Devil is in the Queries: Advancing Mask Transformers for Real-world Medical Image Segmentation and Out-of-Distribution Localization

Mingze Yuan[1,2], Yingda Xia[1,*], Hexin Dong[1,2], Zifan Chen[2], Jiawen Yao[1], Mingyan Qiu[1], Ke Yan[1],
Xiaoli Yin[3], Yu Shi[3], Xin Chen[4], Zaiyi Liu[4], Bin Dong[2], Jingren Zhou[1], Le Lu[1], Ling Zhang[1], Li Zhang[2]

[1]DAMO Academy, Alibaba Group   [2]Peking University
[3]Shengjing Hospital   [4]Guangdong Province People's Hospital

Abstract

Real-world medical image segmentation has tremendous long-tailed complexity of objects, among which tail conditions correlate with relatively rare diseases and are clinically significant. A trustworthy medical AI algorithm should demonstrate its effectiveness on tail conditions to avoid clinically dangerous damage in these out-of-distribution (OOD) cases. In this paper, we adopt the concept of object queries in Mask Transformers to formulate semantic segmentation as a soft cluster assignment. The queries fit the feature-level cluster centers of inliers during training. Therefore, when performing inference on a medical image in real-world scenarios, the similarity between pixels and the queries detects and localizes OOD regions. We term this OOD localization as MaxQuery. Furthermore, the foregrounds of real-world medical images, whether OOD objects or inliers, are lesions. The difference between them is less than that between the foreground and background, possibly misleading the object queries to focus redundantly on the background. Thus, we propose a query-distribution (QD) loss to enforce clear boundaries between segmentation targets and other regions at the query level, improving the inlier segmentation and OOD indication. Our proposed framework is tested on two real-world segmentation tasks, i.e., segmentation of pancreatic and liver tumors, outperforming previous state-of-the-art algorithms by an average of 7.39% on AUROC, 14.69% on AUPR, and 13.79% on FPR95 for OOD localization. On the other hand, our framework improves the performance of inlier segmentation by an average of 5.27% DSC when compared with the leading baseline nnUNet.

1. Introduction

Image segmentation is a fundamental task in medical image analysis. With the recent advancements in computer
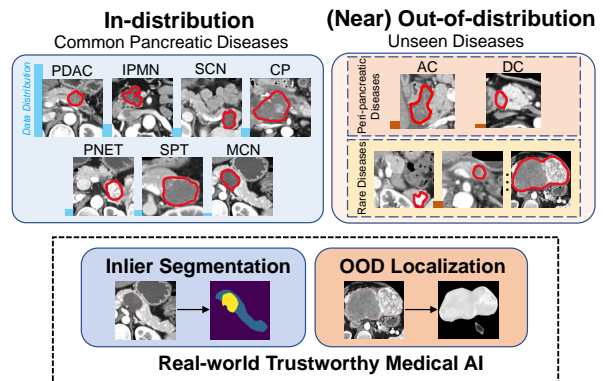

Figure 1. Real-world medical image segmentation. Real-world medical outliers (unseen, usually rare, tumors) are "near" to the inliers (labeled lesions), forming a typical near-OOD problem. A real-world medical OOD detection/localization model should focus more on subtle differences between outliers and inliers than the significant difference between foreground and background..

vision and deep learning, automated medical image segmentation has reached expert-level performance in various applications [3, 28, 52]. Most medical image segmentation methods are based on supervised machine learning that heavily relies on collecting and annotating training data. However, real-world medical images are long-tailed distributed. The tail conditions are outliers and inadequate (or even unable) to train a reliable model [35, 59, 61]. Yet, the model trained with inliers is risky for triggering failures or errors in real-world clinical deployment [41]. For example, in pancreatic tumor image analysis, a miss-detection of metastatic cancer will directly threaten life; an erroneous recognition of a benign cyst as malignant will lead to unnecessary follow-up tests and patient anxiety. Medical image segmentation models should thus demonstrate the ability to detect and localize out-of-distribution (OOD) conditions, especially in some safety-critical clinical applications.

Previous studies have made valuable attempts on med-

---

* Corresponding author. (yingda.xia@alibaba-inc.com)

ical OOD localization [46, 63], including finding lesions apart from normal cases or simulating OOD conditions for model validation. However, the real-world clinical scenario, such as tumor segmentation, is more complex, where either in-distribution or OOD cases have multiple types of tumors. Establishing a direct relationship between image pixels and excessive semantics (types of tumors) is difficult for real-world medical image segmentation. Using this relationship to distinguish inliers and outliers is even more challenging. Fortunately, several works about Mask Transformers [5, 9] have inspired us to split segmentation as a two-stage process of per-pixel cluster assignment and cluster classification [55, 56]. A well-defined set of inlier clusters may greatly benefit in identifying the OOD conditions from the medical images. Therefore, we propose MaxQuery, a medical image semantic segmentation framework that advances Mask Transformers to localize OOD targets. The framework adopts learnable object queries to iteratively fit inlier cluster centers. Since the affinity between OODs and an inlier cluster center should be less than that within the cluster (between inliers and cluster centers), MaxQuery uses the negative of such affinity as an indicator to detect OODs.

Several recent works further define real-world medical image OOD localization as a near-OOD problem [41, 49], where the distribution gaps between inlier and OOD tumors are overly subtle, as shown in Fig. 1. Thus, the near-OOD problems are more difficult. Our pilot experiments show that the cluster centers redundantly represent the large regions of background and organ rather than tumors, compromising the necessary variability of the cluster assignments for OOD localization. To solve this issue, we propose the query-distribution (QD) loss to regularize specific quantities of object queries on background, organ, and tumors. This enforces the diversity of the cluster assignments, benefiting the segmentation and recognition of OOD tumors.

We curate two real-world medical image datasets of (pancreatic and liver) tumor images from 1,088 patients for image segmentation and OOD localization. Specifically, we collect consecutive patients' contrast-enhanced 3D CT imaging with a full spectrum of tumor types confirmed by pathology. In these scenarios, the OOD targets are rare tumors and diseases. Our method shows robust performance across two datasets, significantly outperforming the previous leading OOD localization methods by an average of 7.39% in AUROC, 14.69% in AUPR, 13.79% in FPR95 for localization, and 3.42% for case-level detection. Meanwhile, our framework also improves the performance of inlier segmentation by an average of 5.27% compared with the strong baseline nnUNet [24].

We summarize our main contributions as follows:

- To the best of our knowledge, we are the first to explore the near-OOD detection and localization problem in medical image segmentation. The proposed method

has a strong potential for utility in clinical practice.

- We propose a novel approach, MaxQuery, using the maximum score of query response as a major indicator for OOD localization.

- A query-distribution (QD) loss is proposed to concentrate the queries on important foreground regions, demonstrating superior effectiveness for near-OOD problems.

- We curate two medical image datasets for tumor semantic segmentation/detection of real-world OODs. Our proposed framework substantially outperforms previous leading OOD localization methods and improves upon the inlier segmentation performance.

2. Related Work

Medical Image Segmentation and Diagnosis. U-Net [40] and its variants [31, 34, 37, 54, 62] have been promoting the development of medical image segmentation. A recent self-configuring U-Net (nnUNet) [24, 25] further surpassed existing approaches in various medical image segmentation tasks with minimal manual parameter tuning. Semantic segmentation serves as the core for downstream clinical tasks of disease detection [10], differential diagnosis [12, 59], survival prediction [52], therapy planning [44], and treatment response assessment [28]. Therefore, developing a reliable segmentation method is critical to improving safety in real-world clinical use. After the publication of Vision Transformers (ViTs) [16], integrating subsequent transformer blocks into the backbone of network architecture [7, 17, 18, 45] has been investigated. ViTs achieved improved results over traditional U-Net, particularly for multi-class semantic segmentation tasks. This work greatly focuses on exploring the real-world OOD localization detection problem over medical image segmentation. Current solutions provide limited performance, so we study a novel architecture combining Transformer and nnUNet for improving segmentation performance under clinical tasks, utilizing segmentation to detect and diagnose minority tumors [59].

