in Fig. 4 (a), our learnable fundamental queries focus on distinct foreground organ regions and explicitly encourage the strict boundaries between different categories, preventing mixed representations where the target region and the disturbing regions are grouped together. Such discriminative representation also enhances the localization of unseen tumors, as discussed in Sec. 3.2.

3.2. Stage-II: Advanced Queries for Tumors

In Stage-II, we aim to refine a set of N_A advanced queries $A \in \mathbb{R}^{N_A \times C}$ for tumor segmentation and ensure their generalization to unseen tumors. The core insight is to endow the advanced queries with the ability to capture anomaly information of tumors by utilizing fundamental queries. We propose to reformulate tumor segmentation as a self-prompting process where the advanced queries can be aware of abnormal information related to pathological changes in the feature context via visual prompts. We retain the training datasets in Stage-I to avoid the forgetting problem of the fundamental queries and add several datasets containing tumor labels for Stage-II. It is worth noting that there are novel tumor categories in the testing phase, rendering our method truly "zero-shot".

Self-Generated Visual Prompts. As shown in Fig. 2 (b), we utilize the pretrained V_e, V_d, T_d, and fundamental queries F from Stage-I. The volumes are fed into the pretrained network to obtain the multi-scale visual features $V = \{V_i\}_{i=1}^4$, high-resolution image embeddings $O \in \mathbb{R}^{H \times W \times D \times C}$ and refined fundamental query embeddings $\hat{F} \in \mathbb{R}^{N_F \times C}$. We compute multi-scale query response maps $R_i \in \mathbb{R}^{N_F \times H_i \times W_i \times D_i}$ representing the affinity between visual features and different fundamental queries at each resolution stage. Then we adopt the negative of maximal operation [70] along channel dimension on R_i to generate multi-scale anomaly score $M_i \in \mathbf{R}^{H_i \times W_i \times D_i}$ maps:

$$R_i = \hat{F}_i V_i^T \quad (7)$$

$$M_i = \min_{c \in 1,\ldots,N_F} R_i^c, i \in [1,4]. \quad (8)$$

These anomaly score maps M_i can be further normalized into mask prompts $\hat{M}_i \in [0,1]^{H_i \times W_i \times D_i}$ by min-max normalization and a threshold of 0.5, where $\hat{M}_i^{(h,w,d)} = 1$ and $\hat{M}_i^{(h,w,d)} = 0$ represent that the voxel located at position (h,w,d) in the input 3D volume belongs to an anomalous unseen category and an in-distribution seen organ class, respectively. We use these mask prompts, adaptively derived from the embedding space, to assist the advanced queries to attend to the anomalous context features and learn representations that effectively localize unseen tumors.

Query Refinement Decoder. The Query Refinement Decoder (QRD) takes mask prompts, multi-scale visual features, a set of zero-initialized advanced queries, and embeddings of fundamental queries as inputs. As shown in Fig. 2 (c), the N_A advanced queries are designed to localize and identify unseen tumors on organs corresponding to the N_F fundamental queries. For the i-th block in QRD, the advanced queries A_i are first updated via interactions with multi-scale features V_i and mask prompts $\hat{M}_i$ via:

$$\hat{A}_i = A_i + \text{Softmax}(\mathcal{M}_i + \mathcal{Q}_{A_i}\mathcal{K}_{V_i}^T)\mathcal{V}_{V_i}{}^T, \quad (9)$$

where $\mathcal{Q}_{A_i} = f_Q(A_i) \in \mathbb{R}^{N_A \times C_i}$ denotes embeddings of advanced queries under transformation $f_Q(\cdot)$. $\mathcal{K}_{V_i}$, $\mathcal{V}_{V_i} \in \mathbb{R}^{C_i \times H_i W_i D_i}$ denote 3D image features under transformation $f_\mathcal{K}(\cdot)$ and $f_\mathcal{V}(\cdot)$, respectively. The visual prompt attention mask $\mathcal{M}_i$ at feature location (h,w,d) is defined as:

$$\mathcal{M}_i^{(h,w,d)} = \begin{cases} 0 & \text{if } \hat{M}_i^{(h,w,d)} = 1 \\ -\infty & \text{otherwise} \end{cases}. \quad (10)$$

We refer to this mechanism as **prompt-based masked attention** because it aggregates the visual information highlighted by given mask prompts, allowing advanced queries to concentrate on abnormal regions with pathological changes across various organs, thereby facilitating the detection of previously unseen tumors. Through a series of blocks, we can obtain the refined advanced queries $\hat{A}_i$. Then, QRD concatenates fundamental queries $\hat{F}_i$ with advanced queries $\hat{A}_i$ and performs self-attention among them to encode the relationship between fundamental and advanced queries, facilitating the adjustment of their representations to encourage a clear semantic distinction between organs and tumors. Finally, we use these updated queries to generate corresponding mask proposals as described in 3.1. Organ labels supervise the mask proposals derived from the fundamental queries, whereas tumor labels supervise those from advanced queries.

Query-Knowledge Alignment. To make the model better retain its generalization ability for recognizing unseen tumors, we introduce Query-Knowledge Alignment for the weakly-supervised cross-modal alignment between visual features of queries and high-level semantics of textual knowledge. Instead of a simple description (*e.g.* "a photo of ") in previous methods [13, 35, 36, 67], we utilize detailed knowledge of each class name by prompting GPT4 [44] with an instruction: "Please describe {CLS}. The answer should encompass attributes related to its location, shape, size, and anatomical structure.". Each piece of generated knowledge is checked and modified by professional doctors to ensure correctness. We adopt ClinicalBERT [1] as a pretrained text encoder to get the knowledge embeddings $\mathbf{K}$. The predicted probability distribution over the training classes for the i-th query is calculated as:

$$\mathcal{P}_i = \frac{\exp(\frac{1}{\tau}\zeta(\mathbf{K}_i, \mathbf{Q}_i))}{\sum_{j=0}^{|L|}\exp(\frac{1}{\tau}\zeta(\mathbf{K}_j, \mathbf{Q}_i))}, \tag{11}$$

where ζ is the cosine similarity between two embeddings, and τ is the temperature. $\mathbf{Q}$ is derived from query embeddings via a linear projection layer. During training, the similarities between matched query embedding and text embedding should be maximized. A cross-entropy loss is applied on $\mathcal{P}$ for supervision.

In summary, we combine the Dice loss on mask proposals and the cross-entropy loss for query-knowledge alignment to supervise the learning in Stage-II.

4. Experiments

Dataset Construction. (1) **Training:** In Stage-I, we assemble the training sets of 8 public datasets, including Pancreas-CT [52], AbdomenCT-1K [40], CT-ORG [51], CHAOS [32], AMOS22 [25], BTCV [34], WORD [38] and TotalSegmentator [60]. These datasets exclusively contained organ labels. In Stage-II, we add CT images from the training sets of LiTS [5] and KiTS [20]. The overall seen categories used for training consist of 25 organ classes and 2 tumor classes. (2) **Inference:** We employ the MSD dataset [2] that encompasses a range of segmentation tasks for five tumor types in CTs. Among these, pancreas tumors, lung tumors, colon tumors, and hepatic vessel tumors belong to unseen categories. A real-world, private dataset containing 388 3D CT volumes of four distinct colon tumor subtypes is also utilized for testing. We follow the data pre-processing in [36] to reduce the domain gap among various datasets. Due to page limits, details of all datasets and pre-processing are described in the supplemental material.

Evaluation Metrics. Dice Similarity Coefficient (DSC) is utilized for evaluating organ/tumor segmentation. We also report the area under the receptive-operative curve (AUROC) and the false positive rate at a true positive rate of 95% (FPR_{95}), which are commonly used in OOD detection methods [31, 63, 70]. For all the metrics above, 95% CIs were calculated and the p-value cutoff of less than 0.05 was used for defining statistical significance.

Implementation Details. (1) **Stage-I:** We use the current benchmark model in medical image segmentation, Swin UNETR [19], as the backbone, which consists of a vision encoder and a vision decoder with skip connections. We adopt four transformer decoder blocks in T_d, and each takes image features with output stride 32, 16, 8, and 4, respectively. We employ AdamW optimizer [37] with a warm-up cosine scheduler of 50 epochs. The batch size is set to 2 per GPU with a patch size of $96 \times 96 \times 96$. The training process uses an initial learning rate of $1e^{-4}$, momentum of 0.9 and decay of $1e^{-5}$ on multi-GPU (8) with DDP for 1000 epochs. Extensive data augmentation is utilized on-the-fly to improve the generalization, including random rotation and scaling, elastic deformation, additive brightness, and gamma scaling. The number of the fundamental queries N_F is 25 for 25 organ classes. The loss is the sum of Cross-Entropy loss and Dice loss. (2) **Stage-II:** We adopt the pretrained model in Stage-I. We set the initial learning rate as $4e^{-4}$. QRD has four blocks and each attention layer in the QRD block has eight heads. The number of the advanced queries N_A is 20 for tumors/diseases categories. Other settings are kept the same as in Step-I. We implement ZePT model in PyTorch [45]. All experiments are conducted on 8 NVIDIA A100 GPUs.

Baselines. In this paper, the zero-shot tumor segmentation setting requires that models directly segment unseen tumor types during inference without any fine-tuning or retraining. This is notably challenging as the model has to handle both unseen classes and domain gaps between different datasets. For unseen tumor segmentation, we compare ZePT with a series of representative OVSS methods, including ZegFormer [13], zsseg [67], OpenSeg [17], OVSeg [35],

Method	MSD Dataset												Real-World Colon Tumor Segmentation		
	Pancreas Tumor			Lung Tumor			Hepatic Vessel Tumor			Colon Tumor					
	AUROC↑	FPR_{95}↓	DSC↑	AUROC↑	FPR_{95}↓	DSC↑	AUROC↑	FPR_{95}↓	DSC↑	AUROC↑	FPR_{95}↓	DSC↑	AUROC↑	FPR_{95}↓	DSC↑
ZegFormer [13]	66.45	69.33	14.92	41.31	81.78	9.94	75.39	55.94	30.81	50.13	78.81	11.34	55.64	74.29	12.03
zsseg [67]	53.23	79.27	11.40	34.96	87.54	7.98	71.43	60.35	28.57	47.79	82.68	9.68	50.30	79.52	10.18
OpenSeg [17]	44.56	85.19	10.05	23.49	91.75	6.12	59.23	70.52	23.38	41.76	89.44	7.13	41.92	87.51	7.59
OVSeg [35]	70.22	59.73	19.36	52.93	68.65	14.11	85.77	40.28	35.66	59.94	65.25	15.76	69.95	64.84	16.05
FreeSeg [48]	69.98	60.75	18.19	49.92	70.39	13.26	85.62	41.77	35.08	56.45	68.49	14.71	67.01	66.07	15.30
SynthCP [63]	51.24	81.69	11.33	25.85	90.28	6.43	70.12	63.55	28.01	43.84	87.71	8.74	48.37	84.95	8.72
SML [31]	37.95	89.93	9.72	20.18	93.65	6.02	57.44	70.96	22.97	22.41	92.07	6.65	39.88	88.41	7.21
MaxQuery [70]	68.99	59.93	18.15	48.24	70.47	11.29	83.66	42.45	34.30	50.47	69.88	13.43	64.53	67.65	15.24
ZePT	**86.81**	**35.18**	**37.67**	**77.84**	**44.30**	**27.22**	**91.57**	**20.64**	**52.94**	**82.36**	**40.73**	**30.45**	**84.35**	**38.29**	**36.23**