Mask Transformers. Unlike using Transformers directly as network backbones for natural and medical image segmentation [36, 43, 51, 57, 60], Mask Transformers seek to enhance the CNN-based backbone with stand-alone transformer blocks. MaX-Deeplab [48] interprets object queries in DETR [5] as memory-encoded queries for end-to-end panoptic segmentation. MaskFormer [9] further applies this design to semantic segmentation by unifying the CNN and the transformer branches. Afterward, Mask2Former [8] technically improves over its predecessor. Recently, CMT-Deeplab [55] and KMaX-Deeplab [56] propose to interpret the queries as clustering centers and add regulatory constraints for learning the cluster representations of the

queries. The design of Mask Transformers is intuitively suitable for medical image segmentation, especially for the semantic segmentation and diagnosis of tumors. This task requires the network to be locally sensitive to image textures for tumor segmentation and can globally understand organ-tumor morphological information for tumor sub-type recognition. To our knowledge, we are the first to adapt Mask Transformers for medical image segmentation and further explore its usage of recognizing outliers via queries.

OOD Detection and Localization. OOD Detection aims to detect the out-of-distribution conditions (outliers) that are unseen in the training data. Maximal softmax probability (MSP) [21] serves as a strong baseline. After that, various approaches improved OOD detection from multiple aspects [13, 22, 29, 30]. These approaches focus on image-level OOD detection, and efforts have also been made to localize OOD objects or regions on a large image, e.g., urban driving scenes [4, 6, 14, 21, 26, 32, 50]. Despite the advance of OOD detection and localization on natural images, its application on real-world medical images is challenging. Since the difference between foregrounds in real-world medical images is subtle, their OOD detection/localization becomes a typical near-OOD problem [15, 38, 39, 49]. Therefore, the existing OOD solutions could hardly be recommended for clinical practice [46, 63]. Recent work, HOD [41], paces one step toward real-world OOD detection of rare diseases in dermatology classification.

3. Method

In this section, we first provide an overview of our method and then describe our proposed query-distribution (QD) loss and MaxQuery framework for OOD localization.

3.1. Method Overview

Medical image segmentation aims to segment an image into multiple regions representing anatomical objects of interest. Here, we focus on 3D medical image $\mathbf{X} \in \mathbb{R}^{H \times W \times D}$, and use a segmentation model to partition it into K category-labeled binary masks,

$$\mathbf{G} = \{\mathbf{G}_i\}_{i=1}^{K}, \tag{1}$$

where $\mathbf{G}_i \in \{0, 1\}^{H \times W \times D}$ is the ground truth mask that belongs to the i-th class, and $\sum_{i=1}^{K} \mathbf{G}_i = \mathbf{1}^{H \times W \times D}$. In our problem, class 1 refers to background, class 2 stands for specific organ, and the others for tumors. Since the real-world medical image dataset has a long-tail distribution in quantity, its segmentation task should be divided into supervised inlier segmentation and pixel-level OOD localization.

Inlier Segmentation. As shown in Fig. 2, we build our model with a CNN backbone to extract per-pixel features $\mathbf{P} \in \mathbb{R}^{HWD \times C}$ and a transformer module. The transformer module gradually updates a set of learnable object queries, $\mathbf{C} \in \mathbb{R}^{N \times C}$, to meaningful mask embedding vectors via cross attention between object queries and per-pixel features,

$$\mathbf{C} \leftarrow \mathbf{C} + \operatorname*{argmax}_{N}(\mathbf{Q}^c \times (\mathbf{K}^p)^{\mathrm{T}}) \times \mathbf{V}^p, \tag{2}$$

where the superscripts c and p represent query and pixel features, respectively. We also adopt cluster-wise argmax from KMax-DeepLab [56] to substitute spatial-wise softmax in the original cross attention settings.

Inspired by recent works on cluster analysis of mask transformers [55, 56], we consider semantic segmentation as a two-stage cluster analysis process. First, all pixels are assigned into different clusters. The mask embedding vectors $\mathbf{C}$ from the transformer module are formulated as the cluster centers. The product $\mathbf{R}$ of $\mathbf{C}$ and $\mathbf{P}^{\mathrm{T}}$ represents the query response, which expresses the similarity between each pixel and cluster centers. Then, we use the query-wise softmax activation on query responses $\mathbf{R}$ to generate a mask prediction, which encourage the exclusiveness of cluster assignment. The mask prediction (cluster assignment) $\mathbf{M}$ is defined as,

$$\mathbf{M} = \operatorname*{softmax}_{N}(\mathbf{R}) = \operatorname*{softmax}_{N}(\mathbf{C} \times \mathbf{P}^{\mathrm{T}}). \tag{3}$$

Notably, different from the sigmoid activation used in [8, 9], the query-wise softmax activation could better guide the object queries (cluster centers) to focus on different regions of the image and encourage diversity in real-world medical image segmentation.

Secondly, the grouped pixels are classified under the guidance of cluster classification. We evaluate the cluster centers $\mathbf{C}$ via a multi-layer perceptron (MLP) to predict the K-channel cluster classifications $\mathbf{C}_K \in \mathbb{R}^{N \times K}$ for all N clusters. We then aggregate the cluster assignments $\mathbf{M}$ of grouped pixels and their classifications $\mathbf{C}_K$ for the final semantic segmentation,

$$\mathbf{Z} = (\mathbf{C}_K)^{\mathrm{T}} \times \mathbf{M}, \tag{4}$$

where $\mathbf{Z} \in \mathbb{R}^{K \times HWD}$ represents the final logits. To supervise the final segmentation, we combine the classic segmentation loss and a novel QD loss between final output $\mathbf{Z}$ and ground truth $\mathbf{G}$, more details in Sec 3.2.

OOD Localization. To further segment abnormal regions unseen in training images, an OOD localization process is required when performing inference on a test image. Formally, given a test image $\mathbf{X} \in \mathbb{R}^{H \times W \times D}$, OOD localization evaluates the query response to find the maximal one that represents the similarity between the pixel and its assigned cluster center. Then, the model can generate an pixel-wise anomalous score map $\mathbf{A} \in [0, 1]^{H \times W \times D}$, where $\mathbf{A}_i = 1$ and $\mathbf{A}_i = 0$ represent that i-th pixel in $\mathbf{X}$ belongs
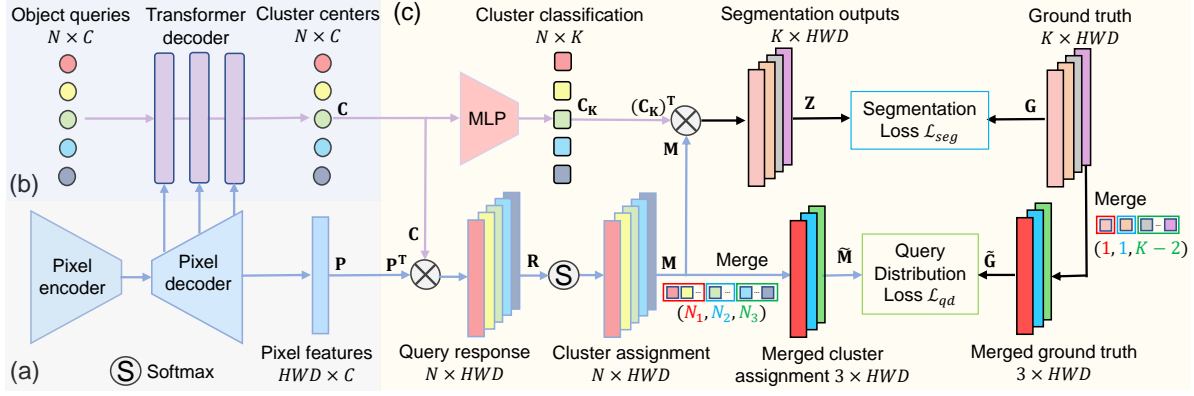
Figure 2. Overview of our proposed framework. (a) A CNN backbone for image segmentation, here we use nnUNet [24]; (b) A transformer decoder interatively updates the object queries to fit the inlier cluster centers; (c) A two-stage cluster analysis: 1) cluster assignment groups the pixels based on the affinity between pixel features and cluster centers; 2) cluster classification guides the grouped pixels to generate segmentation logits. The overall segmentation is supervised by a classic segmentation loss and a novel query-distribution loss.

to an OOD class and an in-distribution class, respectively. More details of this novel OOD localization (MaxQuery) is in Sec 3.3

3.2. Managing Cluster Distribution with QD Loss

Classic segmentation loss serves as an important learning target of our model. We combine the Cross-Entropy and Dice losses between final output $\mathbf{Z}$ and ground truth $\mathbf{G}$ in Eq. (1) as the segmentation loss, *i.e.*, $\mathcal{L}_{seg} = \ell_{ce} + \ell_{dc}$. However, when only using classic segmentation loss, object queries focus majorly on the background and organs rather than the tumors. The significant difference between foreground and background greatly distracts the model from focusing on subtle differences between OOD objects and inliers. As later shown in an example in Fig 6 , some queries may even have mixed representation on background and foreground which is an unsatisfactory phenomenon for discriminative cluster learning. Therefore, we propose query-distribution (QD) loss to manipulate the object queries and guide them to focus on the foreground, especially the tumors, and encourage concentrated cluster learning. The key idea is to use ground-truth $\mathbf{G} \in \mathbb{R}^{K \times HWD}$ to supervise the cluster assignment probability maps $\mathbf{M}$. This motivation also benefits OOD localization as introduced in Sec 3.3.

We thus divide the N channels into three groups, including N_1, N_2, N_3 queries, for background, organ and tumor regions, respectively. Our goal is to associate the first N_1 channels of $\mathbf{M}$ (representing the assignment probabilities of the first N_1 cluster centers) with the background class $\mathbf{G}_1$, the next N_2 channels with the organ class $\mathbf{G}_2$, and the last N_3 channels with the tumor classes $\sum_{i=3}^{K} \mathbf{G}_i$. We define the merged cluster assignments $\tilde{\mathbf{M}}$ and class labels $\tilde{\mathbf{G}}$

as the following,

$$
\begin{aligned}
\tilde{\mathbf{M}} &= (\tilde{\mathbf{M}}_1, \tilde{\mathbf{M}}_2, \tilde{\mathbf{M}}_3) \in \mathbb{R}^{3 \times HWD} \\
&= (\sum_{i=1}^{N_1} \mathbf{M}_i, \sum_{j=1}^{N_2} \mathbf{M}_{N_1+j}, \sum_{k=1}^{N_3} \mathbf{M}_{N_1+N_2+k}),
\end{aligned}
\tag{5}
$$

$$
\tilde{\mathbf{G}} = (\tilde{\mathbf{G}}_1, \tilde{\mathbf{G}}_2, \tilde{\mathbf{G}}_3) = (\mathbf{G}_1, \mathbf{G}_2, \sum_{i=3}^{K} \mathbf{G}_i) \in \mathbb{R}^{3 \times HWD},
\tag{6}
$$

where the merged $\tilde{\mathbf{M}}$ are still probability distributions in each spatial position, *i.e.*, $\sum_{j=1}^{3} \tilde{\mathbf{M}}_i = \mathbf{1}^{H \times W \times D}$.

Finally, we formulate the QD loss as the negative log likelihood loss between $\tilde{\mathbf{M}}$ and $\tilde{\mathbf{G}}$,

$$
\mathcal{L}_{qd} = -\sum_{j=1}^{HWD} \sum_{i=1}^{3} \tilde{\mathbf{G}}_{ij} \log \tilde{\mathbf{M}}_{ij},
\tag{7}
$$

which draws strict boundaries between different types of cluster assignments ($\tilde{\mathbf{M}}_1$, $\tilde{\mathbf{M}}_2$, and $\tilde{\mathbf{M}}_3$) based on the ground truth. The final loss function $\mathcal{L}$ is a combination of segmentation loss $\mathcal{L}_{seg}$ and QD loss $\mathcal{L}_{qd}$ with a balance weight λ, formulated as,

$$
\mathcal{L} = \mathcal{L}_{seg} + \lambda \mathcal{L}_{qd}.
\tag{8}
$$

3.3. Localizing OOD Regions with MaxQuery

Given a test image $\mathbf{X} \in \mathbb{R}^{H \times W \times D}$, our mask transformer will yield the pixel-level query response $\mathbf{R} \in \mathbb{R}^{N \times H \times W \times D}$, representing the affinity of pixel feature and cluster centers. The maximal query response of one pixel then represents the similarity between the pixel and its assigned cluster center. Intuitively, the maximal query response of outliers should be smaller than inliers. We therefore adopt the negative of maximal query response in Eq. (3)
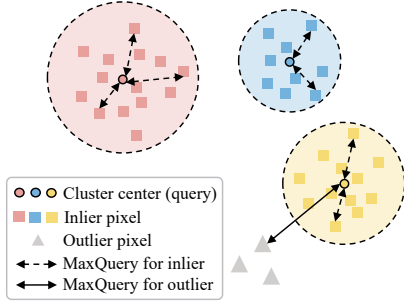
Figure 3. Illustration of how MaxQuery works. MaxQuery, *i.e.*, the negative of maximal query response, reflects the distance of the pixel and its assigned cluster center. MaxQuery of the inlier (dotted arrow) is usually smaller than that of the outlier (solid arrow) and thus is able to identify the anomalous/OOD pixels.

as the pixel-wise anomaly score, called MaxQuery, *i.e.*,

$$\mathbf{A} = -\max_N \mathbf{R}, \qquad (9)$$

where $\mathbf{R} \in \mathbb{R}^{N \times H \times W \times D}$ represents the query response matrix and $\mathbf{A} \in \mathbb{R}^{H \times W \times D}$ indicates the anomaly score map. The anomaly score can be further normalized into $[0, 1]$ by min-max normalization. Figure 3 illustrates the capability of MaxQuery for OOD pixels identification. The subscript N means that we perform maximum operation on the query dimension. We add a minus sign because when the maximal query response of a pixel is larger, it is less likely to be an OOD pixel.

In addition, we compare the results of anomaly score maps according to the maximum of query responses $\mathbf{R}$ (pre-softmax, $\mathbf{A} = -\max_N \mathbf{R}$) and cluster assignments $\mathbf{M}$ (post-softmax, $\mathbf{A}' = -\max_N \mathbf{M}$). $\mathbf{A}$ greatly outperforms $\mathbf{A}'$, since if an inlier pixel is evenly close to multiple cluster centers, the maximal score in $\mathbf{M}$ can be very low and easily be mis-classified as an outlier. But with maximum query response $\mathbf{R}$ (pre-softmax), the score is still high enough for an indication of inlier. Thus we choose the maximal query response to imply the anomalous regions.

4. Experiments

4.1. Datasets and Experiment Setting

We collect two datasets, i.e., pancreas and liver tumor segmentation datasets, which includes contrast-enhanced 3D CT scans from consecutive patients before treatment. We register the multi-phase CT scans into arterial late and venous phase using DEEDS [19], respectively. All types of tumors are confirmed by pathology, except for cysts in the liver (confirmed by a radiologist specialized in liver imaging). All tumors are annotated slice-by-slice manually on the CT phase with the best tumor visibility by experienced radiologists specialized in specific diseases. The or-

gan (pancreas or liver) in each dataset is first annotated automatically by a self-learning approach [58] trained on public datasets (e.g., Medical Decathlon [1]) and then edited by engineers.