Table 1. Detection and segmentation performance of unseen tumors on MSD [2] and real-world colon tumor dataset. ZePT achieves state-of-the-art unseen tumor detection and segmentation performance. More results can be found in the supplemental material.

and Freeseg [48]. We also compare ZePT with OOD detection methods [31, 63, 70], which treat unseen tumors as a single outlier class. We adopt the masked back-propagation in [36] to enable the training of these methods on partially labeled datasets. All baselines are trained with all datasets used in Stage-I and Stage-II. For seen organ/tumor segmentation, we compare ZePT with SOTA benchmark models, including nnUNet [23], Swin UNETR [56] and Universal model [36].

4.1. Main Results

Unseen Tumor Segmentation on MSD Dataset. Tab. 1 shows the segmentation performance of four unseen tumor categoreis from MSD [2]. All available volumes in these four tumor segmentation tasks are directly used for testing. Compared with SOTA OVSS methods, ZePT demonstrates a notable performance enhancement in the context of average unseen tumor localization performance across four tasks, achieving at least a 17.43% improvement in AUROC and a 23.27% increase in FPR_{95}. Regarding the average performance in unseen tumor segmentation across these tasks, ZePT continues to maintain a substantial lead, as evidenced by a notable 15.85% improvement in the DSC. These results indicate that the proposed query-disentangling and self-prompting can effectively help the model capture visual cues related to tumors, thus boosting the ability to recognize unseen ones. Moreover, OVSS methods require an additional frozen CLIP vision encoder to classify each mask proposal, leading to slower inference speeds. In contrast, ZePT removes this process and adopts query-knowledge alignment at the feature level, which maintains a reasonable computation cost. As shown in Tab. 2, ZePT has fewer network parameters and approximately 23% of the FLOPs compared to previous OVSS methods.

We also compare ZePT with SOTA OOD detection methods. ZePT's average performance surpasses the previously best-performing MaxQuery [70] across four tasks by 21.81% in AUROC, 25.47% in FPR_{95}, and 17.78% in DSC. ZePT aligns visual features with linguistic semantics for cross-modal interaction instead of solely exploiting information from visual modality like most OOD region segmentation methods. The results suggest that leveraging features from images together with medical domain knowledge benefits the semantic understanding of unseen tumors.

Efficiency \ Method	ZePT	ZegFormer [13]	OVSeg [35]	FreeSeg [48]
Params	**745.94**M	950.82M	963.44M	1077.85M
FLOPs	**1337.59**G	5766.21G	5929.65G	6893.14G

Table 2. Computational cost comparison between ZePT and current OVSS methods. The FLOPs is computed based on input with spatial size $96 \times 96 \times 96$ on the same single A100 GPU.

Method	BTCV	LiTS	KiTS
nnUNet [23]	82.23±2.07	77.15±3.47	85.18±1.26
Swin UNETR [56]	82.26±2.02	76.79±3.52	85.52±1.13
Universal [36]	86.38±1.61	80.58±3.03	87.05±1.04
ZePT	**87.09±1.54**	**81.66±2.79**	**87.73±0.99**

Table 3. 5-fold cross-validation results on the BTCV [34], LiTS [5], and KiTS [20] validation dataset. We report the average DSC of 13 organs in BTCV, the Dice per case score of liver tumors in LiTS, and the DSC of kidney tumors in KiTS.

Real-World Colon Tumor Segmentation. We further conduct a zero-shot evaluation on a real-world colon tumor dataset. The average results of four colon tumor subtypes are also summarized in Tab. 1. ZePT outperforms the baselines by at least absolute 14.40% in AUROC, 26.55% in FPR_{95}, and 20.18% in DSC, demonstrating much better generalizability and robustness. ZePT reaches a much lower FPR_{95} compared with previous methods, which is crucial for safety-critical medical scenarios. The results indicate ZePT has a strong potential for utility in clinical practice.

Segmentation of Seen Organs and Tumors. As shown in Tab. 3, the segmentation performance of ZePT on seen organs surpasses strong baseline nnUNet [23] and Swin UNETR [56] by at least absolute 4.83% in DSC on BTCV [34], 4.51% in DSC for liver tumors on LiTS [5], and 2.21% in DSC for kidney tumors on KiTS [20]. Notably, ZePT achieves comparable or even better segmentation performance compared with the Universal model [36], which utilizes an assembly of 14 public datasets with a total of 3,410