Pancreatic Multi-type Tumors dataset contains 661 patients. Every patient has five phases of CT scans: noncontrast, arterial-early, arterial-late, venous, and delay. The median spacing is $3 \times 0.419 \times 0.419$ mm. According to previous clinical studies about pancreatic tumor classification [11, 42], we assign the seven most common conditions (PDAC, PNET, SPT, IPMN, MCN, CP, and SCN) as inliers, and allocate AC, DC, and "other" as outliers. We randomly split 590 inlier data into 378(64%) training, 94(16%) validation, and 118(20%) testing, and leave out all 71 outlier data for OOD testing.

Liver Multi-type Tumors dataset contains 427 patients. Each patient has three phases of CT scans: noncontrast, arterial, and venous. The median spacing is $3 \times 0.760 \times 0.760$ mm. Following [53], we assign the five most common conditions (HCC, ICC, metastasis, hemangiomas, and cyst) as inliers, and allocate hepatoblastoma, FNH, and "other" as outliers. Similarly, We randomly split 327 inlier data into 209(64%) training, 52(16%) validation, and 66(20%) testing, and leave out all 100 outlier data for OOD testing. Notice that the "other" class in both datasets contains multiple rare diseases, reflecting the long-tailed distribution of real-world disease incidence.

4.2. Implementation & Evaluation Metrics

Network Architecture. We use the current benchmark model in medical image segmentation, nnUNet [24], as a CNN backbone, which consists of a pixel encoder and a pixel decoder with skip connections. We adopt four transformer decoder blocks, and each takes pixel features with output stride 32, 16, 8, and 4, respectively. The self-attention layer in the block has 8 heads. Since medical image segmentation is sensitive to local textures, we add a decoder block for output stride 4 compared with previous works [48, 56]. To increase numerical stability, we add an InstanceNorm [47] layer and a LayerNorm [2] at the end of pixel-level and transformer decoder modules, respectively.

Training and Testing. Each CT scan is resampled into the median spacing per tumor dataset (*e.g.*, $3 \times 0.419 \times 0.419$ mm for the pancreatic dataset) and normalized into zero mean and unit variance. Our model is trained using a batch size of 2 on one GPU (with $28 \times 192 \times 320$ patch size for pancreatic, $40 \times 192 \times 224$ for liver). We adopt the drop path [23] strategy with a probability of 0.2 for regularization. During training, extensive data augmentation is utilized on-the-fly [24] to improve the generalization, including random rotation and scaling, elastic deformation, additive brightness, and gamma scaling. The network is trained with RAdam [33] with the initial learning rate as

	Pancreatic %				Liver %			
	OOD Localization			OOD_{case}	OOD Localization			OOD_{case}
Methods	AUROC↑	AUPR↑	FPR_{95} ↓	AUC↑	AUROC↑	AUPR↑	FPR_{95} ↓	AUC↑
MC Dropout [27]	49.08	11.47	84.60	72.91	39.61	16.05	91.13	34.05
MSP [21]	53.81	13.44	86.44	73.38	75.14	25.27	70.04	66.76
MaxLogit [20]	58.46	21.93	83.68	73.42	78.60	35.47	48.73	65.68
SynthCP [50]	69.86	26.50	66.65	68.43	74.93	34.03	57.91	63.34
SML [26]	56.10	30.44	77.81	62.26	86.64	44.59	31.04	63.85
Ours (w/o $\mathcal{L}_{qd}$)	63.54	25.25	67.09	74.87	74.95	42.31	53.52	65.91
Ours	**82.52**	**55.60**	**46.19**	**77.97**	**88.75**	**48.80**	**23.93**	**69.04**

Table 1. OOD localization and case-level OOD detection performance on *Pancreatic Tumors* and *Liver Tumors*. Our proposed method achieves state-of-the-art OOD detection performance at both pixel level and case level. All the methods are implemented based on the nnUNet [24] backbone. (OOD_{case}: case-level OOD detection.)

	Pancreatic %								Liver %					
Methods	PDAC	IPMN	PNET	SCN	CP	SPT	MCN	Avg.	HCC	ICC	Meta.	Heman.	Cyst	Avg.
nnUNet [24]	65.65	27.60	32.59	36.46	23.33	31.73	30.96	35.47	57.22	28.16	52.81	77.55	46.49	52.45
Ours (w/o $\mathcal{L}_{qd}$)	65.87	28.3	32.43	40.63	28.93	30.77	30.89	36.84	60.91	30.58	53.21	78.47	46.42	53.92
Ours	67.91	46.92	32.07	42.51	31.36	42.67	28.97	**41.77**	67.61	30.78	60.40	77.07	47.61	**56.69**

Table 2. Inlier segmentation Dice scores (%) on *val* set of *Pancreatic Tumors* and *Liver Tumors* (all methods report results with final checkpoint). Compared with the benchmark model (nnUNet [24]) in medical image segmentation, our method noticeably outperforms the strong baseline for the task of inlier tumor segmentation.

1×10^{-4} and a polynomial learning rate decay. We first pre-train the nnUNet backbone for 1000 epochs and fine-tune the whole architecture jointly for another 200 epochs. During finetuning, we keep the backbone weights fixed for the first 50 epochs, and then set it with a learning rate multiplier of 0.1 for the next 150 epochs. The number of object queries (*i.e.*, cluster centers) N is 32, and the query distribution (N_1, N_2, N_3) is set as (16, 4, 12). We follow KMax-DeepLab [56] to directly add deep supervision on the attention map of (k-means) cross attention to align it with the final segmentation after the segmentation output head. The loss weight λ for QD loss is 0.1.

Evaluation Metrics. For OOD localization, we follow the standard metrics for anomaly segmentation [14, 26, 50]. We compute the area under receptive-operative curve (AUROC) and the area under precision-recall curve (AUPR). We also report FPR at the TPR level of 0.95 (FPR95) as OOD localization metrics since the false positive rate is safety-critical in clinical practice. For case-level OOD detection, we compute the average of anomaly scores in predicted tumor regions as the case-level anomaly score and choose AUC as the case-level OOD detection metric. Meanwhile, we report the average Dice Score of inlier tumors to evaluate the segmentation performance on inlier classes.

Baselines. For OOD localization, we compare our work with a series of representative anomaly segmentation methods in multiple aspects, including uncertainty statistics-based (MSP [21], MaxLogit [20], SML [26]), Bayesian deep learning-based (MC Dropout [27]) and image re-synthesis-based (SynthCP [50]) methods. All of them are implemented using nnUNet [24] backbone. For inlier segmentation, we compare our work with the benchmark model (nnUNet [24]) and previous leading model (Swin UNETR [45]) implemented by their officially released code and pre-trained model with same settings.

4.3. Main Results

Comparisons on the real-world datasets, including *Pancreatic Tumors* and *Liver Tumors*, are summarized in Tabs. 1 and 2. We also present visualization examples in Figs. 4 and 5 to better understand the role of object queries in our proposed mask transformer and compare different anomaly segmentation methods.

Pancreatic Tumors. In Tab. 1, we compare MaxQuery with other baselines on *Pancreatic Tumors*. Our framework shows the best performance in all metrics. Specifically, our framework outperforms the previous best method SML [26] by a large margin of 12.66% in AUROC, 25.16% in AUPR, 20.46% in FPR95 for OOD localization, and 4.55% in AUC for case-level OOD detection. For qualitative analysis, we present four visual examples from *Pancreatic Tumors* by visualizing the anomaly score map of MSP [21], MaxLogit [20], SML [26], and ours. As shown in Fig. 4, our method maintains a high anomaly score in the OOD pixels (outlier tumor), while a low anomaly score in the in-distribution pixels (organ). Moreover, the previ-
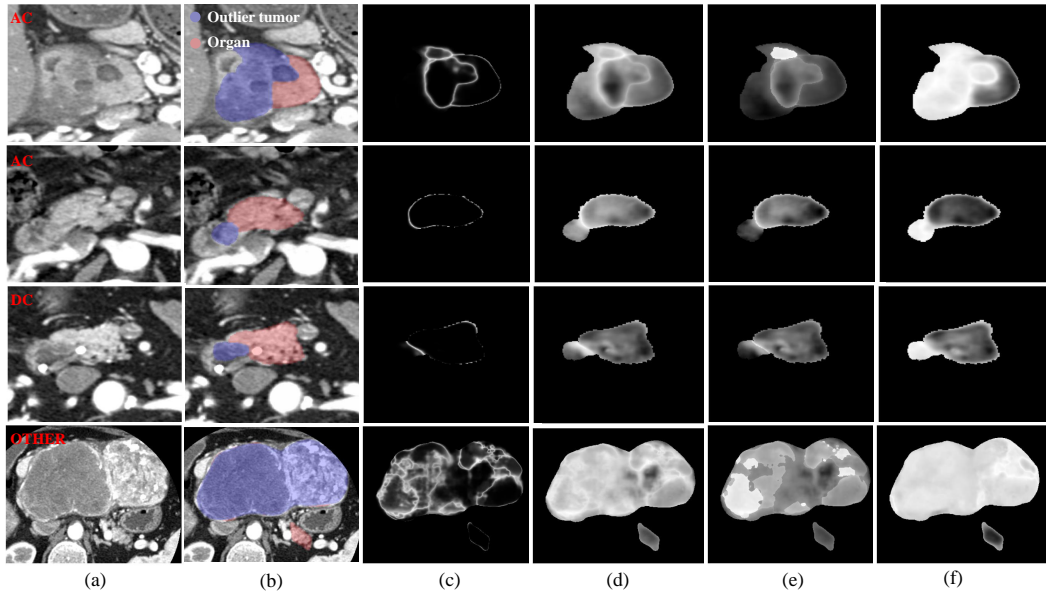
Figure 4. Visualization results of anomaly score map for OOD localization on *Pancreatic Tumors*: (a) CT slice, (b) ground truth (red: pancreas, blue: outlier tumor), (c) MSP [21], (d) MaxLogit [20], (e) SML [26] and (f) Ours. The grayscale level indicates the anomaly score. Our approach maintains a high anomaly score in the OOD pixels (outlier tumor), while a low anomaly score in the in-distribution pixels (organ). The four cases are selected from three different unknown diseases to show our method's robustness to tumor type.

(a) An in-distribution example.
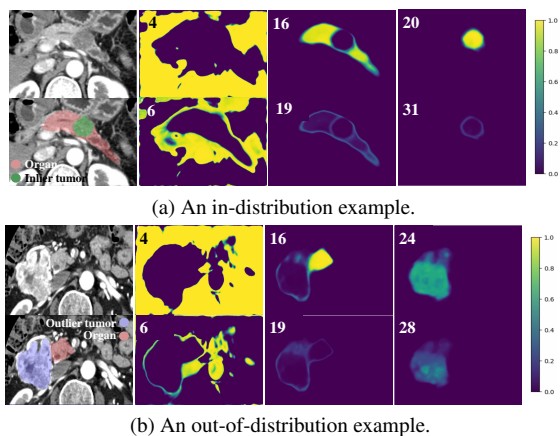

(b) An out-of-distribution example.

Figure 5. Visual examples of cluster assignments for (a) an in-distribution and (b) an out-of-distribution (OOD) sample. From left to right: (Column 1) image and ground truth with red: organ, green: inlier tumor, blue: outlier tumor; (Columns 2-4) representative object queries for background (C2), organ (C3) and tumor (C4), respectively. Query IDs are at the upper-left corners.

ous methods underestimate the anomalous score map. They tend only to highlight the boundaries of the OOD region, but our method preserves a high anomalous score on the entire OOD region.

In Tab. 2, our segmentation performance for inliers surpasses nnUNet by 6.30% in DSC. We find that Swin UN-ETR [45] could not converge to reasonable tumor segmentations, that might be due to its difficulty in identifying subtle tumor differences without sufficient data samples. These

improvements demonstrate that our framework can simultaneously detect common diseases with high accuracy and identify rare diseases in pixel-level localization and case-level diagnosis without requiring very large data samples.

We visualize the mask predictions of in-distribution and OOD examples to illustrate the working mechanism of object queries as cluster centers and how MaxQuery identifies the OOD condition. As shown in Fig. 5, for either in-distribution or OOD example, the background and organ regions are confidently activated by specific queries (Queries 4 and 6 for background, Query 16 for the target organ). Interestingly, regions with distinguishing features, such as the aorta or other abdominal organs, are not activated by the major cluster center (Query 4) but by an independent center (Query 6). This supports that the queries gradually converge to different meaningful centers. Furthermore, the corresponding queries of specific in-distribution tumors usually concentrate at a single center (Query 20 in Fig. 5a). Yet, queries corresponding to the OOD tumors seem to split into multiple centers with lower responses (Query 24 and 28 in Fig. 5b). The visual examples fulfill the motivation of the proposed MaxQuery that no inlier cluster centers can dominantly fit the OOD pixels.

Liver Tumors. Table 1 also shows the quantitative result on *Liver Tumors*. Our method outperforms the baselines in all evaluation metrics. Note that SML [26] improves the performance in OOD localization while dropping its performance in case-level OOD detection compared with MaxLogit [20], whereas our method performs well in both
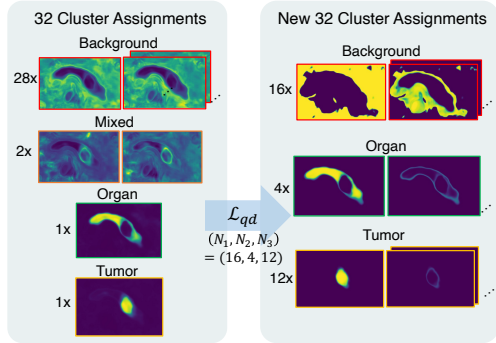
7

Figure 6. The effect of QD loss by visualizing the cluster assignment maps of the 32 queries on an inlier. **Left:** without QD loss, most queries redundantly focus on the background and some queries mix the background with foreground. **Right:** after using QD loss, we can manage the query distribution on the background, organ, and tumor with better seperation. The clear boundaries and high responses shows that QD loss encourages discriminative representation learning of the queries which will benefit both segmentation and OOD localization.

Query Dist. (N_1, N_2, N_3)	AUROC↑	AUPR↑	FPR↓	DSC_{inlier} ↑
SML [26]	56.10	30.44	77.81	35.47
(8, 4, 20)	84.44	51.32	42.10	36.50
(8, 20, 4)	83.73	49.76	43.32	39.79
(16, 4, 12)	82.52	**55.60**	46.90	**41.77**
(20, 4, 8)	85.66	55.17	37.24	38.19
(24, 4, 4)	**86.41**	52.58	**33.70**	39.43

Table 3. Ablation study on the distribution of queries. (DSC_{inlier}: mean Dice Score of inlier tumors.)