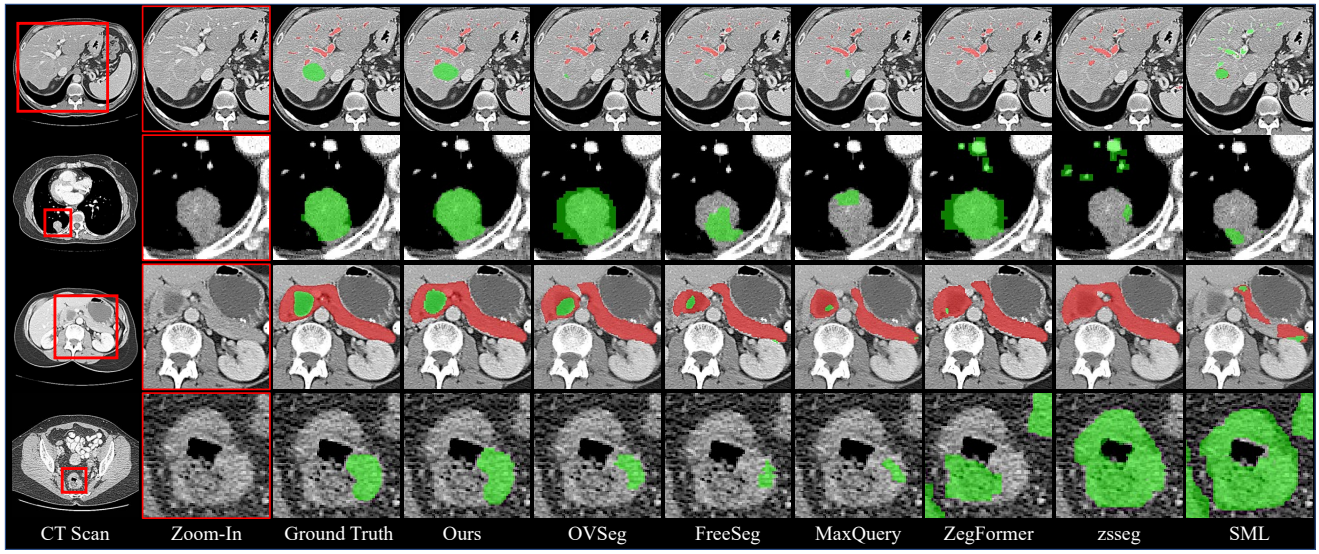


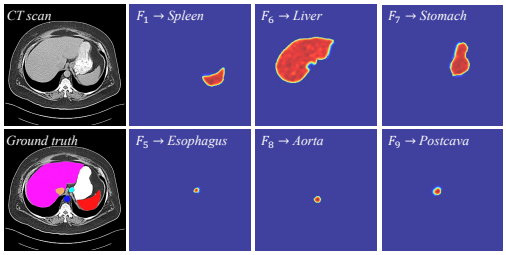
Figure 3. Qualitative visualizations on MSD [2] dataset. We compare ZePT with other advanced OVSS methods and OOD detection methods in a zero-shot manner. The segmentation results presented from rows one to four correspond, in order, to hepatic vessel tumors, lung tumors, pancreatic tumors, and colorectal tumors. We present the visualizations on other datasets in the supplemental material.

CT scans for training. The Universal model [36] adopts a CLIP text encoder, which processes the organ and tumor names into text embeddings, to introduce the semantic relationship between anatomical structures. ZePT takes one step further by designing a more advanced architecture for visual feature extraction and incorporating additional medical domain knowledge to polish feature representations with diverse and fine-grained cues. These improvements demonstrate that ZePT can also segment seen organs and tumors with high accuracy.

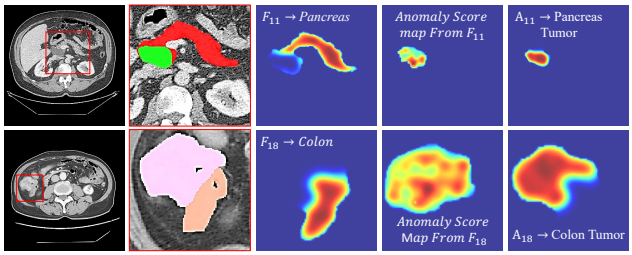
Qualitative Analysis. Fig. 3 shows the qualitative results and demonstrates the merits of ZePT. Most competing methods suffer from segmentation target incompleteness-related failures and misclassification of background regions as tumors (false positives). ZePT produces sharper boundaries and generates results that are more consistent with the ground truth in comparison with all other models.

We also visualize the query response maps of different seen organs and unseen tumors, as well as the anomaly score maps to illustrate the working mechanism of fundamental queries and advanced queries in ZePT. As shown in Fig. 4 (a), the different organ regions are confidently activated by fundamental queries F_i (F_1 for spleen, F_5 for esophagus, F_6 for liver, *etc*.). This advantage is attributed to the object-aware feature grouping, which enables each fundamental query to represent a corresponding organ.

In Row 1 of Fig. 4 (b), we can observe the pancreas region is captured by the fundamental query F_{11}, and the anomaly score maps derived from F_{11} maintains high responses around the pancreas tumor region to indicate our model the potential location of unseen tumors. Furthermore, the advanced queries A_{11} capture the region corresponding to the pancreas tumor according to the guidance of mask prompts derived from anomaly score maps. The same phenomenon can also be observed in the example of



(a) A testing example containing seen categories (From BTCV)



(b) Testing examples containing unseen tumors

Figure 4. Visualization of query response maps. (a) A test sample containing seen categories from BTCV [34] evaluation set. (b) Two test samples, one from the MSD's pancreas tumor task [2] and the other from the real-world colon tumor segmentation dataset. We can observe the query distribution on the different organs and tumors with obvious separation. The clear boundaries and high responses show the advantages of encouraging discriminative and disentangled queries to represent different objects, which benefits the segmentation of both seen and unseen categories.

colon tumor segmentation in Row 2 of Fig. 4 (b). These visual examples fulfill the motivation of the self-prompting strategy that uses the anomaly score maps derived from fundamental queries as visual prompts to guide the advanced queries to segment unseen tumors.

4.2. Ablation Study and Discussions

Significance of Object-Aware Feature Grouping. We remove the object-aware feature grouping (OFG) and use the vanilla MaskFormer [10] to update the fundamental queries. We denote this variant as ZePT (w/o OFG) in Tab. 4, which is observed to have a performance drop on both seen and unseen categories. This shows that OFG is crucial in learning object queries for region-level information.