Level	Softmax	AUROC↑	AUPR↑	FPR95↓
Category	post	58.14	16.28	79.29
	pre	52.70	24.59	88.40
Query	post	76.88	33.98	55.82
	pre	**82.52**	**55.60**	**46.19**

Table 4. Comparison of category- and query-level anomaly scores. With the same network, the query-level anomaly scores show superiority over the category-level ones for OOD localization. Meanwhile, MaxQuery from the pre-softmax query-level scores outperforms that from post-softmax ones.

pixel and case level. Particularly, our method reaches a significantly lower FPR95 of 23.93% compared with previous approaches, which is crucial to localizing the OOD regions in medical scenarios. As shown in Tab. 2, our segmentation performance for inliers surpasses nnUNet by 4.24% in DSC. Swin UNETR [45] achieves Dice scores of 50.48% (HCC), 32.62% (ICC), 36.06% (Meta.), 71.82% (Heman.) and 15.30% (Cyst) on *Liver Tumors*, resulting in the average score of 41.26%. The qualitative analysis on *Liver Tumors* is in the Appendix.

4.4. Ablation Study

The Effect of the Query-Distribution Loss. Without the QD loss, the mean inlier tumor DSC of our framework increases only by a small margin compared to the nnUNet [24] baseline (Tab. 2). Fig. 6 presents query visualizations to show benefits from query-level guidance. Most queries redundantly represent the large and heterogeneous region of the background rather than the tumors without the QD loss (Fig. 6 left). With the QD loss, our framework is manipulated to provide fixed resources (queries for tumors) on distinguishing subtle differences of foregrounds for a near-OOD problem (Fig. 6 right). Final results are thus further improved on all metrics by large margins using QD loss (Tabs. 1 and 2). The results reveal that managing the object queries with QD loss contributes to masking transformers to improve both segmentation and OOD localization/detection performance.

The Distribution of Queries. We also perform an in-depth analysis of query distribution, as shown in Tab. 3. Our method shows robustness to different settings of query distribution. On all settings, our method outperforms the

previous leading method, SML [26], by a large margin in OOD localization and inlier segmentation. Eventually, we choose the hyper-parameter (N_1, N_2, N_3) as (16, 4, 12).

Pre-softmax versus Post-softmax for MaxQuery. As shown in Tab. 4, MaxQuery with pre-softmax score $\mathbf{R}$ exceeds the one with post-softmax $\mathbf{M}$ by 21.62% in AUPR for OOD localization, which agrees with our explanation in Sec. 3.3.

Query-level versus Category-level Anomaly Score. The debate of pre-softmax versus post-softmax corresponds to the one of MaxLogit [21] versus MSP [20]. Specifically, MSP calculates the post-softmax score in the final category level, while MaxLogit calculates the pre-softmax one. Unlike MSP and MaxLogit, our MaxQuery produces an anomalous score at the query level. For a fair comparison, we apply MSP and MaxLogit based on the Mask transformer we used in our model. As shown in Tab. 4, MaxQuery (post-softmax) outperforms MSP (category, post-softmax) by 17.70% and MaxQuery (pre-softmax) exceeds MaxLogit (category, pre-softmax) by 32.01% in AUPR. This comparison indicates the superiority of our query-level anomaly score over the category-level ones.

5. Conclusion

Processing a large collection of medical imaging data with long-tailed distributions has always been challenging. The significant performance improvement of our method on two real-world datasets validates its effectiveness. This result proves that interpreting segmentation as (query) cluster assignment is valid and effective. Our novel MaxQuery and QD loss are also evidently helpful for inlier segmentation

and (near-)OOD detection/localization, performing in practical scenarios. We believe that the proposed method has the good potential to further boost the adoption of medical image segmentation in designing various clinical applications.

6. Appendix

6.1. Dataset Details

We provide the abbreviation and full name for each disease from *Pancreatic Tumors* and *Liver Tumors* in Tabs. 1 and 2, respectively. Meanwhile, we report their incidence count in our datasets.

We determine the data splitting of known (inliers) and unknown classes (outliers) according to the real-world medical scenario and previous clinical studies [11,42]. For *Pancreatic Tumors*, we assign seven common pancreatic diseases (PDAC, PNET, SPT, IPMN, MCN, CP, and SCN) as inliers, and allocate two peri-pancreatic diseases (AC, DC) and "other" as outliers. The two peri-pancreatic diseases (AC, DC) are relatively difficult to distinguish from PDACs by radiologists, but clinical studies of pancreatic lesion diagnosis [11, 42] did not include them because they are not inside the pancreas. Thus we regard them as OOD in our model. For *Liver Tumors*, we assign five common liver tumors [53] (HCC, ICC, metastasis, hemangiomas, and cyst) as inliers, and allocate hepatoblastoma, FNH, and "other" as outliers, due to their low incidental rate.

Note that "other" class represents rare neoplasms or tumors in the real-world dataset, which reflects the long-tailed distribution of real-world disease incidence. Since these rare diseases are individually infrequent, it is impossible to collect them completely. Therefore, we address the thorny problem by OOD detection and localization.

Abbr.	Full name	Count
PDAC	Pancreatic ductal adenocarcinoma	366
IPMN	Intraductal papillary mucinous neoplasms	61
PNET	Pancreatic neuroendocrine tumor	35
SCN	Serous cystic neoplasms	46
CP	Chronic pancreatitis	43
SPT	Solid pseudopapillary tumor	32
MCN	Mucinous cystadenoma	7
AC	Ampullary cancer	46
DC	Bile duct cancer	12
"other"	Other rare neoplasms	13

Table 1. Dataset details of real-world *Pancreatic Tumors*. This full-spectrum dataset consists of ten pancreatic diseases, among which we assign the top seven as inlier tumors and the bottom three as outlier tumors, based on the real-world medical scenario and previous clinical studies [11, 42].

Abbr.	Full name	Count
HCC	Hepatocellular carcinoma	162
ICC	Intrahepatic cholangiocarcinoma	51
Meta.	Metastasis	97
Heman.	Hemangiomas	75
Cyst	Cyst	146
Hepato.	Hepatoblastoma	17
FNH	Focal nodular hyperplasia	27
"other"	Other rare tumors	60

Table 2. Dataset details of real-world *Liver Tumors*. This full-spectrum dataset includes seven liver tumors, among which we assign the top five as inlier tumors and the bottom three as outlier tumors, according to the real-world medical scenario and previous clinical studies [53].

6.2. Qualitative Results on Liver Tumors

For qualitative analysis on *Liver Tumors*, we present visual examples of anomaly score map for OOD localization in Fig. 1. This shows that our approach achieves a high anomaly score in the OOD pixels (outlier tumor), while a low anomaly score in the in-distribution pixels (organ), compared with other methods.

6.3. Baselines for Inlier Segmentation

Comparison with (nn)TransUNet. For a fair comparison with our method, we re-implement TransUNet [7] based on the framework of nnUNet [24], which is thus termed as (nn)TransUNet. TransUNet adopts transformer modules as pixel encoder, whereas our method uses CNN as the pixel-level backbone and leverages stand-alone transformer modules to interact with it. As presented in Tabs. 3 and 4, our method shows superiority on inlier segmentation compared with strong baselines, including nnUNet [24] and (nn)TransUNet [7]. This demonstrates that the distinctive architecture of our newly designed mask transformers leads to better performance on real-world medical image segmentation.

Methods	PDAC	IPMN	PNET	SCN	CP	SPT	MCN	Avg.
nnUNet [24]	65.65	27.60	32.59	36.46	23.33	31.73	30.96	35.47
(nn)TransUNet [7]	64.91	31.18	26.78	38.96	22.39	29.87	30.27	34.91
Ours	67.91	46.92	32.07	42.51	31.36	42.67	28.97	**41.77**

Table 3. Inlier segmentation Dice scores (%) on *val* set of *Pancreatic Tumors*.