Importance of Query Refinement Decoder. We replace the Query Refinement Decoder (QRD) with a vanilla transformer decoder to directly update all object queries together. It should be noted that this operation also eliminates the query-disentangling scheme since object queries are entangled into a single set and directly matched across all categories. We denote this variant as ZePT (w/o QRD) in Tab. 4, which is observed to have a significant performance drop of 11.89% in AUROC and 10.39% in DSC for unseen colon tumor segmentation. This confirms the efficacy of QRD, designed to use visual prompts for updating advanced queries and to introduce interactions between the two sets of queries. This design enables both query sets to modify their representations for enhanced performance.

Effectiveness of Prompt-Based Masked Attention. As illustrated in Tab. 4, we replace the prompt-based masked attention with a vanilla self-attention layer and directly concatenate the visual prompts (anomaly score maps) with image features to update the advanced queries. We denote this variant as ZePT (w/o PMA). This variant leads to decreased performance on both seen and unseen tumors, verifying that prompt-based masked attention is an effective way to leverage the visual prompts.

Efficacy of the Query-Knowledge Alignment. We remove the query-knowledge alignment and employ the CLIP image encoder to classify mask proposals in the same manner as conventional OVSS methods. We denote this variant as ZePT (w/o QKA) in Tab. 4, which is observed to have a performance drop of 10.13% in AUROC and 5.19% in DSC on unseen colon tumors. Compared with an additional tuned CLIP image encoder, our feature-level cross-modal alignment between queries and domain knowledge directly and explicitly introduces high-level linguistic semantics into the visual representation, which is more effective and efficient.

Investigation on Self-Generated Visual Prompts. Recently, many methods based on SAM [33] leverage visual prompts for medical image segmentation [7, 39, 62]. ZePT, using self-generated prompts adaptively derived from the embedding space, consistently outperforms these SAM-based methods, which adopt bounding boxes or points provided by the users as prompts (Tab. 5). Moreover, our prompt-based masked attention can also handle box prompts. We observe that self-generated visual prompts can match or even surpass the performance of strong manual prompts, like relaxed boxes, a finding corroborated by our visualization of anomaly score maps in Fig. 4 (b). Such flexible and adaptive self-generated prompts are crucial for unseen tumor segmentation scenarios, where acquiring even bounding box prompts is challenging.

Discussions about Limitations. The zero-shot segmentation performance of ZePT on unseen tumors still falls behind supervised models fully trained on these tumors. Although the comparison is unfair, it indicates there is still much room for improvement. We hope our work can shed some light on designing models with zero-shot abilities for medical imaging tasks.

Method	Liver Tumor in LiTS (Seen)		Real-World Colon Tumor (Unseen)		
	AUROC↑	DSC↑	AUROC↑	FPR_{95}↓	DSC↑
ZePT (w/o OFG)	95.96±0.85	80.84±3.17	75.39	48.24	31.28
ZePT (w/o QRD)	94.53±1.75	80.11±3.98	72.46	58.93	25.84
ZePT (w/o PMA)	96.25±0.62	81.00±3.01	78.01	44.35	33.06
ZePT (w/o QKA)	95.04±1.09	80.61±3.34	74.22	50.31	31.04
ZePT	**96.82±0.47**	**81.66±2.79**	**84.35**	**38.29**	**36.23**

Table 4. Ablation study of different network components on LiTS and the real-world colon tumor segmentation dataset.

Methods with Different Prompts	DSC↑	NSD↑
MedSAM [39] (relaxed 3d bbx prompt)	26.18	36.25
MSA [62] (1 point prompt)	27.88	39.06
MA-SAM [7] (relaxed 3d bbx prompt)	29.39	41.11
ZePT (relaxed 3d bbx prompt)	35.89	48.04
ZePT	**36.23**	48.78

Table 5. Comparisons between ZePT and SAM-based [33] medical image segmentation methods [7, 39, 62] on real-world colon tumor segmentation dataset. We report DSC and Normalized Surface Distance (NSD).

5. Conclusions

In this work, we propose ZePT, a novel framework based on query-disentangling and self-prompting for zero-shot pan-tumor segmentation. We disentangle the object queries into two subsets and decouple their learning process into two stages. ZePT exploits discriminative and object-level feature representation for organs and tumors. We introduce a self-prompting strategy to adaptively localize abnormalities for guiding the queries to be aware of the pathological changes among visual contexts. Additionally, we perform query-knowledge alignment at the feature level to further enhance the model's generalization capabilities. The significant performance improvements of ZePT on various organ and tumor segmentation tasks validate its effectiveness.

References

[1] Emily Alsentzer, John R Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. Publicly available clinical bert embeddings. *arXiv preprint arXiv:1904.03323*, 2019. 5, 2, 4

[2] 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. 2, 5, 6, 7, 1

[3] Donghyeon Baek, Youngmin Oh, and Bumsub Ham. Exploiting a joint embedding space for generalized zero-shot semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9536–9545, 2021. 1

[4] Cheng Bian, Chenglang Yuan, Kai Ma, Shuang Yu, Dong Wei, and Yefeng Zheng. Domain adaptation meets zero-shot learning: an annotation-efficient approach to multi-modality medical image segmentation. *IEEE Transactions on Medical Imaging*, 41(5):1043–1056, 2021. 8

[5] Patrick Bilic, Patrick Christ, Hongwei Bran Li, Eugene Vorontsov, Avi Ben-Cohen, Georgios Kaissis, Adi Szeskin, Colin Jacobs, Gabriel Efrain Humpire Mamani, Gabriel Chartrand, et al. The liver tumor segmentation benchmark (lits). *Medical Image Analysis*, 84:102680, 2023. 2, 5, 6, 1

[6] Maxime Bucher, Tuan-Hung Vu, Matthieu Cord, and Patrick Pérez. Zero-shot semantic segmentation. *Advances in Neural Information Processing Systems*, 32, 2019. 1

[7] Cheng Chen, Juzheng Miao, Dufan Wu, Zhiling Yan, Sekeun Kim, Jiang Hu, Aoxiao Zhong, Zhengliang Liu, Lichao Sun, Xiang Li, et al. Ma-sam: Modality-agnostic sam adaptation for 3d medical image segmentation. *arXiv preprint arXiv:2309.08842*, 2023. 2, 8

[8] Jieneng Chen, Yingda Xia, Jiawen Yao, Ke Yan, Jianpeng Zhang, Le Lu, Fakai Wang, Bo Zhou, Mingyan Qiu, Qihang Yu, et al. Towards a single unified model for effective detection, segmentation, and diagnosis of eight major cancers using a large collection of ct scans. *arXiv preprint arXiv:2301.12291*, 2023. 1, 2, 8

[9] Bowen Cheng, Alex Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. *Advances in Neural Information Processing Systems*, 34:17864–17875, 2021. 2, 3, 5

[10] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1290–1299, 2022. 2, 3, 8, 4, 5

[11] Junlong Cheng, Jin Ye, Zhongying Deng, Jianpin Chen, Tianbin Li, Haoyu Wang, Yanzhou Su, Ziyan Huang, Jilong Chen, Lei Jiang, et al. Sam-med2d. *arXiv preprint arXiv:2308.16184*, 2023. 2

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 1

[13] Jian Ding, Nan Xue, Gui-Song Xia, and Dengxin Dai. Decoupling zero-shot semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11583–11592, 2022. 2, 3, 5, 6

[14] Yuhang Ding, Xin Yu, and Yi Yang. Rfnet: Region-aware fusion network for incomplete multi-modal brain tumor segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3975–3984, 2021. 1

[15] Deng-Ping Fan, Tao Zhou, Ge-Peng Ji, Yi Zhou, Geng Chen, Huazhu Fu, Jianbing Shen, and Ling Shao. Inf-net: Automatic covid-19 lung infection segmentation from ct images. *IEEE transactions on medical imaging*, 39(8):2626–2637, 2020. 1

[16] Xi Fang and Pingkun Yan. Multi-organ segmentation over partially labeled datasets with multi-scale feature abstraction. *IEEE Transactions on Medical Imaging*, 39(11):3619–3629, 2020. 2

[17] Golnaz Ghiasi, Xiuye Gu, Yin Cui, and Tsung-Yi Lin. Scaling open-vocabulary image segmentation with image-level labels. In *European Conference on Computer Vision*, pages 540–557. Springer, 2022. 2, 3, 5, 6

[18] Zhangxuan Gu, Siyuan Zhou, Li Niu, Zihan Zhao, and Liqing Zhang. Context-aware feature generation for zero-shot semantic segmentation. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1921–1929, 2020. 1

[19] Ali Hatamizadeh, Vishwesh Nath, Yucheng Tang, Dong Yang, Holger R Roth, and Daguang Xu. Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In *International MICCAI Brainlesion Workshop*, pages 272–284. Springer, 2021. 2, 5

[20] Nicholas Heller, Sean McSweeney, Matthew Thomas Peterson, Sarah Peterson, Jack Rickman, Bethany Stai, Resha Tejpaul, Makinna Oestreich, Paul Blake, Joel Rosenberg, et al. An international challenge to use artificial intelligence to define the state-of-the-art in kidney and kidney tumor segmentation in ct imaging., 2020. 2, 5, 6, 1

[21] Nicholas Heller, Fabian Isensee, Klaus H Maier-Hein, Xiaoshuai Hou, Chunmei Xie, Fengyi Li, Yang Nan, Guangrui Mu, Zhiyong Lin, Miofei Han, et al. The state of the art in kidney and kidney tumor segmentation in contrast-enhanced ct imaging: Results of the kits19 challenge. *Medical image analysis*, 67:101821, 2021. 1, 2

[22] Ahmed Hosny, Chintan Parmar, John Quackenbush, Lawrence H Schwartz, and Hugo JWL Aerts. Artificial intelligence in radiology. *Nature Reviews Cancer*, 18(8):500–510, 2018. 1

[23] Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 18(2):203–211, 2021. 2, 6, 7

[24] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. In *ICLR*, 2017. 4

[25] Yuanfeng Ji, Haotian Bai, Chongjian Ge, Jie Yang, Ye Zhu, Ruimao Zhang, Zhen Li, Lingyan Zhanng, Wanling Ma, Xiang Wan, et al. Amos: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation.

Advances in Neural Information Processing Systems, 35: 36722–36732, 2022. 5, 1, 2

[26] Zhanghexuan Ji, Dazhou Guo, Puyang Wang, Ke Yan, Jia Ge, Xianghua Ye, Minfeng Xu, Jingren Zhou, Le Lu, Mingchen Gao, et al. Continual segment: Towards a single, unified and accessible continual segmentation model of 143 whole-body organs in ct scans. *arXiv preprint arXiv:2302.00162*, 2023. 2

[27] Jue Jiang, Yu-Chi Hu, Chia-Ju Liu, Darragh Halpenny, Matthew D Hellmann, Joseph O Deasy, Gig Mageras, and Harini Veeraraghavan. Multiple resolution residually connected feature streams for automatic lung tumor segmentation from ct images. *IEEE transactions on medical imaging*, 38(1):134–144, 2018. 2

[28] Yankai Jiang, Shufeng Xu, Hongjie Fan, Jiahong Qian, Weizhi Luo, Shihui Zhen, Yubo Tao, Jihong Sun, and Hai Lin. Ala-net: Adaptive lesion-aware attention network for 3d colorectal tumor segmentation. *IEEE transactions on medical imaging*, 40(12):3627–3640, 2021. 1, 2