Methods	HCC	ICC	Meta.	Heman.	Cyst	Avg.
nnUNet [24]	57.22	28.16	52.81	77.55	46.49	52.45
(nn)TransUNet [7]	52.26	25.50	42.31	70.90	47.52	47.70
Ours	67.61	30.78	60.40	77.07	47.61	**56.69**

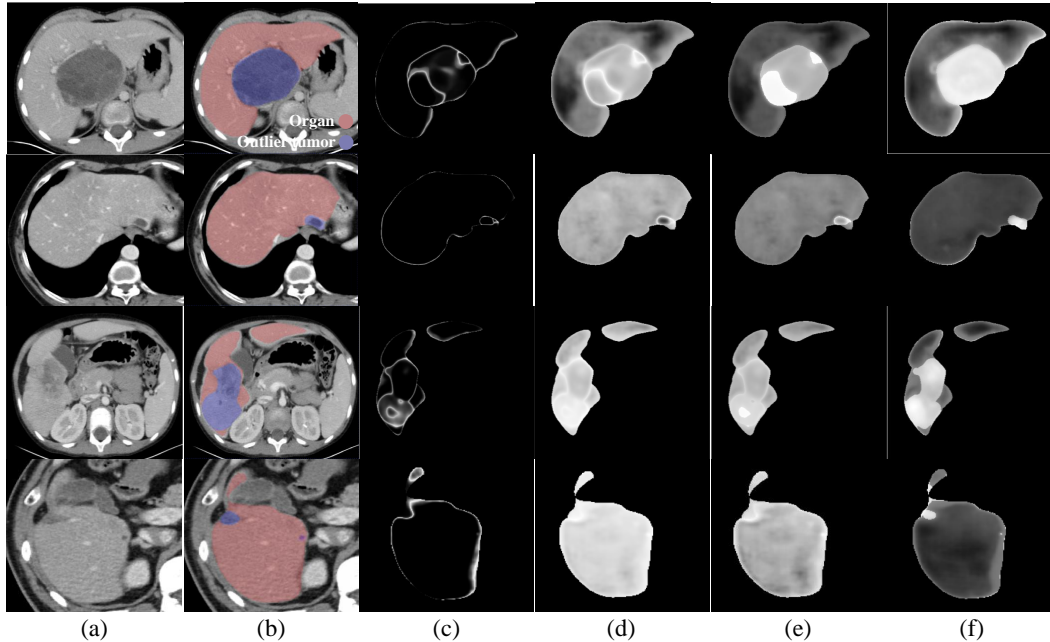Table 4. Inlier segmentation Dice scores (%) on *val* set of *Liver Tumors*.

Figure 1. Visualization results of anomaly score map for OOD localization on *Liver Tumors*: (a) 2D slices of the CT image, (b) ground truth annotation (red: liver, blue: outlier tumor), (c) MSP [21], (d) MaxLogit [20], (e) SML [26] and (f) Ours. The grayscale level indicates the anomaly score. Our method reaches a high anomaly score in the OOD pixels (outlier tumor), while a low anomaly score in the in-distribution pixels (organ).

References

[1] 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):1–13, 2022. 5

[2] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *NeurIPS*, 2016. 5

[3] Yun Bian, Zhilin Zheng, Xu Fang, Hui Jiang, Mengmeng Zhu, Jieyu Yu, Haiyan Zhao, Ling Zhang, Jiawen Yao, Le Lu, et al. Artificial intelligence to predict lymph node metastasis at CT in pancreatic ductal adenocarcinoma. *Radiology*, page 220329, 2022. 1

[4] Hermann Blum, Paul-Edouard Sarlin, Juan Nieto, Roland Siegwart, and Cesar Cadena. Fishyscapes: A benchmark for safe semantic segmentation in autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019. 3

[5] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pages 213–229. Springer, 2020. 2

[6] Robin Chan, Krzysztof Lis, Svenja Uhlemeyer, Hermann Blum, Sina Honari, Roland Siegwart, Pascal Fua, Mathieu Salzmann, and Matthias Rottmann. Segmentmeifyoucan: A benchmark for anomaly segmentation. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021. 3

[7] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L Yuille, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*, 2021. 2, 9

[8] 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

[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

[10] Linda C Chu, Seyoun Park, Satomi Kawamoto, Yan Wang, Yuyin Zhou, Wei Shen, Zhuotun Zhu, Yingda Xia, Lingxi Xie, Fengze Liu, et al. Application of deep learning to pancreatic cancer detection: lessons learned from our initial experience. *Journal of the American College of Radiology*, 16(9):1338–1342, 2019. 2

[11] Linda C Chu, Seyoun Park, Sahar Soleimani, Daniel F Fouladi, Shahab Shayesteh, Jin He, Ammar A Javed, Christopher L Wolfgang, Bert Vogelstein, Kenneth W Kinzler, et al. Classification of pancreatic cystic neoplasms using radiomic feature analysis is equivalent to an experienced academic radiologist: a step toward computer-augmented diagnostics for radiologists. *Abdominal Radiology*, pages 1–12, 2022. 5, 9

[12] Jeffrey De Fauw, Joseph R Ledsam, Bernardino Romera-Paredes, Stanislav Nikolov, Nenad Tomasev, Sam Black-

well, Harry Askham, Xavier Glorot, Brendan O'Donoghue, Daniel Visentin, et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nature Medicine*, 24(9):1342–1350, 2018. 2

[13] Terrance DeVries and Graham W Taylor. Learning confidence for out-of-distribution detection in neural networks. *arXiv preprint arXiv:1802.04865*, 2018. 3

[14] Hexin Dong, Zifan Chen, Mingze Yuan, Yutong Xie, Jie Zhao, Fei Yu, Bin Dong, and Li Zhang. Region-aware metric learning for open world semantic segmentation via meta-channel aggregation. In *31th International Joint Conference on Artificial Intelligence (IJCAI-22)*, 2022. 3, 6

[15] Xin Dong, Junfeng Guo, Ang Li, Wei-Te Ting, Cong Liu, and HT Kung. Neural mean discrepancy for efficient out-of-distribution detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19217–19227, 2022. 3

[16] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021. 2

[17] 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, 2022. 2

[18] Ali Hatamizadeh, Yucheng Tang, Vishwesh Nath, Dong Yang, Andriy Myronenko, Bennett Landman, Holger R Roth, and Daguang Xu. Unetr: Transformers for 3d medical image segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 574–584, 2022. 2

[19] Mattias P Heinrich, Mark Jenkinson, Michael Brady, and Julia A Schnabel. Mrf-based deformable registration and ventilation estimation of lung ct. *IEEE Transactions on Medical Imaging*, 32(7):1239–1248, 2013. 5

[20] Dan Hendrycks, Steven Basart, Mantas Mazeika, Mohammadreza Mostajabi, Jacob Steinhardt, and Dawn Song. Scaling out-of-distribution detection for real-world settings. *ICML*, 2022. 6, 7, 8, 10

[21] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *International Conference on Learning Representations, ICLR*, 2017. 3, 6, 7, 8, 10

[22] Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. *International Conference on Learning Representations, ICLR*, 2019. 3

[23] Gao Huang, Yu Sun, Zhuang Liu, Daniel Sedra, and Kilian Q Weinberger. Deep networks with stochastic depth. In *European conference on computer vision*, pages 646–661. Springer, 2016. 5

[24] 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, 4, 5, 6, 8, 9

[25] Fabian Isensee, Jens Petersen, Andre Klein, David Zimmerer, Paul F Jaeger, Simon Kohl, Jakob Wasserthal, Gregor Koehler, Tobias Norajitra, Sebastian Wirkert, et al. nnu-net: Self-adapting framework for u-net-based medical image segmentation. *arXiv preprint arXiv:1809.10486*, 2018. 2