[29] Yankai Jiang, Mingze Sun, Heng Guo, Xiaoyu Bai, Ke Yan, Le Lu, and Minfeng Xu. Anatomical invariance modeling and semantic alignment for self-supervised learning in 3d medical image analysis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15859–15869, 2023. 2

[30] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*, 2016. 1

[31] Sanghun Jung, Jungsoo Lee, Daehoon Gwak, Sungha Choi, and Jaegul Choo. Standardized max logits: A simple yet effective approach for identifying unexpected road obstacles in urban-scene segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15425–15434, 2021. 5, 6

[32] A Emre Kavur, N Sinem Gezer, Mustafa Barış, Sinem Aslan, Pierre-Henri Conze, Vladimir Groza, Duc Duy Pham, Soumick Chatterjee, Philipp Ernst, Savaş Özkan, et al. Chaos challenge-combined (ct-mr) healthy abdominal organ segmentation. *Medical Image Analysis*, 69:101950, 2021. 5, 1, 2

[33] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023. 2, 8

[34] Bennett Landman, Zhoubing Xu, J Igelsias, Martin Styner, T Langerak, and Arno Klein. Miccai multi-atlas labeling beyond the cranial vault–workshop and challenge. In *Proc. MICCAI Multi-Atlas Labeling Beyond Cranial Vault—Workshop Challenge*, page 12, 2015. 2, 5, 6, 7, 1

[35] Feng Liang, Bichen Wu, Xiaoliang Dai, Kunpeng Li, Yinan Zhao, Hang Zhang, Peizhao Zhang, Peter Vajda, and Diana Marculescu. Open-vocabulary semantic segmentation with mask-adapted clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7061–7070, 2023. 2, 3, 5, 6, 4

[36] 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. 1, 2, 5, 6, 7, 8

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

[38] Xiangde Luo, Wenjun Liao, Jianghong Xiao, Jieneng Chen, Tao Song, Xiaofan Zhang, Kang Li, Dimitris N Metaxas, Guotai Wang, and Shaoting Zhang. Word: A large scale dataset, benchmark and clinical applicable study for abdominal organ segmentation from ct image. *Medical Image Analysis*, 82:102642, 2022. 5, 1, 2

[39] Jun Ma and Bo Wang. Segment anything in medical images. *arXiv preprint arXiv:2304.12306*, 2023. 2, 8

[40] Jun Ma, Yao Zhang, Song Gu, Cheng Zhu, Cheng Ge, Yichi Zhang, Xingle An, Congcong Wang, Qiyuan Wang, Xin Liu, et al. Abdomenct-1k: Is abdominal organ segmentation a solved problem. *IEEE TPAMI*, 2021. 5, 1, 2

[41] Shiqiang Ma, Xuejian Li, Jijun Tang, and Fei Guo. A zero-shot method for 3d medical image segmentation. In *2021 IEEE international conference on multimedia and expo (ICME)*, pages 1–6. IEEE, 2021. 8

[42] Chris J Maddison, Andriy Mnih, and Yee Whye Teh. The concrete distribution: A continuous relaxation of discrete random variables. In *ICLR*, 2017. 4

[43] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26, 2013. 1

[44] OpenAI. Gpt-4 technical report, 2023. 5, 2

[45] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *NeurIPS*, 32, 2019. 5

[46] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014. 1

[47] Walter HL Pinaya, Petru-Daniel Tudosiu, Robert Gray, Geraint Rees, Parashkev Nachev, Sebastien Ourselin, and M Jorge Cardoso. Unsupervised brain imaging 3d anomaly detection and segmentation with transformers. *Medical Image Analysis*, 79:102475, 2022. 3

[48] Jie Qin, Jie Wu, Pengxiang Yan, Ming Li, Ren Yuxi, Xuefeng Xiao, Yitong Wang, Rui Wang, Shilei Wen, Xin Pan, et al. Freeseg: Unified, universal and open-vocabulary image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19446–19455, 2023. 2, 3, 6

[49] Taiping Qu, Xiuli Li, Xiheng Wang, Wenyi Deng, Li Mao, Ming He, Xiao Li, Yun Wang, Zaiyi Liu, Longjiang Zhang, et al. Transformer guided progressive fusion network for 3d pancreas and pancreatic mass segmentation. *Medical Image Analysis*, 86:102801, 2023. 1, 2

[50] 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. 2, 3, 4

[51] Blaine Rister, Darvin Yi, Kaushik Shivakumar, Tomomi Nobashi, and Daniel L Rubin. Ct-org, a new dataset for multiple organ segmentation in computed tomography. *Scientific Data*, 7(1):381, 2020. 5, 1, 2

[52] Holger R Roth, Le Lu, Amal Farag, Hoo-Chang Shin, Jiamin Liu, Evrim B Turkbey, and Ronald M Summers. Deeporgan: Multi-level deep convolutional networks for automated pancreas segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part I 18*, pages 556–564. Springer, 2015. 5, 1, 2

[53] Abhijit Guha Roy, Jie Ren, Shekoofeh Azizi, Aaron Loh, Vivek Natarajan, Basil Mustafa, Nick Pawlowski, Jan Freyberg, Yuan Liu, Zach Beaver, et al. Does your dermatology classifier know what it doesn't know? detecting the long-tail of unseen conditions. *Medical Image Analysis*, 75:102274, 2022. 3

[54] Hyunseok Seo, Charles Huang, Maxime Bassenne, Ruoxiu Xiao, and Lei Xing. Modified u-net (mu-net) with incorporation of object-dependent high level features for improved liver and liver-tumor segmentation in ct images. *IEEE transactions on medical imaging*, 39(5):1316–1325, 2019. 1