[26] 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. 3, 6, 7, 8, 10

[27] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems*, 30, 2017. 6

[28] Philipp Kickingereder, Fabian Isensee, Irada Tursunova, Jens Petersen, Ulf Neuberger, David Bonekamp, Gianluca Brugnara, Marianne Schell, Tobias Kessler, Martha Foltyn, et al. Automated quantitative tumour response assessment of MRI in neuro-oncology with artificial neural networks: a multicentre, retrospective study. *The Lancet Oncology*, 20(5):728–740, 2019. 1, 2

[29] Kimin Lee, Honglak Lee, Kibok Lee, and Jinwoo Shin. Training confidence-calibrated classifiers for detecting out-of-distribution samples. *International Conference on Learning Representations, ICLR*, 2018. 3

[30] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *Advances in Neural Information Processing Systems*, 2018. 3

[31] Xiaomeng Li, Hao Chen, Xiaojuan Qi, Qi Dou, Chi-Wing Fu, and Pheng Ann Heng. H-DenseUNet: Hybrid densely connected UNet for liver and liver tumor segmentation from CT volumes. *IEEE Transactions on Medical Imaging*, 2017. 2

[32] Krzysztof Lis, Krishna Nakka, Pascal Fua, and Mathieu Salzmann. Detecting the unexpected via image resynthesis. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2152–2161, 2019. 3

[33] Liyuan Liu, Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Jiawei Han. On the variance of the adaptive learning rate and beyond. In *International Conference on Learning Representations*, 2019. 5

[34] Siqi Liu, Daguang Xu, S Kevin Zhou, Olivier Pauly, Sasa Grbic, Thomas Mertelmeier, Julia Wicklein, Anna Jerebko, Weidong Cai, and Dorin Comaniciu. 3d anisotropic hybrid network: Transferring convolutional features from 2d images to 3d anisotropic volumes. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 851–858. Springer, 2018. 2

[35] Yuan Liu, Ayush Jain, Clara Eng, David H Way, Kang Lee, Peggy Bui, Kimberly Kanada, Guilherme de Oliveira Marinho, Jessica Gallegos, Sara Gabriele, et al. A deep learning system for differential diagnosis of skin diseases. *Nature Medicine*, 26(6):900–908, 2020. 1

[36] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In

Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 10012–10022, 2021. 2

[37] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 Fourth International Conference on 3D Vision (3DV)*, pages 565–571. IEEE, 2016. 2

[38] Hossein Mirzaei, Mohammadreza Salehi, Sajjad Shahabi, Efstratios Gavves, Cees GM Snoek, Mohammad Sabokrou, and Mohammad Hossein Rohban. Fake it till you make it: Near-distribution novelty detection by score-based generative models. *arXiv preprint arXiv:2205.14297*, 2022. 3

[39] Jie Ren, Stanislav Fort, Jeremiah Liu, Abhijit Guha Roy, Shreyas Padhy, and Balaji Lakshminarayanan. A simple fix to mahalanobis distance for improving near-ood detection. *arXiv preprint arXiv:2106.09022*, 2021. 3

[40] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 2

[41] 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. 1, 2, 3

[42] Simeon Springer, David L Masica, Marco Dal Molin, Christopher Douville, Christopher J Thoburn, Bahman Afsari, Lu Li, Joshua D Cohen, Elizabeth Thompson, Peter J Allen, et al. A multimodality test to guide the management of patients with a pancreatic cyst. *Science Translational Medicine*, 11(501):eaav4772, 2019. 5, 9

[43] Robin Strudel, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid. Segmenter: Transformer for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7262–7272, 2021. 2

[44] Hao Tang, Xuming Chen, Yang Liu, Zhipeng Lu, Junhua You, Mingzhou Yang, Shengyu Yao, Guoqi Zhao, Yi Xu, Tingfeng Chen, et al. Clinically applicable deep learning framework for organs at risk delineation in ct images. *Nature Machine Intelligence*, 1(10):480–491, 2019. 2

[45] Yucheng Tang, Dong Yang, Wenqi Li, Holger R Roth, Bennett Landman, Daguang Xu, Vishwesh Nath, and Ali Hatamizadeh. Self-supervised pre-training of swin transformers for 3d medical image analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20730–20740, 2022. 2, 6, 7, 8

[46] 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 *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 128–140. Springer, 2021. 2, 3

[47] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*, 2016. 5

[48] Huiyu Wang, Yukun Zhu, Hartwig Adam, Alan Yuille, and Liang-Chieh Chen. Max-deeplab: End-to-end panoptic segmentation with mask transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5463–5474, 2021. 2, 5

[49] Jim Winkens, Rudy Bunel, Abhijit Guha Roy, Robert Stanforth, Vivek Natarajan, Joseph R Ledsam, Patricia MacWilliams, Pushmeet Kohli, Alan Karthikesalingam, Simon Kohl, et al. Contrastive training for improved out-of-distribution detection. *arXiv preprint arXiv:2007.05566*, 2020. 2, 3

[50] Yingda Xia, Yi Zhang, Fengze Liu, Wei Shen, and Alan L Yuille. Synthesize then compare: Detecting failures and anomalies for semantic segmentation. In *European Conference on Computer Vision*, pages 145–161. Springer, 2020. 3, 6

[51] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, 34:12077–12090, 2021. 2

[52] Jiawen Yao, Kai Cao, Yang Hou, Jian Zhou, Yingda Xia, Isabella Nogues, Qike Song, Hui Jiang, Xianghua Ye, Jianping Lu, et al. Deep learning for fully automated prediction of overall survival in patients undergoing resection for pancreatic cancer: A retrospective multicenter study. *Annals of Surgery*, 2022. 1, 2

[53] Koichiro Yasaka, Hiroyuki Akai, Osamu Abe, and Shigeru Kiryu. Deep learning with convolutional neural network for differentiation of liver masses at dynamic contrast-enhanced CT: a preliminary study. *Radiology*, 286(3):887–896, 2018. 5, 9

[54] Lequan Yu, Xin Yang, Hao Chen, Jing Qin, and Pheng-Ann Heng. Volumetric convnets with mixed residual connections for automated prostate segmentation from 3D MR images. In *AAAI*, 2017. 2

[55] Qihang Yu, Huiyu Wang, Dahun Kim, Siyuan Qiao, Maxwell Collins, Yukun Zhu, Hartwig Adam, Alan Yuille, and Liang-Chieh Chen. Cmt-deeplab: Clustering mask transformers for panoptic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2560–2570, 2022. 2, 3

[56] Qihang Yu, Huiyu Wang, Siyuan Qiao, Maxwell Collins, Yukun Zhu, Hartwig Adam, Alan Yuille, and Liang-Chieh Chen. k-means mask transformer. In *European Conference on Computer Vision*, pages 288–307. Springer, 2022. 2, 3, 5, 6

[57] Qihang Yu, Yingda Xia, Yutong Bai, Yongyi Lu, Alan L Yuille, and Wei Shen. Glance-and-gaze vision transformer. *Advances in Neural Information Processing Systems*, 34:12992–13003, 2021. 2

[58] Ling Zhang, Vissagan Gopalakrishnan, Le Lu, Ronald M Summers, Joel Moss, and Jianhua Yao. Self-learning to detect and segment cysts in lung ct images without manual annotation. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 1100–1103. IEEE, 2018. 5

12

[59] 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, 2

[60] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip HS Torr, et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6881–6890, 2021. 2

[61] S Kevin Zhou, Hayit Greenspan, Christos Davatzikos, James S Duncan, Bram Van Ginneken, Anant Madabhushi, Jerry L Prince, Daniel Rueckert, and Ronald M Summers. A review of deep learning in medical imaging: Imaging traits, technology trends, case studies with progress highlights, and future promises. *Proceedings of the IEEE*, 109(5):820–838, 2021. 1

[62] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: Redesigning skip connections to exploit multiscale features in image segmentation. *IEEE Transactions on Medical Imaging*, 39(6):1856–1867, 2019. 2

[63] 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. 2, 3