[55] Gonglei Shi, Li Xiao, Yang Chen, and S Kevin Zhou. Marginal loss and exclusion loss for partially supervised multi-organ segmentation. *Medical Image Analysis*, 70: 101979, 2021. 2

[56] 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 *CVPR*, pages 20730–20740, 2022. 2, 6, 7

[57] Yu Tian, Guansong Pang, Fengbei Liu, Yuanhong Chen, Seon Ho Shin, Johan W Verjans, Rajvinder Singh, and Gustavo Carneiro. Constrained contrastive distribution learning for unsupervised anomaly detection and localisation in medical images. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part V 24*, pages 128–140. Springer, 2021. 3

[58] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017. 4, 5

[59] Haoyu Wang, Sizheng Guo, Jin Ye, Zhongying Deng, Junlong Cheng, Tianbin Li, Jianpin Chen, Yanzhou Su, Ziyan Huang, Yiqing Shen, et al. Sam-med3d. *arXiv preprint arXiv:2310.15161*, 2023. 2

[60] Jakob Wasserthal, Hanns-Christian Breit, Manfred T Meyer, Maurice Pradella, Daniel Hinck, Alexander W Sauter, Tobias Heye, Daniel Boll, Joshy Cyriac, Shan Yang, et al. Totalsegmentator: robust segmentation of 104 anatomical structures in ct images. *arXiv preprint arXiv:2208.05868*, 2022. 5, 1, 2

[61] Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Towards generalist foundation model for radiology. *arXiv preprint arXiv:2308.02463*, 2023. 2

[62] Junde Wu, Rao Fu, Huihui Fang, Yuanpei Liu, Zhaowei Wang, Yanwu Xu, Yueming Jin, and Tal Arbel. Medical sam adapter: Adapting segment anything model for medical image segmentation. *arXiv preprint arXiv:2304.12620*, 2023. 2, 8

[63] Yingda Xia, Yi Zhang, Fengze Liu, Wei Shen, and Alan L Yuille. Synthesize then compare: Detecting failures and anomalies for semantic segmentation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pages 145–161. Springer, 2020. 5, 6

[64] Yongqin Xian, Subhabrata Choudhury, Yang He, Bernt Schiele, and Zeynep Akata. Semantic projection network for zero-and few-label semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8256–8265, 2019. 1

[65] Yutong Xie, Jianpeng Zhang, Yong Xia, and Qi Wu. Unimiss: Universal medical self-supervised learning via breaking dimensionality barrier. In *ECCV*, pages 558–575. Springer, 2022. 2

[66] Jiarui Xu, Shalini De Mello, Sifei Liu, Wonmin Byeon, Thomas Breuel, Jan Kautz, and Xiaolong Wang. Groupvit: Semantic segmentation emerges from text supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18134–18144, 2022. 4

[67] Mengde Xu, Zheng Zhang, Fangyun Wei, Yutong Lin, Yue Cao, Han Hu, and Xiang Bai. A simple baseline for open-vocabulary semantic segmentation with pre-trained vision-language model. In *European Conference on Computer Vision*, pages 736–753. Springer, 2022. 2, 3, 5, 6

[68] Lisha Yao, Yingda Xia, Haochen Zhang, Jiawen Yao, Dakai Jin, Bingjiang Qiu, Yuan Zhang, Suyun Li, Yanting Liang, Xian-Sheng Hua, et al. Deepcrc: Colorectum and colorectal cancer segmentation in ct scans via deep colorectal coordinate transform. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 564–573. Springer, 2022. 2

[69] Qihang Yu, Dong Yang, Holger Roth, Yutong Bai, Yixiao Zhang, Alan L Yuille, and Daguang Xu. C2fnas: Coarse-to-fine neural architecture search for 3d medical image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4126–4135, 2020. 2

[70] Mingze Yuan, Yingda Xia, Hexin Dong, Zifan Chen, Jiawen Yao, Mingyan Qiu, Ke Yan, Xiaoli Yin, Yu Shi, Xin Chen, et al. Devil is in the queries: Advancing mask transformers for real-world medical image segmentation and out-of-distribution localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23879–23889, 2023. 3, 4, 5, 6

[71] Hui Zhang and Henghui Ding. Prototypical matching and open set rejection for zero-shot semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6974–6983, 2021. 1

[72] Jianpeng Zhang, Yutong Xie, Yong Xia, and Chunhua Shen. Dodnet: Learning to segment multi-organ and tumors from

multiple partially labeled datasets. In *CVPR*, pages 1195–1204, 2021. 2

[73] Shaoting Zhang and Dimitris Metaxas. On the challenges and perspectives of foundation models for medical image analysis. *arXiv preprint arXiv:2306.05705*, 2023. 1

[74] Tianyi Zhao, Kai Cao, Jiawen Yao, Isabella Nogues, Le Lu, Lingyun Huang, Jing Xiao, Zhaozheng Yin, and Ling Zhang. 3d graph anatomy geometry-integrated network for pancreatic mass segmentation, diagnosis, and quantitative patient management. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13743–13752, 2021. 1

[75] 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*, 2023. 2

[76] Chaoyang Zhu and Long Chen. A survey on open-vocabulary detection and segmentation: Past, present, and future. *arXiv preprint arXiv:2307.09220*, 2023. 2

[77] David Zimmerer, Peter M Full, Fabian Isensee, Paul Jäger, Tim Adler, Jens Petersen, Gregor Köhler, Tobias Ross, Annika Reinke, Antanas Kascenas, et al. Mood 2020: A public benchmark for out-of-distribution detection and localization on medical images. *IEEE Transactions on Medical Imaging*, 41(10):2728–2738, 2022. 3